Using Morpho-Syntactic and Semantic Information to Improve Statistical Machine Translation

Von der Fakultät für Informatik, Elektrotechnik und Informationstechnik
der Universität Stuttgart zur Erlangung der Würde eines Doktors
der Philosophie (Dr. phil.) genehmigte Abhandlung

Vorgelegt von
Marion Di Marco
aus Frankfurt am Main

Hauptberichterin:	PD Dr. Sabine Schulte im Walde
1. Mitberichter:	Prof. Dr. Alexander M. Fraser
2. Mitberichter:	Prof. Dr. Jonas Kuhn

Tag der mündlichen Prüfung: 29.09.2017

Institut für Maschinelle Sprachverarbeitung
der Universität Stuttgart

2018

"Sometimes, if you pay real close attention to the pebbles you find out about the ocean."

Terry Pratchett

Abstract

Machine translation is the process of automatically translating a document from one language into another by means of a computer program. Typically, statistical machine translation (SMT) systems are learned from phrase-translation pairs observed in parallel corpora. While such translation models work very well when trained on large corpora, they have two general shortcomings: they cannot generalize from the observed training instances, and the learned translational equivalents are often imprecise or contextually inadequate due to unaccounted for differences in the underlying language pair. This is particularly the case if the involved languages are different in terms of morphological complexity or syntactic structure.

For example, a translation system translating into a morphologically complex language like German does not "know" that the different forms *blau, blaue, blaues, blauem, blauen, blauer* are all instances of the lemma *blau* ('blue'), but considers inflectional variants as unrelated words and stores them as separate and unconnected translation options: this is far from optimal in terms of modeling, and also restricts the system to forms observed in the training data. Furthermore, words or phrases observed in one context are not necessarily applicable in other contexts: while the translation pair *the blue car* → *das blaue Auto* is correct when used as subject or direct object, it needs to be inflected differently when used as indirect object (*dem blauen Auto*) or as possessive construction (*des blauen Autos*). Further differences between languages may occur on the structural level, for example by using a noun phrase in one language and a prepositional phrase in the other language, such as *to remember [sth.]*$_{NP}$ → *(sich) [an etw.]*$_{PP}$ *erinnern*, and on a more global level, such as adhering to a subject-verb-object sentence structure in English, in contrast to a more flexible word order in German.

The more different the source language and the target language are, the more difficult it becomes to find balanced and precise phrase-translation pairs as a basis to learn good models. As a translation system has only knowledge about phrase-translation pairs and their (immediate) context, it cannot fully grasp the complex interactions between the source and target language, but only fragments thereof. For example, without understanding the relevant features and conditions applying to inflection, it is impossible to make an informed decision when selecting one out of the many forms of *blau* observed as possible translations for *blue*. This results in issues that do not only concern the *morpho-syntactic level*, but also entail a *semantic dimension*, as for example grammatical case is tied to the perception of semantic roles.

Many errors in machine translation can be attributed to the fact that the complexity of language is not sufficiently represented in the translation model. The relevant criteria to support the generation of a good translation are often not immediately evident in the training data used to build an SMT system, but can be formulated on a *linguistic level*. It may thus be a promising approach to integrate explicit linguistic information into a statistical translation system, with the objective to provide the model with structured information to improve the modeling of linguistic phenomena. This thesis explores the integration of morpho-syntactic and semantic information to improve the translation quality of an English–German statistical machine translation system. By adjusting the representation of the parallel training data, differences between the source and the target side are reduced such that the contained linguistic information is more similar. Additional information can then be added at different stages in the translation process, either as pre- or post-processing, or into the phrase-table to be used at decoding time. The experiments in this thesis show that the integration of explicit linguistic information can improve the translation quality.

The basis of the presented research is a two-step translation system that handles target-side morphology by first translating into an underspecified representation, and then generating fully inflected forms in a post-processing step. In the translation model, all inflected forms *blau, blaues, blauer, ...* are replaced by the abstract form `blau<ADJ>`, which allows for a better generalization at the lexical level. Then, the generation component has access to information relevant for target-side inflection to predict the features left underspecified in the translation step as basis for the generation of inflected forms with an external morphological resource. The generation step allows in particular to produce inflected forms not observed in the training data. This basic inflection prediction model is then extended with syntactically and semantically motivated information to improve the modeling of subcategorization, with a particular focus on the selection of prepositions, such that the model learns to recognize whether a phrase should be realized as a subject or an object (*das blaue Auto* vs. *dem blauen Auto*), or rather as a prepositional phrase with an appropriate preposition (*[an das blaue Auto]*$_{PP}$), as required in the context of *remember* → *erinnern*.

Zusammenfassung

Maschinelle Übersetzung ist das automatische Übersetzen von Texten von einer Sprache in eine andere Sprache mit Hilfe eines Computerprogramms. Statistische maschinelle Übersetzungssysteme (SMÜ) werden üblicherweise anhand von Phrasen-Übersetzungspaaren gelernt, die aus parallelen Daten abgeleitet werden. Solche Übersetzungsmodelle können sehr gut funktionieren, wenn sie auf einer entsprechend großen Datengrundlage gelernt werden, haben aber zwei Haupteinschränkungen: Sie können nicht von den gesehenen Trainingsinstanzen abstrahieren, und die gelernten Übersetzungspaare sind häufig nicht sehr präzise, da beim Lernen des Systems Unterschiede in dem zugrundeliegenden Sprachpaar nicht berücksichtigt werden. Das trifft ganz besonders bei solchen Sprachpaaren zu, die sich in ihrer morphologischen Komplexität oder syntaktischen Struktur unterscheiden.

Wenn man zum Beispiel ein Übersetzungssystem hat, das in eine morphologisch komplexe Sprache wie das Deutsche übersetzt, dann "weiß" das Modell nicht, dass die unterschiedlichen Formen *blau, blaue, blaues, blauem, blauen* und *blauer* alle zu dem gleichen Lemma *blau* gehören und eine Übersetzung von *blue* darstellen, sondern behandelt sie wie vollkommen unterschiedliche Wörter, und die verschiedenen Formen werden als separate und nicht verwandte Übersetzungssmöglichkeiten gespeichert. Diese Art der Modellierung ist nicht optimal, und zudem auch auf die Menge der Formen beschränkt, die in den parallel Daten vorkommen. Weiterhin kommt es vor, dass Wörter und Wortsequenzen, die während des Trainings in einem bestimmten Kontext vorkamen, nicht in jedem anderen Kontext zum Übersetzen eingesetzt werden können: das Übersetzungspaar *the blue car → das blaue Auto* ist zwar korrekt, aber nur, wenn es als Subjekt oder direktes Objekt benutzt wird; als Possessivkonstruktion (*des blauen Autos*) oder als indirektes Objekt (*dem blauen Auto*) muss es anders flektiert werden. Weiterhin können sich Sprachen auch strukturell unterscheiden, zum Beispiel indem eine Präpositionalphrase in der einen Sprache einer Nominalphrase in der anderen Sprache entspricht (*to remember [sth.]*$_{NP}$ *→ (sich) [an etw.]*$_{PP}$ *erinnern*), oder auf globaler Ebene durch unterschiedliche Satzstellungen, wie zum Beispiel Subjekt-Verb-Objekt im Englischen, im Vergleich zu einer relativ flexiblen Satzstellung im Deutschen.

Je mehr sich die Quellsprache und Zielsprache voneinander unterscheiden, desto schwieriger ist es, ausgewogene und präzise Übersetzungspaare als Grundlage für das Übersetzungsmodell zu finden. Da ein Übersetzungsmodell nur Informationen über die beobachteten Phrasen-Übersetzungspaare und deren (unmittelbaren) Kontext hat, kann

es die komplexen Zusammenhänge zwischen Quell- und Zielsprache nicht vollständig erfassen. Zum Beispiel ist es unmöglich, ohne Wissen über Flektionsmerkmale und weitere relevante Kriterien die passende Form von *blau* aus der Menge der möglichen Formen auszuwählen, die als Übersetzung von *blue* in Frage kommen. Das führt zu Problemen, die nicht nur die *morpho-syntaktische Ebene* betreffen, sondern mit denen auch eine *semantische Dimension* einhergeht, da zum Beispiel die Realisierung von Argumenten an die Wahrnehmung von semantischen Rollen geknüpft ist.

Viele der Fehler in maschineller Übersetzung können darauf zurückgeführt werden, dass das Übersetzungsmodell die Komplexität natürlicher Sprache nicht ausreichend widerspiegelt. Die Kriterien, die für eine gute Übersetzung notwendig sind, sind häufig nicht direkt aus den Trainingsdaten ersichtlich, können aber *linguistisch* definiert werden. Die Integration von expliziter linguistischer Information in ein Übersetzungssystem ist also eine interessante und vielversprechende Methode, dem Modell strukturierte Information zur Verfügung zu stellen, und somit eine bessere Modellierung linguistischer Phänomene zu ermöglichen. Die vorliegende Arbeit untersucht Strategien, mit der Integration von morpho-syntaktischer und semantischer Information ein Englisch–Deutsches Übersetzungssystem zu verbessern. Durch das Anpassen der Repräsentation der Trainingsdaten werden Unterschiede zwischen der Quell- und Zielsprache so reduziert, dass sich die jeweils enthaltenen linguistischen Informationen angleichen. Weitere Informationen können dann an unterschiedlichen Stellen im Übersetzungsablauf hinzugefügt werden, beispielsweise in einem Vor- oder Nachverarbeitungsschritt, oder in der Phrasen-Tabelle, die beim Übersetzen benutzt wird. Die Versuche in dieser Arbeit zeigen, dass die Integration von expliziter linguistischer Information die Übersetzungsqualität verbessern kann.

Die Grundlage dieser Arbeit ist ein Übersetzungssystem zur Modellierung von Flektion in der Zielsprache, das in zwei Schritten arbeitet: auf das Übersetzen in eine unterspezifizierte Repräsentation folgt ein Nachverarbeitungsschritt, in dem vollständig flektierte Formen generiert werden. Beim Übersetzen werden alle flektierten Formen *blau, blaues, blauer, ...* durch die abstrakte Form `blau<ADJ>` ersetzt, wodurch das Modell auf lexikalischer Ebene besser generalisieren kann. Im Generierungsschritt werden dann genau die Informationen eingesetzt, die erforderlich sind, um die Merkmale zu berechnen, die im Übersetzungsschritt unterspezifiziert geblieben sind. Mit einem externen morphologischen Werkzeug werden dann schließlich flektierte Formen erzeugt, was insbesondere auch die Generierung von Flektionsvarianten ermöglicht, die nicht in den Trainingsdaten vorkommen.

Dieses Übersetzungsmodell wird dann mit syntaktisch und semantisch motivierten Informationen erweitert, um die Modellierung von Subkategorisierung, und besonders die Wahl von Präpositionen, zu verbessern – das Modell soll erkennen, ob eine Phrase ein Subjekt oder Objekt ist (*das blaue Auto* vs. *dem blauen Auto*), oder als Präpositionalphrase mit einer bestimmten Präposition realisiert werden muss, wie etwa *[an das blaue Auto]*$_{PP}$ im Kontext von *remember* → *erinnern*.

List of Publications

The thesis includes the research published in the papers listed below. With the exception of the work described in Fraser et al. (2012), I was the main person responsible for implementing the presented approaches, carrying out the experiments, and writing the actual papers. My co-authors contributed through discussions and feedback. Regarding the publication Fraser et al. (2012) I specify my contributions below.

- Alexander Fraser, Marion Weller, Aoife Cahill and Fabienne Cap (2012). *Modeling Inflection and Word-Formation in SMT.* In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL). Avignon/France, pp. 664–674.

- Marion Weller, Alexander Fraser and Sabine Schulte im Walde (2013). *Using Subcategorization Knowledge to Improve Case Prediction for Translation to German.* In Proceedings of the Association for Computational Linguistics (ACL). Sofia/Bulgaria, pp. 593–603.

- Marion Weller, Sabine Schulte im Walde and Alexander Fraser (2014). *Using Noun Class Information to Model Selectional Preferences for Translating Prepositions in SMT.* In Proceedings of the Association for Machine Translation in the Americas (AMTA). Vancouver/Canada, pp. 275–287.

- Marion Weller, Alexander Fraser and Sabine Schulte im Walde (2015). *Target-Side Generation of Prepositions for SMT.* In Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT). Antalya/Turkey, pp. 177–184.

- Marion Weller-Di Marco, Alexander Fraser, and Sabine Schulte im Walde (2016). *Modeling Complement Types in Phrase-Based SMT.* In Proceedings of the First Conference of Machine Translation (WMT16). Berlin/Germany, pp. 43–53.

- Marion Weller-Di Marco, Alexander Fraser, and Sabine Schulte im Walde (2017). *Combining Approaches to Model Morphology, Syntax and Lexical Choice.* In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Short Papers (EACL). Valencia/Spain, pp. 625–630.

Notes My contribution to the Fraser et al. (2012) paper consisted in designing the stemmed representation and refining the stem markup used for feature prediction, as well as the pipeline to generate inflected forms in the post-processing step.

As first experiments with Hidden Markov Models (HMMs) for the feature prediction did not work sufficiently well, the prediction setup was changed to use more powerful Conditional Random Fields (CRFs). To assess the influence of joint vs. individual feature prediction, I compared the prediction accuracy of four separate HMMs and a single joint HMM to justify the use of four separate CRFs, as one single CRF predicting all features jointly would not have been tractable.

Acknowledgements

I would like to thank Sabine Schulte im Walde and Alex Fraser for being great Ph.D. advisors – thank you for so many interesting and constructive discussions, always having an open ear, and for cheering me up during those phases when none of my experiments seemed to work ... and last, but not least, for all the cookies, coffee and chocolate!

A big thank you goes to Ulrich Heid, who gave me the possibility to carry out interesting research projects as a *Hiwi* – it was especially this pleasant work experience that encouraged me to stay at the university and to pursue my PhD.

During our time as *Hiwis*, Fabienne Cap and I started to work in our first collaboration, with many more to come during our time together at the IMS. It was always a great pleasure to work with Fabienne, and often more fun than work!

Boris Haselbach has been the best office mate I could have asked for – *takk beaucoup* for so many enjoyable conversations about work-related and not so work-related topics. Going for lunch with the *early Mensagroup* was always a fun part of the day.

I would also like to thank Sibylle Laderer, Peggy Hobmaier and Ingrid Trojan for taking great care of all administrative issues, both in Stuttgart and in Munich. At this point, I would like to thank Sabine Schulte im Walde and Jonas Kuhn for sorting out unexpected bureaucratic problems at the very end of the thesis.

The biggest thank you goes to Daniel for always being there for me – this thesis would not have been possible without you!

This project has received funding from the DFG grants *Distributional Approaches to Semantic Relatedness* and *Models of Morphosyntax for Statistical Machine Translation* (Phase One and Phase Two), from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402 (HimL) and from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n. 248005 (TTC).

Contents

List of Abbreviations

BLEU	**B**i**L**ingual **E**valuation **U**nderstudy
CRF	**C**onditional **R**andom **F**ield
HMM	**H**idden **M**arkov **M**odel
LL	**L**og-**L**ikelihood
LM	**L**anguage **M**odel
ME	**M**aximum **E**ntropy
MEMM	**M**aximum **E**ntropy **M**arkov **M**odel
MERT	**M**inimum **E**rror **R**ate **T**raining
NLP	**N**atural **L**anguage **P**rocessing
NMT	**N**eural **M**achine **T**ranslation
NP	**N**oun **P**hrase
POS	**P**art-**o**f-**S**peech
PP	**P**repositional **P**hrase
RBMT	**R**ule-**B**ased **M**achine **T**ranslation
SMT	**S**tatistical **M**achine **T**ranslation

Chapter 1

Introduction

Machine translation is the process of automatically translating a document from one language into another by means of a computer program. Machine translation applications are becoming indispensable in many everyday and commercial applications. Usage scenarios include made-to-measure translation systems for a particular domain to translate documents that are then revised by a professional translator, or just a quick translation by non-expert casual users with online systems such as Google Translate.

Statistical machine translation systems are built on statistics derived from phrase-to-phrase translations observed in parallel corpora. While such translation models work very well when trained on large amounts of data, they cannot generalize from the seen training instances. For example, words are simply treated as strings, and inflectional variants of the same lemma are not recognized as related words. The lack of generalization in a standard translation system applies not only to the *morphological* level, but also extends through the *syntactic* and *semantic* levels, which is reflected in translation errors concerning, for example, subcategorization and semantic roles. While some errors in machine translation can simply be addressed by the addition of more training data, the most challenging – and persistent – errors in machine translation are based in the complexity of language that translation systems fail to fully grasp.

The relevant criteria to support the generation of a good translation are complex and depend on many factors, which are often not immediately evident in the data, but are obvious on a *linguistic* level. It may thus be a promising and interesting approach to integrate explicit linguistic information into SMT systems, with the objective to provide the translation system with structured information to improve the modeling of a particular translation problem.

This thesis explores the integration of rich linguistic information into machine translation: relying on a translation system that handles *morphology* on the target side, *syntactically* and *semantically* motivated information is added to model translation-relevant aspects concerning subcategorization with a particular focus on the choice of prepositions, which continue to be a challenging problem in machine translation.

1.1 Motivation

Statistical translation systems are typically trained on parallel word-aligned corpora, with an additional language model built on large quantities of target-side data. While monolingual data is typically abundantly available, parallel data is restricted to much smaller corpora. As a result, translation systems can run into data-sparsity issues, in particular when languages with rich morphology are involved, as the richness in inflected forms prevents the model from efficiently learning lexical relationships between source and target side: the connection between different inflectional realizations of one base form is simply ignored.

However, data sparsity is not the only problem, as becomes clear when looking at how translation models are created: The bilingual translation model, a core part of SMT systems, is learned from source-target phrase pairs extracted from word-aligned parallel data. Based on these pairs, phrase translation probabilities, in form of relative frequencies, and other statistics are derived. The quality of the translation pairs and the accompanying translation scores depends on how well word alignments between the source and the target language can be established, and how precise a source fragment represents the respective target fragment: the more different the source and the target language are, the more difficult it is to find well-mapped word or phrase pairs as a basis to train good models.

Languages differ on many linguistic levels, in particular if they are positioned at opposite sides in the spectrum between analytic (such as English) and synthetic (such as German) languages. For example, differences can occur at the *syntactic level*, in form of contrasting word order structures. While English follows a rather strict subject-verb-object ordering, German is more flexible and allows for several sentence structures, including “verb-second” and “verb-final” structures. Differences can also apply to the *morphological level*: source language and target language may differ with respect to the degree of morphological complexity, and also regarding the type of information that is represented through morphology. The syntactic and morphological level often interact with each other; for example, information about a constituent’s role in the sentence can be encoded at the syntactic level (through the position in the sentence in English) or at the morphological level (through grammatical case in German). This adds a *semantic dimension*: the translation system needs to reproduce the content of the input sentence – *who did what when, where and why* – by deriving the relevant information from the source sentence and realizing it appropriately in the target sentence such that the translation is a well-formed equivalent of the input sentence.

Standard machine translation systems do not make use of explicit linguistic information, but output the most probable sequence of target-language words without taking into account semantic plausibility and target-side morpho-syntactic constraints. This is often not optimal, as the translation process and (target-side) inflectional processes are commingled in a way that the system is cluttered with information that is irrelevant at translation time, while relevant information is not or only indirectly accessible.

This thesis explores strategies to reduce differences between the source and target side by modifying the representation of source- and target-side information to correspond to the needs of the respective steps of *training*, *translation* and *generation*. This allows to handle information more efficiently during the translation process, and to furthermore integrate rich linguistic information at the morpho-semantic level. A recurrent concept in this thesis is a *two-step underspecification-generation approach* that first creates a meaningful intermediate representation to minimize the differences between source and target language. In a second step, the intermediate representation undergoes a prediction process in which surface word forms are generated based on predicted features conditioned on the respective context. This two-step strategy evokes the research question of how to efficiently arrange information within the translation process, namely

- what are differences between source side and target side, and how do they affect the translation?
- how to efficiently include relevant information and how to represent the data in order to minimize these differences at training time?
- which processes should be handled during translation, as opposed to being modeled in an external (post-processing) component?
- how to integrate linguistic information, potentially obtained from external resources, into the generation step?

The main focus of this thesis is the modeling of target-side morphology with a focus on subcategorization and the choice of prepositions. The basis is provided by a *morphology-aware translation system* that operates on a morphologically reduced representation annotated with inflection-relevant features; a post-processing step then generates inflected forms relying on predicted inflectional features and a morphological resource. Separating the translation step and morphological generation allows the system to generate target-side inflected forms as needed by the context and independently from their occurrence in the parallel training data, while benefiting from a more general translation model at translation time. The inflection prediction model is then extended to tackle two challenging problems in machine translation: *subcategorization* and the translation of *prepositions*. Using rich source- and target-side features improves the prediction of grammatical case in the inflection step, and thus the realization of subcategorization by preserving source-side syntactic functions while meeting target-side subcategorization requirements and selectional preferences. A first attempt to model prepositions aims at enforcing selectional preferences by integrating noun class information to provide a better basis for the choice of prepositions. A second attempt goes back to the concept of predicting surface forms from an intermediate placeholder: prepositions are treated as a target-side generation problem and are predicted in the translation output. These strategies to model prepositions show that (i) prepositions cannot be modeled without also looking at a broader context, including all subcategorized elements by a verb; and that (ii) modeling prepositions in a post-processing step is insufficient. A third strategy thus consists in modeling subcategorization across *complement types* – noun phrases and prepositional phrases – by creating *synthetic phrases* containing the optimal complement type realization for the given context to be used at decoding time.

At the end of the thesis, the focus shifts to a higher level by investigating combinations of strategies to handle problems across linguistic levels: approaches to address differences at the morpho-syntactic level, namely morphological complexity and structural differences between the source and target side, are combined with an approach aiming at the lexical level. The experiments show that the investigated techniques are complementary.

1.2 Contributions

The main objective of my thesis research is the improvement of statistical machine translation by adjusting the representation of the source and target sides of the training data to be more similar in structure and linguistic information. Furthermore, rich linguistic information that is not directly accessible to a standard translation system is integrated. The main contributions of the thesis are summarized below:

Inflection prediction system The system separates target-side inflection from the translation step, and thus combines the advantages of being more generic during translation with the ability to generate context-appropriate forms in a post-processing step. In many scenarios, depending on domain and training data size, this system outperforms baseline systems trained on standard surface data. Furthermore, the abstract representation and the flexibility to model inflectional features based on the sentence context provide the basis for the research in the following parts of the thesis.

The inflection prediction system was developed in several stages: I started working on modeling target-side morphology already in my *Studienarbeit* (Weller, 2009), although in a monolingual setting only, and the resulting system was not yet functioning in an actual machine translation system. In Fraser et al. (2012), the inflection prediction pipeline outlined in the *Studienarbeit* was adapted and extended such that it improved translation quality when used in a machine translation system. The main modification consisted in departing from the simple Hidden Markov Models (HMMs) previously used for the feature prediction to more powerful Conditional Random Fields (CRFs), which however had tractability issues.

My contribution to Fraser et al. (2012) consisted in designing the stemmed representation and refining the stem markup, as well as the pipeline to generate inflected forms in the post-processing step. Furthermore, I compared the prediction accuracy of four separate HMMs and a single joint HMM to justify the use of four separate CRFs, as one single CRF predicting all features jointly would not have been tractable.

The inflection prediction system presented in this thesis is essentially equivalent to the system in Fraser et al. (2012), even though I modified some implementational details since, namely

- the handling of the features strong/weak which are predicted based on the predicted values of the other features, whereas they were treated as an independent feature in Fraser et al. (2012);
- an adjustment of the representation of out-of-vocabulary words in the language model (and other models such as *Operation Sequence Models*) to allow for a match with uncovered words transferred from the input sentence.

Other modifications have the main purpose of simplifying the inflection prediction process, in particular the handling of the inflectional features.

Combining external target-side features with source-side features to improve the prediction of case While the inflection prediction system works well, the prediction component is restricted to local context, which is problematic for the prediction of grammatical case. I combined projected source-side features with target-side subcategorizational information obtained from an external database to provide the case prediction model with structured information relevant for the case prediction (Weller et al., 2013b). While there was only a minor improvement in BLEU, a manual evaluation showed improvement in the inflection step. The usage of this combination of resources, and in particular external subcategorization knowledge, is novel in a machine translation scenario.

Modeling complement types The research in the first part of the thesis showed that subcategorized noun phrases (NPs) and prepositional phrases (PPs) need to be modeled simultaneously to account for potential structural changes between NPs and PPs. I showed that the integration of "placeholder prepositions" at the beginning of noun phrases to technically transform NPs into PPs reduces structural differences, which improved the translation quality.

Furthermore, I modeled complement types by means of contextually conditioned synthetic phrase-table entries containing prepositions, including a special "empty" preposition to represent NPs (Weller-Di Marco et al., 2016). These phrase-table entries aim at providing an optimal selection of phrase-translation pairs for the realization of complements with regard to their grammatical case and the choice of preposition, either as empty or overt prepositions. The generation of prepositions allows to create *new sequences* containing empty/overt prepositions.

While the use of synthetic phrases itself is not new, I applied this technique to the generation of function words (i.e. prepositions, including "empty" prepositions as heads of NPs) that carry semantic content and can be determined only on global information derived from source and target sentence. In contrast, previous approaches generated inflectional variants (Chahuneau et al., 2013), or determiners (Tsvetkov et al., 2013), i.e. semantically void function words refined to a local context.

A secondary contribution stemming from my work on modeling complement types is the insight that prepositions, despite being a closed-word class, are not always well-covered in a translation system: while prepositions are not under-represented in the same way as low-frequency words, their "meaning" or subcategorized use in a particular setting might be not well-attested in the training data.

Combining approaches to model syntax, lexical choice and morphology I combined established strategies to address problems across linguistic levels, namely source-side reordering for syntactic differences, the use of a discriminative classifier to improve decisions at the lexical level, and inflection prediction to model target-side morphology (Weller-Di Marco et al., 2017). While none of these approaches is new, their combination within one system is novel. The main findings of this set of experiments are (i) that the investigated strategies are complementary and that individual gains add up; and (ii) the fact that source-side reordering can have negative influence on the lexical level. The second outcome is particularly interesting when taking into consideration that reordering approaches are a popular strategy to handle long-distance structural differences.

1.3 Outline

Background Chapter 2 presents some background on the key components of machine translation systems, as well as on the widely used automatic evaluation metric BLEU.

Inflection prediction Chapter 3 presents a morphology-aware translation system that separates translation from the generation of target-side morphology. In a first step, a translation model is built on a simplified, underspecified representation of the target-side data that allows the model to generalize over inflectional variants during the training phrase. In a second step, the underspecified translation output is inflected: after predicting the relevant inflectional features of a word in its sentence context, fully inflected forms can be generated using a morphological resource.

Modeling grammatical case To model grammatical case, both source-side and target-side features need to be consulted: by looking at the source-side syntactic function, the role of a phrase can be realized accordingly in the translation. Taking into account target-side subcategorization information at the same time ensures that target-side requirements are met. In chapter 4, source-side features are combined with subcategorization information obtained from an external database with the objective to improve the prediction of case.

Using noun class information to model selectional preferences The translation of prepositions is a challenging problem in machine translation: prepositions are highly ambiguous, and the choice of prepositions depends on many factors. In chapter 5, noun class information is used to model selectional preferences of prepositions. The annotation of noun class information into the parse trees used to train a hierarchical machine translation system aims at obtaining more precise translation rules: by grouping training instances according to their annotation, the resulting translation rules do now allow for any PP to be applied in its context, but restrict the selection to a specific semantic class.

Target-side generation of prepositions In chapter 6, prepositions are treated as a target-side generation problem, which allows to also model structural differences between the source and target side: in a pre-processing step prior to training the SMT system, all NPs and PPs are transformed into a PP with underspecified preposition by substituting overt prepositions with placeholders, and by inserting placeholders for "empty" prepositions at the beginning of NPs. After translation, actual prepositions, including the empty prepositions to form NPs, are predicted on the translation output using rich source- and target-side features.

Modeling complement types with synthetic phrases Chapter 7 takes the idea from the previous chapter a step further: Using the placeholder representation, the prediction step is shifted from post-processing to an earlier stage: synthetic phrase-translation pairs containing generated prepositions conditioned mainly on rich source-side context are integrated into the translation model. These newly generated phrases provide an optimal selection of phrase-translation pairs for the given context.

Combining strategies to address problems across linguistic levels Chapter 8 investigates the combination of different strategies to handle structural differences, lexical choice and target-side morphology within one system.

Conclusion and future work Chapter 9 summarizes the main findings and results, and outlines interesting ideas for related future work.

Chapter 2

Background – SMT in a Nutshell

To train a phrase-based translation model, phrase translation probabilities and other statistics are derived from word-aligned parallel data. During translation, the input sentence is segmented into phrases that are then translated based on the statistics in the translation model. To ensure a good translation, the translation system combines statistics that take into account the relation between the source language and the target language, as well as a component that optimizes the quality of the translation hypotheses on the target side. The explanations below give a short overview of the functioning of the components required to train a Moses[1] translation system (Koehn et al. (2007b), Koehn (2010)).

2.1 Key components in a translation system

The key components in a translation system are a *phrase-table* that contains translation probabilities between source and target phrases, a *reordering-table* with statistics telling the system when and how to reorder the sequences of phrases, and a *target-side language model* to score the translation hypotheses. The individual parameters of these models are weighted, and optimal feature weights are estimated as part of the training process.

2.1.1 Phrase-table

A prerequisite for building a phrase-table is a parallel corpus aligned on word-level, as illustrated by the example in figure 2.1. The first step to build a phrase-table is the extraction of all source-target phrase pairs from the parallel corpus; table 2.1 shows examples for phrase pairs that can be extracted from the sentence in figure 2.1. The extracted phrases are not linguistically motivated, but their extraction is solely restricted

[1] `http://www.statmt.org/moses/`

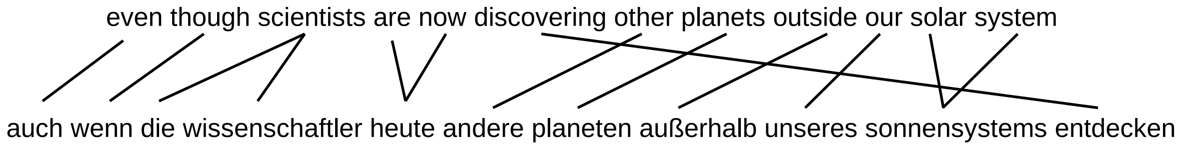
even though scientists are now discovering other planets outside our solar system
auch wenn die wissenschaftler heute andere planeten außerhalb unseres sonnensystems entdecken

FIGURE 2.1: Example for word alignment in parallel data

planets	planeten	outside our	außerhalb unseres
other planets	andere planeten	solar system	sonnensystems
planets outside	planeten außerhalb	our solar system	unseres sonnensystems

TABLE 2.1: Sample of source-target phrase pairs that can be extracted from the sentence in figure 2.1.

by the alignment structure and a defined maximum phrase length (often set to values between 5 and 7). As can be seen in the examples, phrases can be as short as one word, such as *planets – planeten*, as well as neighbouring phrases combined to blocks, such as *other planets – andere planeten*. The extracted phrase pairs must be contiguous; for example, it is impossible to extract the pair *discovering other – andere entdecken*, as the words *andere* and *entdecken* are separated. Furthermore, multi-alignments as in *solar system – sonnensystems* must be preserved; thus a phrase pair such as *solar – sonnensystems* cannot be extracted from this phrase.

After having extracted all phrase-translation pairs from the parallel data, the relevant translation features are computed. The phrase-table contains four scores: the translation probabilities $p(e|f)$ and $p(f|e)$, as well as the lexical weights $lex(e|f)$ and $lex(f|e)$, where f denotes the source language, and e denotes the target language.

The phrase translation probability $p(e|f)$ is calculated as the relative frequency of the target phrase e over all target phrases e_i that were observed as possible translations for the source phrase f:

$$p(e \mid f) = \frac{count(f,e)}{\sum_{e_i} count(f,e_i)}$$

The reverse translation probability $p(f|e)$ is calculated equivalently.

The lexical weights are an additional measure to estimate the reliability of a source-target phrase pair. Conceptually, the lexical weight is a back-off to combined single word statistics in comparison to estimating the translation probability of the entire phrase pair. The motivation to adding lexical weights to the phrase-table is due to the fact that some phrase-translation pairs, especially longer ones, are observed only once or very few times, which results in comparatively high translation probabilities.

source	**target**	**p(f|e)**	**lex(f|e)**	**p(e|f)**	**lex(e|f)**
solar system	sonnensystem	0.6923	0.1437	0.2706	0.0067
solar system	sonnensystems	0.3717	0.0930	0.2180	0.0052
solar system	solarsystem	0.6	0.045	0.0225	0.0008
solar system	solarsysteme	0.1666	0.048	0.0150	0.0014
solar system	planetensystem	0.0666	0.0125	0.0075	0.0001
solar system	sonnensystem zu	1	0.0718	0.0075	9.74e-06
solar system	sonnensystemkörper	0.1428	0.0003	0.0075	0.0005
solar system	sind	2.80e-06	1.92e-09	0.0075	0.0003

TABLE 2.2: Example entries from a phrase-table of an English–German translation system for the source phrase *solar system*.

Estimating the lexical plausibility in form of individual, word-per-word probabilities is a valuable indicator for the system to decide whether to trust such phrases. The lexical weight $lex(e|f,a)$ is computed as follows (a describes the alignment within the phrase):

$$lex(e \mid f,a) = \prod_{i=1}^{length(e)} \frac{1}{\mid \{j \mid (i,j) \in a\} \mid} \sum_{\forall (i,j) \in a} w(e_i|f_j)$$

The reverse lexical weight $lex(f \mid e,a)$ is calculated equivalently.

Phrase-table entries typically do not give only one or two translation options as a dictionary might, but typically offer a wide range of synonyms, near synonyms, related words and variations thereof, and, unfortunately, also unrelated translation candidates. This is illustrated by the phrase-table entries in table 2.2: the first five entries are valid translations of the English source phrase *solar system*, including the translation option *sonnensystems* as observed in figure 2.1. The translation option *sonnensystem zu* is obviously a related translation, but does not exactly correspond to the source phrase – nevertheless, the added *zu* ('to') might be useful for the system in some contexts. It is not unusual that "extra" words, often function words and punctuation, are found in either the source or the target phrase. Similarly, the translation option *sonnensystemkörper* ('solar system body') is not an exact translation of the source phrase, but is due to either incorrect alignment (missing the equivalent for 'body'), or an imprecise translation. The last translation option, *sind* ('are') is just wrong, which is also visible in its low scores.

The phrase extraction routine does not consider the context of a phrase or any restrictions besides the non-linguistically motivated alignment structure. Thus, the table contains German inflectional variants of the same base form, e.g. *sonnensystem*$_{Nom,Acc,Dat}$

– *sonnensystems*$_{Gen}$, and *solarsystem*$_{Sg}$ – *solarsysteme*$_{Pl}$. In addition to hindering generalization, the model has to select one of the forms at translation time: the genitive form, for example, is a valid option in many contexts (including in the sentence in figure 2.1), but is of course not always correct. The issue of asymmetry between source and target-side, as well as the selection of inflected target-side forms, in particular with regard to case, will be a relevant topic in this thesis.

Looking again at the phrase-table entries, it is important to keep in mind that the translation process is often not "exact", but that the sentence is often segmented into non-intuitive phrases, that are often translated by non-exact matches. However, as languages are different and many constructions are not translatable in an isomorphic way, the richness of the translation options can provide useful variants that lead to good translations, but it also leads to noise in the translation model. One goal of the thesis is to minimize differences between phrase pairs, and to systematically generate context-appropriate variants that take into account particular types of structural differences.

2.1.2 Reordering-table

The order of the phrases in the source sentence is often different from the expected order in the target language. Thus, the translation model needs to be able to reorder phrases during translation. A typical example for reordering is the position of French adjectives, which usually are to the right of a noun. When translating French into German or English, adjectives need to be reordered to be positioned before the noun. Similarly, verbs in German are systematically positioned differently than verbs in English, cf. figure 2.1. The reordering model learns statistics to predict potential reordering movements for the phrases listed in the phrase-table. In practice, however, it is limited to local reordering movements over small distances, as the costs for longer reordering movements are too high for the system. This means that reordering over short distances, such as switching the position of an adjective, is well handled by the reordering model, but reordering over long distances, such as moving a verb from the end to the beginning of a clause, is often impossible.

2.1.3 Language model

For a good translation, the translation model needs to generate a sentence that (i) conveys the same meaning as the source sentence, and (ii) is well-formed and respects the grammatical requirements of the target-language. Reproducing the source content in the target language is addressed by the bilingual statistics in the phrase-table. A target-side language model then judges the quality of a generated target-language sentence; this task is monolingual and does not take into consideration whether the sentence is a good translation, but only whether it is a good sentence of the target language.

Language models are trained on large monolingual corpora by learning a probability distribution over sequences of words. The order of a language model, for example $n = 5$, indicates the size of the word n-grams that are considered in the language model. To score a translation candidate, the language model estimates the probability of observing the sentence. For a language model of order n and a sentence of length m, this can be approximated as follows:

$$P(w_1, ..., w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, ..., w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, ..., w_{i-1})$$

The probability to observe a word w_i in the context of the preceding words $w_1 ... w_{i-1}$ is approximated by the shorter context $w_{i-(n-1)...w_{i-1}}$, where n corresponds to the considered window. The probability $P(w_i \mid w_{i-(n-1)}, ..., w_{i-1})$ can be calculated as relative frequency of observing the n-grams $w_{i-(n-1)}, ..., w_i$ and $w_{i-(n-1)}, ..., w_{i-1}$.

As unseen n-grams lead to probabilities of zero, which is unpractical when applying the model, language models usually make use of smoothing algorithms to reserve a small amount of probability mass for unseen n-grams.

2.1.4 Parameter tuning

The features presented above are combined in a log-linear model of the following form:

$$P(x) = argmax_x \; exp \sum_{i=1}^{n} \lambda_i h_i(x)$$

where n is the number of feature functions h_i with the respective weight λ_i. The variable x denotes the source sentence and the target sentence and its segmentation into phrases. The modeled features h contain the probabilities in phrase-table and reordering-table, as well as the language model score; further features can be added as desired.

In the training phrase of the system, optimal weights are determined by minimizing a standard error metric (e.g. BLEU) on a development set (MERT: Minimum Error Rate

Training). The weights λ_i are modified such that they maximize the BLEU score (i.e. minimizing errors) when repeatedly translating the development set until convergence.

2.2 Tree-based translation models

Tree-based models are not applied to flat text as the phrase-based model explained above, but assume a tree spanning over the sentence to extract translation rules. Such a tree can, but does not necessarily need to be, a syntactic tree as output by a linguistic parser. There are several variants of tree-based models[2]:

- hierarchical phrase-based: no linguistic syntax
- string-to-tree: linguistic syntax only in the target language
- tree-to-string: linguistic syntax only in the source language
- tree-to-tree: linguistic syntax in both languages

In tree-based translation models, translation rules can have gaps, i.e. non-terminal symbols that can be filled with other rules, as illustrated by the following example where X is a generic, non-linguistically motivated non-terminal:

(1) ate X → hat X gegessen
cake → kuchen

In a system using linguistic information, non-terminals such as NP or VP are used instead of the generic X non-terminal, resulting in a translation rule as in (2).

(2) ate NP → hat NP gegessen

Some of the experiments carried out in this thesis are based on a *string-to-tree model*, i.e. using linguistic parse trees on the target side, and generic non-terminals on the source side. While tree-to-tree models make most use of linguistic information, syntax trees on both sides also represent a certain restriction: for example, the extraction of source-target phrase pairs prior to building the phrase-table is restricted to phrase pairs that meet the alignment conditions while at the same time being valid sub-trees in the parsing structure. Being mainly interested in modeling target-side phenomena, using syntactic trees on the target side in a string-to-tree setting is a good compromise.

[2] Definitions as found on `http://www.statmt.org/moses/?n=Moses.SyntaxTutorial`

source	[X] though [X] are now discovering other planets [X] [X]
target	[ADV] wenn [NP] heute andere Planeten [PP] entdecken [S]
alignment	0-0 1-1 2-2 3-3 4-3 6-4 7-5 8-6 5-7
source	even though scientists [X] discovering [X] [X]
target	auch wenn die Wissenschaftler [ADV] [NP] entdecken [S]
alignment	0-0 1-1 2-2 2-3 3-4 5-5 4-6
source	[X] though [X] [X] discovering [X] [X]
target	[ADV] wenn [NP] [ADV] [NP] entdecken [S]
alignment	0-0 1-1 2-2 3-3 5-4 4-5
source	[X] though [X] are now [X] [X] [X]
target	[ADV] wenn [NP] heute [NP] [VVFIN] [S]
alignment	0-0 1-1 2-2 3-3 4-3 6-4 5-5

TABLE 2.3: Some string-to-tree translation rules extracted from the sentence in table 2.1: while the source side uses generic "X" nodes as non-terminals, the target side uses the labels as provided by the parser.

Table 2.3 shows examples for string-to-tree translation rules extracted from the sentence in figure 2.1. The non-terminals on the very right ([X] for the source, [S] for the target side) denote the respective left side of the phrase ($S \rightarrow \dots$). Finally, the alignment indicates how the words and non-terminals on the source side and the target side are related. For example, the last non-terminal [X] on the source side in the first entry (at position 8) corresponds to the [PP] at position 6 in the target phrase. The example illustrates how tree-based translation systems can capture large "gaps" – the alignment structure in figure 2.1 makes the extraction of flat phrases containing the verb impossible; besides the pair *discovering – entdecken*, only the entire chunk can be extracted, but a phrase of that length is unlikely to be of much use. A tree-based approach can effectively reduce the gap by representing a lengthy NP or PP by a non-terminal. This does not only allow the model to learn more general (syntactic) translation patterns, but also to handle long-distance reordering at decoding time.

2.3 Automatic evaluation metric: BLEU

When evaluating the output of a machine translation system, the two main criteria are *adequacy* and *fluency*. These criteria are motivated by the idea that a good translation should convey the same meaning as the source sentence (adequacy), while at the same time being a good sentence of the target language in terms of word choice and grammatical correctness (fluency).

Ideally, each system should be rated and evaluated by a competent human annotator – of course, this is unfeasible. In addition to being hugely unpractical, the evaluation of machine translation output is also difficult even for human annotators; furthermore, it is not guaranteed that two annotators (or even the same annotator after a certain amount of time) would agree in their annotation. Instead, it is more practical to evaluate the output of a machine translation system with an automatic metric that is cheap and easy to use, as well as reproducible. A widely used automatic evaluation metric is BLEU (**B**i**L**ingual **E**valuation **U**nderstudy, Papineni et al. (2002)) that compares the translation output with one or more reference translations. The comparison between translation and reference sentence is carried out in terms of *n-gram precision* for values of n up to a predefined length. As translation systems have a tendency to over-generate high frequency stop words (such as articles), BLEU uses a *modified unigram precision* that takes into account the actual number of words that are matched, and thus avoids to credit over-generated words. Example (3) (taken from Papineni et al. (2002)) illustrates the concept.

(3) MT the the the the the the the

REF the cat is on the mat

The translation output in (3) is obviously bad, but would still obtain a unigram precision of $\frac{7}{7}$, as all words in the translation also occur in the reference. In the modified unigram precision, only two occurrences of *the* are credited ($\frac{2}{7}$), corresponding to the number of occurrences in the reference. As the n-gram precision favours short sentences, BLEU uses a *brevity penalty* that discounts the score for translations that are shorter than the reference sentence. BLEU is computed as follows:

$$\text{BLEU}_n = \text{brevity-penalty} \cdot exp \sum_{i=1}^{n} \lambda_i \cdot log(\text{precision}_i)$$

$$\text{brevity-penalty} = \begin{cases} 1 & \text{if } |c| > |r| \\ e^{1-\frac{r}{c}} & \text{if } |c| \leq |r| \end{cases}$$

where λ_i are weights for the n-gram precision for different values of n, $|c|$ is the length of the translation candidate, and $|r|$ is the length of the reference translation.

BLEU ranges from 0 (no match) to 1 (exact match with the reference sentence), and is typically computed on document-level instead of sentence-wise. The calculated

BLEU score is specific for the document; and as a BLEU score itself is not particular meaningful, BLEU is rather used to compare the performance of two systems.

BLEU has the advantage of being easy and quick to compute, and is independent from additional resources. This makes it a widely used evaluation metric, as well as an often used optimization criterion in parameter tuning (MERT). However, it also has some shortcomings: for example, it strongly depends on the reference translation(s), and while a match between the candidate translation and the reference is likely to be an indicator for the translation quality, a candidate translation that does not match with the reference is not necessarily a bad translation, as there are often many good translations. In particular, BLEU does not recognize synonyms and does not give credit to "near matches" such as *decide – (make a) decision*.

Furthermore, BLEU is a precision-oriented measure and less focused on measuring adequacy; it is also not well suited to measure small (but meaningful) changes between translation systems that are manifested only in few words, such as modifying the inflection of a phrase, which often can be done by changing only one word.

Of course, BLEU is not the only evaluation metric – there are countless variants and other approaches to automatically assess the quality of machine translation output that are beyond the scope of this thesis. For example, there are variants of BLEU itself, e.g. computed on lemmas or POS tags to generalize over inflectional variants or to estimate syntactic soundness (Popović et al., 2009), or a variant of BLEU allowing "fuzzy matches" to account for partial matches in compounding languages (Virpioja et al., 2015). On the other hand, METEOR (Agarwal et al., 2008) is designed to capture near matches and semantically closely related words by making use of ontology resources such as *WordNet* (Fellbaum, 1998). Looking at a larger picture, MEANT (Lo et al., 2011) applies semantic parsing to the translation output and compares the realization of semantic roles between the translation output and the reference.

As is common in SMT research, the translation experiments will be evaluated in BLEU. However, for many experiments in this thesis, BLEU is not an optimal evaluation metric; some results are thus additionally evaluated by a manual annotation, either in the form of ranking sentences or by measuring the accuracy of the modeled phenomenon.

Chapter 3

Inflection Prediction: Modeling Target-Side Inflectional Morphology

This chapter presents a morphology-aware English–German translation system that uses a two-step approach to separate translation from the generation of target-side morphology. First, a translation model is trained on a simplified, underspecified representation of the target-side data that allows the model to generalize over inflectional variants in the training phrase. Then, the underspecified translation output is inflected: after predicting the relevant inflectional features of a word in its sentence context, a morphological resource is used to generate fully inflected forms. This inflection prediction approach allows to address the main problems when translating into a morphologically complex language: (i) the underspecified representation allows for generalization at training time; (ii) inflectional features are predicted in the sentence context to ensure a correct inflection; and (iii) the use of a morphological resource enables the system to generate inflected forms independently from their occurrence in the parallel training data. In the presented experiments, the inflection prediction system improves over a baseline trained on standard surface form data.

The abstract representation used in the inflection prediction system is an essential prerequisite for the experiments in the subsequent chapters, as it enables a flexible modification of inflection. Such a functionality is necessary when modeling grammatical case and prepositions; for example, changing a preposition may entail a change in grammatical case that needs then to be reflected in the inflection of the entire PP. This chapter gives a detailed overview of the relevant inflectional features and their prediction in conditional random fields (CRFs) that will be extended in later chapters.

A first study on modeling morphology in a simulated translation system has already been presented in Weller (2009); a functioning variant of the inflection prediction technique in an English–German translation system is published in Fraser et al. (2012).

3.1 Motivation

Translating into a morphologically complex language is challenging for several reasons: during training, the different inflectional variants of one base form are not recognized as related, but are treated as different forms, resulting in a sub-optimally estimated translation model. When translating a sentence, the system has to select the correct inflection among all forms that have been seen as a possible translation of a source word. To make this problem even harder, it may be the case that the required form does not occur in the training data at all, making it impossible to generate a correct translation for a particular context, despite having lexical evidence for a translation.

Between English and German, there is an imbalance at the level of morphological complexity: German has a comparatively rich morphology, whereas English expresses only few features morphologically. In particular, German nominal morphology is more complex than English nominal morphology. German determiners, adjectives and nouns are inflected for the features *number*, *gender*, *case* and *strong/weak inflection* and have to agree[1] according to the respective features within an NP or PP, whereas in English, only nouns exhibit morphological marking, and only for the feature *number*. The left sentence in figure 3.1 illustrates agreement in German NPs and PPs (no explicit inflectional suffix means that the surface form is identical with the lemma, as in *bank*). It can be easily seen that nearly all eligible words in the German phrases are inflected, whereas on the English side, only the noun *interests* is explicitly marked for plural. It becomes clear from the example that the morphological features relevant for German, except for *number*, are not represented on the English side.
Furthermore, the structure of German NPs and PPs itself can be complex, as is illustrated in the right sentence in figure 3.1: the NP contains an inserted PP which makes the outer NP discontinuous; theoretically, there is no limit to the number of nested phrases and the depth of nesting.

In a machine translation scenario, such differences are problematic as they prevent the model from learning exact and generally applicable translation equivalents, but often lead to inexact translation pairs, or pairs that are bound to specific contexts. Consider, for example, the translation pair *the blue car* → *der blaue Wagen*: while this is a valid translation pair, its German side is restricted to NPs in subject positions, as the

[1] Technically, there is no agreement for the feature *strong/weak*, but the particular values of the three other features as well as some other conditions such as the determiner require a certain order of *strong/weak* inflection within the phrase; cf. section 3.2.2 for more details.

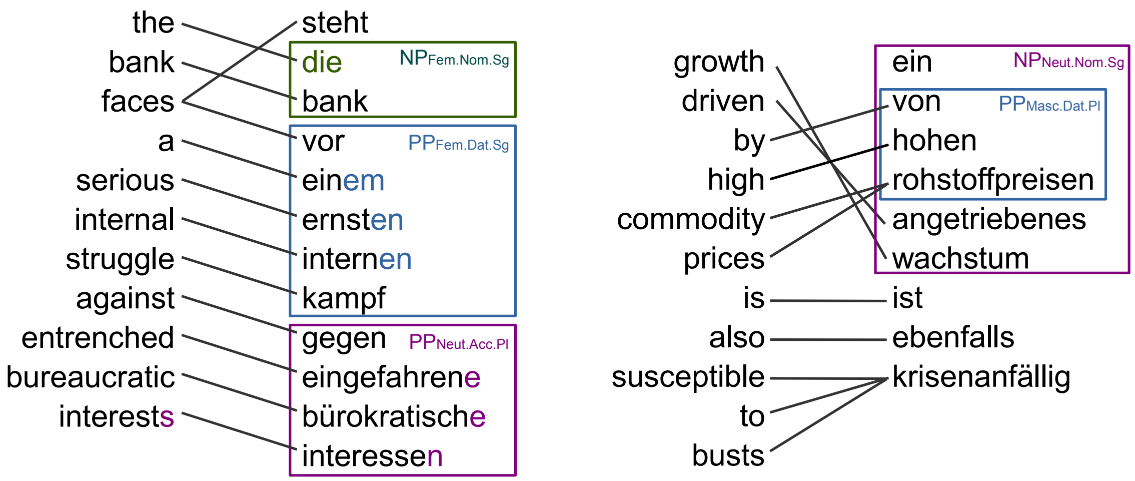

FIGURE 3.1: Examples for German noun phrases (NPs) and prepositional phrases (PPs) illustrating agreement in the morphological features *gender*, *number* and *grammatical case*; inflectional suffixes are colored. The sentence on the right side shows a complex NP with an inserted PP.

corresponding variants for object positions should be *den blauen Wagen* (direct object) or *dem blauen Wagen* (indirect object).

During the translation process, rich target-side morphology leads to two problems: first, the system has to select the correctly inflected form, even though the relevant clues are not necessarily contained in the source language (*target-side generation*). Second, it is not guaranteed that the required form does occur in the parallel training data (*data sparsity*). This is illustrated by the example in table 3.1 which lists all occurrences of the lemma *neokeynesianisch* ("neo-keynesian") in 4.5 million parallel sentences: while there are 7 attested forms, the inflection paradigm is not complete; a translation requiring for example the form *neokeynesianisches*, as in *a neo-keynesian growth model* → *ein neokeynesianisches wachstumsmodell*, is impossible. Furthermore, the model learns separate statistics for each form of *neokeynesianisch*, even though they have the same lexical meaning.

3.1.1 Modeling target-side morphology

The problems described above are to a large extent due to the fact that target-side inflectional processes are commingled with the lexical translation process in a way that does not benefit the translation system as it contains useless information at the

der neokeynesianische ansatz	*the neo-keynesian approach*
und neokeynesianischer begeisterung	*and neo-keynesian enthusiasm*
durchsetzenden neokeynesianischen ansatzes	*prevailing neo-keynesian approach*
den neokeynesianischen empfehlungen	*the neo-keynesian recommendations*
der neokeynesianische ansatz	*the neo-keynesian approach*
eine neokeynesianische politik	*a neo-keynesian policy*
der neokeynesianischen nachfragesteuerung	*of neo-keynesian demand management*
ein neokeynesianisches wachstumsmodell	*a neo-keynesian growth model*
nach neokeynesianischem muster	*accordingto neo-keynesian pattern*

TABLE 3.1: Example for *data sparsity*: the upper part of the table shows trigrams of all 7 occurrences of the lemma *neokeynesianisch* in a corpus of 4.5 million parallel sentences. The lower part shows unobserved inflected forms (*neokeynesianisches, neokeynesianischem*) in a possible context.

cost of generalization. A solution to better handle this situation is the separation of the *translation process* from *target-side inflection* in a morphology-aware translation system that works in two steps: first, a translation system is built to translate into an underspecified representation of German (“*stemmed representation*”) with relevant inflectional markup. In the second step, the output of this MT system (i.e. non-inflected word *stems*) is inflected in order to obtain a sequence of fully inflected surface forms.

In the underspecified representation, morphological features that are not relevant in the translation step are removed, whereas translation-relevant features (such as the number of nouns, which is also present on the English side) are kept in an abstract representation. In the inflection step, a careful combination of annotated features preserved through translation and features predicted on the translation output allows to determine the values of all inflection-relevant features. Based on the stems and the set of feature values, a morphological resource can be used to generate inflected forms.

Modeling target-side morphology in statistical machine translation has several advantages that will be explained and explored within this thesis:

- By introducing an abstract representation, fully inflected forms are reduced to their stems or lemmas, and in some cases annotated with translation-relevant features. This removes unnecessary complexity, as for example, all adjective forms of an inflection paradigm are mapped onto one form:
  {*blau, blaue, blaues, blauer, blauen, blauem*} → `blau<+ADJ><Pos>[ADJ]`.
  Simplified this way, the resulting models are more general as the amount of forms has been reduced, and only information relevant to the translation process is kept.

- The approach provides a direct solution to the problem of selecting the correct inflection and data sparsity: The target-side forms are inflected depending on the respective contextual requirements, and as the forms are generated using a morphological resource, the inflected forms are not restricted to those observed in the parallel training data, but can be generated as needed. The ability to generate new forms is particularly important when dealing with under-resourced domains, such as e.g. translating medical or technical data sets.

- In addition to the more straightforward benefits of reducing the model complexity and generating surface forms as required in the sentence, the abstract representation of the target-side data holds many possibilities for a focused modeling of inflectional features: First, source-side and target-side features can be separated if they are independent from each other; thus, target-side features can be modeled separately more efficiently. Second, the abstract target-language representation allows to integrate different types of externally obtained (linguistic) information, in particular with regard to grammatical case, which will be explored in the later chapters of this thesis.

3.1.2 Overview of the training and translation procedure

Figure 3.2 outlines the process of training the translation system and prediction model, and the generation of inflected forms for the translation output.

The training (upper part in figure 3.2) starts with pre-processing the German data in order to obtain the stemmed representation with feature markup. To this end, the data is parsed and the surface forms are analyzed with the morphological tool SMOR (Schmid et al. (2004), cf. section 3.2.1), yielding a *stem* for each word. Based on the stems output by the morphological resource SMOR and the morphological features annotated by the parser, the underspecified representation with feature markup is generated. This step is applied to all used target-side data, i.e. the German part of parallel data and the language model data. The English data does not undergo any pre-processing except for standard procedures such as tokenizing. Then, a standard Moses system translating from English surface forms to German stems is built[2]. Furthermore, a sequence model to

[2] There are no restrictions concerning the type of Moses system, i.e. the phrase-based and hierarchical variants of Moses are both possible; preliminary experiments indicated that the inflection prediction process can even be applied in neural machine translation (NMT).

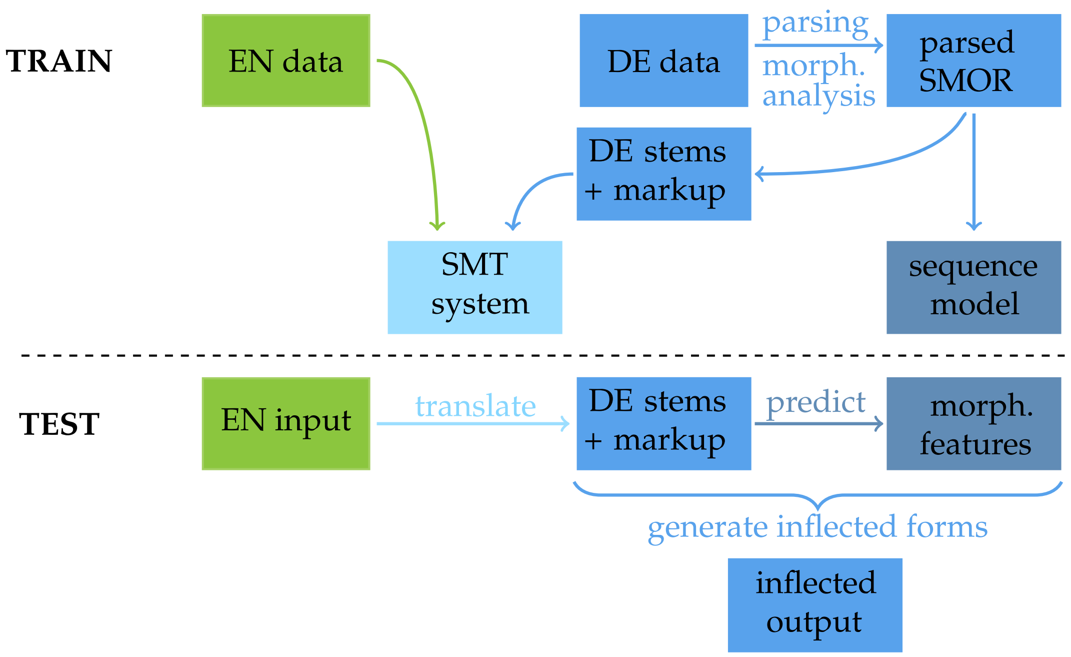

FIGURE 3.2: Illustration of the inflection prediction process: the upper part pictures the training of the SMT system and the models for feature prediction on parsed and morphologically analyzed data. The lower part shows the re-inflection step based on stems and the set of predicted values for the inflection-relevant morphological features.

predict the full set of morphological features for the underspecified machine translation output is trained based on the word stems and the respective inflectional features.

When translating (lower part in figure 3.2), the English input is translated into the stemmed underspecified representation. Using features specified in the markup as input to the prediction step, all values of the inflectional features are determined with the sequence model. Based on the full set of inflectional features and word stems, inflected surface forms are generated with SMOR.

3.2 Modeling inflectional features

In this section, the modeling of the underspecified abstract target-side representation is outlined. After a brief introduction of the linguistic resources, the creation of the stemmed representation is explained in more detail.

```
(TOP
  (S-TOP
    ...
    (VVFIN-HD-Sg steht)                          VVFIN-steht
    (NP-SB\/Sg
      (ART-HD-Nom.Sg.Fem die)                    ARTdef-Nom.Sg.Fem.St
      (NN-HD-Nom.Sg.Fem Bank)                    NN-Nom.Sg.Fem.Wk
      (PP-MNR\/N
        (APPR-AC\/Dat vor)                       APPR-vor-Dat
        (ART-HD-Dat.Sg.Masc einem)               ARTindef-Dat.Sg.Masc.St
        (ADJA-HD-Sup.Dat.Sg.Masc ernsten)        ADJA-Dat.Sg.Masc.Wk
        (ADJA-HD-Pos.Dat.Sg.Masc internen)       ADJA-Dat.Sg.Masc.Wk
        (NN-HD-Dat.Sg.Masc Kampf)                NN-Dat.Sg.Masc.Wk
      )
    )
    (PP-OP\/V
      (APPR-AC\/Acc gegen)                       APPR-gegen-Acc
      (ADJA-HD-Pos.Acc.Pl.Neut eingefahrene)     ADJA-Acc.Pl.Neut.St
      (ADJA-HD-Pos.Acc.Pl.Neut bürokratische)    ADJA-Acc.Pl.Neut.St
      (NN-HD-Acc.Pl.Neut Interessen)             NN-Acc.Pl.Neut.St
    )
    ...
  )
)
```

FIGURE 3.3: This example illustrates the output of BitPar (left side) and the features extracted for the stemmed representation and the training data for the sequence model to predict morphological features for generating inflected forms.

3.2.1 Linguistic resources

The key linguistic knowledge sources used for the inflection prediction process are the constituency parser BitPar (Schmid, 2004), (Schmid, 2006) to annotate the relevant inflectional features, and the morphological tool SMOR (Schmid et al., 2004) to analyze and generate inflected German surface forms.

BitPar is a parser for highly ambiguous probabilistic context-free grammars. Figure 3.3 shows a fragment of BitPar output: brackets indicate the structure of the sentence, and the features *grammatical case, number* and *gender* are annotated onto noun phrases and prepositional phrases. For preparing the stemmed representation, a flat structure, as shown on the right side of figure 3.3, is sufficient – for this tag sequence, the morphological annotation (part-of-speech (POS) tag with inflectional features) are extracted from the parse output. For training the inflection prediction sequence models, the feature *strong/weak inflection* is necessary. As this feature is not part of the parse output, it is additionally annotated following the rules outlined in section 3.2.2.

SMOR is a morphological tool for German inflection and word formation implemented in finite-state technology. It relies on a lexicon, and also covers productive word formation processes such as compounding or derivation using a concatenative approach. SMOR can be applied in two directions:

- analysis: *surface form* → *stem+features*
- generation: *stem+features* → *surface form*

Analysis The morphological analyses contain the inflectional features *number, gender, strong/weak* and *grammatical case*, as illustrated in (1) and (2). Additionally, morphologically complex words are decomposed in sequences of morphemes[3], as in (2). The feature marked with "+" (`<+NN>` or `<+ADJ>` in the examples) indicates the word class of the analyzed word.

(1)

```
analyze> Interessen
Interesse<+NN><Neut><Acc><Pl>
Interesse<+NN><Neut><Dat><Pl>
Interesse<+NN><Neut><Gen><Pl>
Interesse<+NN><Neut><Nom><Pl>
```

(2)

```
analyze> eingefahrene
ein<VPART>fahren<V><PPast><SUFF><+ADJ><Pos><Neut><Acc><Sg><Wk>
ein<VPART>fahren<V><PPast><SUFF><+ADJ><Pos><Neut><Nom><Sg><Wk>
ein<VPART>fahren<V><PPast><SUFF><+ADJ><Pos><Masc><Nom><Sg><Wk>
ein<VPART>fahren<V><PPast><SUFF><+ADJ><Pos><NoGend><Acc><Pl><St>
ein<VPART>fahren<V><PPast><SUFF><+ADJ><Pos><NoGend><Nom><Pl><St>
ein<VPART>fahren<V><PPast><SUFF><+ADJ><Pos><Fem><Acc><Sg>
ein<VPART>fahren<V><PPast><SUFF><+ADJ><Pos><Fem><Nom><Sg>
```

The feature *strong/weak inflection* is only part of the analysis if there are two distinct word forms for the feature values *strong* and *weak* – hence, this feature is indicated in the analysis of *eingefahrene*[4], whereas it is not needed for the analysis of *Interessen*. Furthermore, SMOR uses the feature value `NoGend` in contexts where the respective

[3] In example (2): `VPART`: verb particle; `PPast`: past participle; `SUFF`: suffix

[4] The respective form with weak inflection is *eingefahrenen*, for example when used with a definite article "*gegen die eingefahrenen bürokratischen Interessen*" ('against the entrenched bureaucratic interests'), which typically entails a weak context for the subsequent words in the phrase.

word forms for all values of gender (*masculine, feminine and neuter*) are the same. This is often the case for adjectives in plural.

The analyses listed above also illustrate that SMOR returns the complete set of all possible analyses: for example, the word *Interessen* is always analyzed with the feature `<Pl>`, but the values for *grammatical case* vary (`<(Acc|Dat|Gen|Nom)>`). Thus, when analyzing the surface forms, the morphological features from the parse output are consulted in order to disambiguate the set of possible SMOR analyses: the bold-faced analyses correspond to the context in figure 3.3. The disambiguation of feature values is important for those features that are part of the stem markup, namely number and gender of nouns. While it is relatively rare that a word has two analyses differing in gender (such as *Schild*$_{Neut}$: 'sign' and *Schild*$_{Masc}$: 'shield'), multiple analyses with different values for number (e.g. *Minister*$_{Sg/Pl}$: 'minister/ministers') are rather frequent.

Generation Used in the reverse direction, SMOR generates inflected word forms given a valid stem with the complete set of morphological features, as illustrated in the examples (3) and (4).

(3)
```
generate> ein<VPART>fahren<V><PPast><SUFF><+ADJ><Pos>
          <NoGend><Acc><Pl><Wk>
eingefahrenen
```

(4)
```
generate> Neo<KSF>keynesianisch<+ADJ><Pos><Neut><Dat><Sg><St>
neokeynesianischem
```

This enables to generate forms as required by the target-side context, e.g. the form *eingefahrenen* with weak inflection (as opposed to strong inflection as in figure 3.3), or the form *neokeynesianischem* (which does not occur in the German part of a large parallel corpus, cf. table 3.1). The generation of inflected forms relies on a word stem and the set of feature values, and is thus independent of word forms occurring in the parallel training data. This means that for a word to be translated, it is sufficient to have at least one form occurring on the target-side. Reduced to its stem, it is then used as a basis to generate the required inflected form, including forms not occurring in the parallel training data.

Over-generation While SMOR is a high-quality morphological resource, it is not perfect due to the complexity of natural language. One typical problem is that for morphologically complex words, SMOR tends to over-generate, as illustrated by the following example:

(5)
```
generate> Klima<NN>Wandel<+NN><Masc><Acc><Sg>
Klimaswandel
Klimatawandel
Klimawandel
```

For compound words, such as the word *Klimawandel* ('climate change'), there is a set of potential *Fugenelemente* (transitional elements), which all lead to paths in the finite-state implementation of SMOR, even though usually only one *Fugenelement* is correct for a given compound. For the stem `Klima<NN>Wandel<+NN>`, the first two variants, *Klima**s**wandel* and *Klima**ta**wandel*, are incorrect, whereas the third form, *Klima**∅**wandel*, is correct. In case of several surface forms output in the generation step, word frequency heuristics are used to filter out incorrect forms.

Generally, SMOR generates the set of all potentially possible forms using all transitional elements defined for a noun. However, some compounds are already lexicalized, leading to a "simple" analysis. For example, the noun *Tageszeitung* ('daily newspaper') can be analyzed as either `Tageszeitung<+NN><Fem>` (lexicalized) or `Tag<NN>Zeitung<+NN><Fem>` (compound). Generating surface forms based on the complex stem can lead to over-generation (*Tag**es**zeitung, Tag**e**zeitung, Tag**s**zeitung, Tag**∅**zeitung*), whereas the simple stem only produces the correct form. For the stemmed representation, stems with the least complexity are thus preferred.

3.2.2 Morphological features for nominal inflection

For German nominal inflection, there are four relevant features: *number*, *gender*, *grammatical case* and *strong/weak inflection*. In a noun phrase or prepositional phrase, typically all elements (such as nouns, articles, adjectives, possessive pronouns) are inflected for these features and have to agree within the phrase. While some of these features are also expressed in English, most have no equivalent or are expressed to a lesser degree. The general idea of the underspecified stemmed representation is to remove all features that are not immediately relevant in the translation step, either because they are not represented on the English side, or because they are subject to target-side conditions only. In the stem markup, all inflectional features that are dependent on the source side or that cannot be predicted based on (target-side) context are annotated. These features serve as input for the feature prediction model in the post-processing step. The four inflectional features for German nominal inflection are listed below:

Number The number feature can take the values *singular* and *plural*. In English, number is typically also expressed for nouns, but not for adjectives, articles or possessive pronouns. The number values of a noun phrase or prepositional phrase usually correspond between English and German.[5] This feature is thus generally shared between English and German, even though it is only explicitly expressed on English nouns. As a shared feature, it needs to be reflected in the stem markup in order to guarantee that the number feature is preserved during translation.

Gender The gender feature can take the values *masculine*, *feminine* and *neuter*. The gender of a noun can be considered as lexicalized, even though there are some patterns that allow to derive a noun's gender, for example words ending with the suffix *-ung* are always feminine. In contrast, English does not exhibit the feature gender, and save for a few exceptions such as *actor*$_{Masc}$ – *actress*$_{Fem}$, there is no overt gender morpheme. Gender is thus a feature that is determined only by the target-side. For word stems where the gender value is innate, this feature is also reflected in the stem-markup.

Grammatical case The case of a noun phrase depends on its function in the sentence and the subcategorization frame of the verb. The case feature can take the values *nominative*, *accusative*, *dative* and *genitive*. Typically, *nominative* indicates the subject, *accusative* a direct object and *dative* an indirect object. *Genitive* rarely occurs in subcategorized contexts, but mostly in the role of a modifying phrase corresponding to English *of-phrases* (as in *Unterzeichnung des*$_{Gen}$ *Vertrags*$_{Gen}$: 'signing of the treaty').

For prepositional phrases, the case is determined by the preposition. While some prepositions subcategorize only one grammatical case (such as *für+Acc*: 'for'), others allow for two values of case, which is often tied to a a difference in meaning[6]. Many prepositions (e.g. *in* ('in'), *zwischen* ('between') or *unter* ('under')), can alternate between *accusative* and *dative*, where *accusative* typically expresses a directional meaning, whereas *dative* stands for a locational meaning. Thus, in *in*$_{Acc}$ *den See springen* ('to jump

[5] There are some lexicalized exceptions, such as *Möbel*$_{Pl}$ – *furniture*$_{Sg}$ or *Polizei*$_{Sg}$ – *police*$_{Pl}$.

[6] Some prepositions also can occur with two values of cases without leading to a change in meaning. For example, the preposition *wegen* ('because of') typically subcategorizes the genitive case, but the use of dative case is also accepted.

in(to)$_{directional}$ the lake') the use of accusative indicates a directional meaning[7], whereas in *im*$_{Dat}$ *See schwimmen* ('to swim in$_{locational}$ the lake'), the meaning is locational.

Grammatical case is (nearly) not expressed on the English side; it is mostly subject to target-side requirements. In contrast to the two previous features, which are innate to the stem (gender) or determined by the source-side (number), case rather depends on the role of a phrase in the sentence. Thus, it generally cannot be "pre-determined" via stem markup, but has to be predicted for the resulting, translated sentence. While the source side generally can give clues about the role of a phrase in the target language, the target-side requirements are often more direct. For example, the choice of a verb (and its subcategorization frame with specific case requirements) during translation, or the decision to translate either in active voice (with a direct object in accusative case) or in passive voice (with a subject in nominative case) have an immediate influence on the setting of grammatical case. As will become clear in the following sections, the prediction of grammatical case is one of the most difficult parts in the inflection prediction approach; chapter 4 will be entirely dedicated to this very problem.

Strong/weak inflection Nominal inflection is not only determined by the three previously presented features, but also depends on the particular setting of definite/indefinite determiner and number/gender/case in the respective phrase. The general principle is that the elements in a phrase adopt a *weak inflection* if the features number, gender and case are already unequivocally expressed by a determiner. If this is not the case, *strong inflection* is required. Typically, a definite article (*der/die/das/...*: 'the') entails a weak inflection pattern, whereas no article entails strong inflection. In the case of an indefinite article (*ein/eine/...*: 'a(n)'), some quantifying articles (e.g. *kein*: 'no') or possessive articles, the phrase is inflected according to a *mixed inflection pattern*: The settings *Sg/Masc/Nom*, *Sg/Neut/Nom* and *Sg/Neut/Acc* entail strong inflection, whereas all other settings adopt weak inflection. Two adjacent adjectives are typically inflected according to the same inflection type[8]. Unlike the previous features, the strong/weak inflection feature does not, strictly speaking, exhibit agreement within the phrase in the

[7] The difference between a locational and directional meaning can be expressed in English by using the prepositions *into/onto* which convey a directional reading, as opposed to *in/on*. However, they are often used interchangeably: for example, the phrase "jump in the lake" obtained 2.87 million search results in Google, whereas "jump into the lake" got 329.000 results. A similar outcome was found for the phrases "get in the car" (7.64 million results) and "get into the car" (1.5 million results).

[8] There are several exceptions to this rule, as well as settings in which multiple inflection variants are acceptable. The website `www.canoo.net/services/OnlineGrammar/Wort/Adjektiv/Deklinationstyp` presents a comprehensive overview, including lists of determiners entailing either strong or weak inflection.

sense that all elements adopt the same value, but rather follows an "inflection pattern" where strong inflection can be followed by weak inflection. This feature can be regarded as a merely "technical feature" that is semantically vacuous. It is not present on the English side, and it is not considered in the stem-markup. In the prediction step, it can be determined based on the type of determiner used in the phrase and the values of the other three features. In many cases, the strong/weak feature is redundant, as the respective inflected forms have the same surface form.

3.2.3 Underspecified stemmed representation and feature markup

The stemmed representation used for translation has two main objectives: first, it aims at being as general as possible by not containing information that is not relevant to the translation step. A second goal is the preservation of all necessary information in an abstract form in order to ensure that translation-relevant information, such as number of English noun phrases, is preserved during the translation step. The idea of the stem-markup is to "set" those feature values that are already known (and not variable in the sense that they depend on particular target-side constellations) at translation time, and to then use them as input to the feature prediction step.

The stem-markup is based on word classes: for example, features that are "innate" to a particular word class, e.g. gender for nouns, are annotated as stem-markup. In contrast, there is no innate gender for adjectives or determiners, as they take their gender value from the noun that they modify – thus, receiving their feature values from their respective context, they cannot receive stem markup before translation.

In the following, the type of stem markup is listed for each word class:

Nouns The stems of nouns are marked with gender and number. Gender is considered as part of the stem and is obtained from the SMOR analysis. The feature number is determined from English nouns: when training the translation system, the word alignment typically links noun phrases and prepositional phrases with corresponding number, and consequently, the phrase-table can be expected to contain mostly entries where the number feature is preserved. This allows the system to correctly pass on the number feature during translation.

Adjectives The inflection of adjectives depends entirely on the noun that that they modify. Thus, no stem-markup is applied to adjectives.

Articles Like adjectives, the inflection of articles complies with the noun in the phrase. Thus, articles receive no stem markup. Quantifiers such as *kein* ('no') or *irgendein* ('some'), demonstrative articles and possessive pronouns are also grouped into this set.

Prepositions Prepositions are marked with the case their argument takes: while some prepositions only subcategorize one case, others can occur with two values. Typically, the different values for case are tied to a difference in meaning (locational vs. directional), which makes the prediction of case a difficult problem. By making the case information accessible in form of stem-markup, the difficulty of determining the case from the prediction step is moved to the stem translation process, such that *PREP+Acc* can be translated as a preposition conveying a prepositional meaning, whereas *PREP+Dat* can be translated as a preposition conveying a locational meaning. This assumes that the translation model, having access to both source-side and target-side lexical information, is better equipped to select an appropriate *PREP+Case* combination.

Personal pronouns Personal pronouns are annotated with number and gender, which are considered as part of the stem. Additionally, the case values *nominative* and *non-nominative* are annotated, because English pronouns, with the exception of *you* and *it*, distinguish between subject pronouns (*I, he/she, we, they*) and object pronouns (*me, him/her, us, them*). English does, however, not differentiate between direct and indirect object pronouns. Thus, the prediction model still has to determine whether a pronoun labeled with non-nominative is a direct (accusative) or indirect (dative) object.[9]

Relative pronouns Relative pronouns are marked with case: with case being dependent on the pronoun's role in a sentence, the prediction of case was found quite difficult and is thus kept as part of the feature markup.

Verbs As the inflection prediction approach only handles nominal inflection, verbs are represented using their inflected surface form. Additionally, it is assumed that having access to the inflected verb has a positive influence on the case prediction through subject-verb agreement, which allows, at least in some constellations, to identify the subject. In principle, there are two interacting effects: accessible number information on nouns may trigger the translation model to select a matching verb form in the translation step, and the selected verb may later contribute to the feature prediction step; cf. section 3.3.1 for an example.

[9] Genitive personal pronouns exist, as in *wir gedenken seiner*$_{Gen}$: 'we commemorate him', but are extremely rare.

Surface	Stemmed + morphological features	Gloss
steht	steht[VVFIN]	*stands*
die	die<+ART><Def>[ARTdef]	*the*
Bank∅	Bank<+NN><**Fem**><**Sg**>[NN]	*bank*
vor	vor[APPR-vor-**Dat**]	*before*
einem	eine<+ART><Indef>[ARTindef]	*a*
ernsten	ernst<+ADJ><Pos>[ADJA]	*serious*
internen	intern<+ADJ><Pos>[ADJA]	*internal*
Kampf∅	Kampf<+NN><**Masc**><**Sg**>[NN]	*struggle*
gegen	gegen[APPR-gegen-**Acc**]	*against*
eingefahrene	ein<VPART>fahren<V><PPast><SUFF><+ADJ><Pos>[ADJA]	*entrenched*
bürokratische	bürokratisch<+ADJ><Pos>[ADJA]	*bureaucratic*
Interessen	Interesse<+NN><**Neut**><**Pl**>[NN]	*interests*

TABLE 3.2: Example for stemmed representation with translation-relevant morphological features. In the column 'surface', inflectional suffixes are highlighted. In the column with the stemmed representation, the stem-markup is highlighted. Other components, such as <+ADJ><Pos>, are considered as part of the stem, as per the output of SMOR.

All other word classes All remaining word classes, such as conjunctions or adverbs, do not undergo inflection[10] and are thus just represented with their surface form.

Table 3.2 shows the stemmed representation for the sentence already discussed in previous examples. The final stemmed representation is obtained by concatenating the stem or surface form with morphological markup, if applicable, and the part-of-speech tag of the respective word, which is necessary in the feature prediction models. To allow for easier processing, the part-of-speech tag is added in square brackets to the end of each stem, even though this information is already contained in some of the stems.

In table 3.2, the stem markup in the example is highlighted, as are the inflectional suffixes in the surface forms. Comparing the surface forms with the stemmed representation illustrates how the stemmed representation is more general, in particular when looking at the adjectives. At first glance, adding stem-markup to nouns might increase data sparsity. However, considering gender as an innate feature, only the number feature is actually added to the representation. Thus, for nouns having distinct forms for singular and plural, the markup does not add new forms, but the two already distinct forms are just represented in a different way. Only for nouns with identical singular and plural surface forms, annotating the number feature might introduce a second form

[10] The set of substituting pronouns as in *diese sind neu*: ('these are new') needs to be inflected according to the word they refer to, but are ignored in the current implementation and thus are just represented by their surface form.

where there was previously only one. However, in the stemmed representation, the case of nouns is not represented. Nouns are thus reduced to two forms (singular/plural) at the most, whereas the inflection paradigm of nouns typically consists of more than two forms: often distinct forms for number, and additionally distinct forms for grammatical case[11]. For prepositions, the annotation of case does not add data sparsity; even when two values for case are possible, both variants (dative/accusative) are typically similarly distributed, and as a closed word class, they belong to the set of high-frequency words. A somewhat surprising source of new forms are the part-of-speech tags added to verb surface forms: as infinitives and 3rd person plural forms in present tense are often identical, the distinction of VVFIN and VVINF actually creates two forms where there was previously only one form. Verbal inflection is beyond the scope of this work, and it was not further analyzed whether the annotation of part-of-speech tags to verbs has any influence. It might even be the case that the tag annotation provides useful information to the SMT system.

The example in table 3.2 shows that the stems itself can be complex and might contain information about the word's derivation, as e.g. the adjective stem `ein<VPART>fahren <V><PPast><SUFF><+ADJ><Pos>`. The derivation information itself is not considered during the translation and inflection prediction process, but each stem is just used as an intermediate representation based on which inflected forms are generated. In particular, the component `<Pos>` (positive form of adjectives, in contrast to comparative or superlative form), is just regarded as part of the stem. The same applies to negation prefixes; for example, the analysis `lösen<V>bar<SUFF><+ADJ><Pos>` (*lösbar*: 'solvable') is treated as individual stem while the stem with the prefix *un*- `un<PREF>lösen<V>bar<SUFF><+ADJ><Pos>` (*unlösbar*: 'unsolvable') is treated as another stem. Even though the stems are morphologically closely related, this is not exploited by the translation model.

3.2.4 Portmanteaus and word formation

Portmanteaus are a contraction of preposition and article. The merging is restricted and can occur only for *definite* articles in a *singular* phrase, with a preposition subcategorizing either *dative* or *accusative* case. Furthermore, there are restrictions with regard to gender, depending on the preposition. For example, zu+der_{Fem} is merged (*zur*: 'to+the'), while

[11] For singular nouns, the forms inflected for different case values are often identical except for genitive (and distinguished by accompanying determiners/adjectives). In plural, the forms for accusative/nominative occasionally are identical, whereas dative and genitive forms usually are distinct.

in+der$_{Fem}$ remains un-merged, even though *in* can be merged in a masculine or neuter phrase (*im/ins*: 'in+the'). In contrast to e.g. French, where the creation of a portmanteau is required in some settings (such as *à+le* → *au*), merging preposition and article is not obligatory in German. Both variants can be considered as "technically correct", even though there is often a strong preference for the one or the other realization for a sentence to be considered fluid and natural sounding[12].

For the stemmed representation, portmanteau prepositions are split into article and preposition in a pre-processing step: Based on the parser output, portmanteau prepositions (with the part-of-speech tag APPR) are reduced to their base preposition, and a definite article is inserted, which "inherits" the morphological features of the respective phrase. In the inflection step, prepositions and articles are merged[13] using a rule-based approach if they occur in a constellation that allows for merging.

There are several reasons to split portmanteau prepositions in the stemmed representation. A straightforward reason is that the merging decision depends on the inflection of the article, which is only determined in the feature prediction process, and is not known during the translation step. Thus, there is no point in having inflected portmanteaus in an otherwise stemmed representation. Another reason is that of data sparsity: portmanteaus belong to the set of highly frequent closed-class words, and they are well-covered by the translation model on unigram level. However, translating with larger units is generally preferable, and by generalizing portmanteau prepositions to basic prepositions and articles, rarer n-grams obtain more matching opportunities. For example, the sequence ART ADJ NN obtained after portmanteau splitting can match with such sequences occurring either with other prepositions or within noun phrases. The third reason is rather technical: one objective for the abstract representation besides better generalization is to provide a means to integrate linguistic information. Later in this thesis, the modeling of prepositions plays an important role, and this is greatly simplified if prepositions that are essentially the same are represented consistently.

There are other word formation processes that can be modeled with the inflection prediction approach. For example, compounds are highly productive and lead to data sparsity in SMT. With the derivational analysis given by SMOR, one strategy to

[12] Typically prepositions and articles are merged, unless the prepositional phrase, and thereby the article itself, is stressed or contrasted, or referred to by a relative clause.

[13] The currently implemented approach of *always* merging if possible seems to be working fairly well in the inflection prediction approach. While no detailed analysis for the question "merging vs. not merging" was carried out, an experiment in which all permutations of merged/non-merged portmanteaus in a sentence were ranked via language model entropy lead to no change in BLEU.

model compounding is to split compounds in the stemmed representation and train a translation model on split and stemmed data. This translation system has then access to the individual components of compounds, which greatly reduces data sparsity. In the translation output, the individual components are merged to form a compound, resulting in the original stemmed representation, that then undergoes feature prediction and generation of inflected forms. A first experiment on modeling compounds using such a strategy can be found in Fraser et al. (2012). Cap et al. (2014) present a functioning system combining compound handling and inflection prediction.

3.3 Feature prediction and generation of inflected forms

To generate inflected surface forms for the stemmed translation output, a two-stage approach is applied: in the first stage, the full set of inflectional features is predicted using sequence models (linear chain CRFs) having access to lexical information (stems) and the linguistic features. In the second stage, the morphological tool SMOR generates full surface forms based on stems annotated with all required morphological features. This section outlines the architecture of the prediction models and illustrates how to re-inflect stemmed translation output.

3.3.1 Feature dependencies

Before going into details about the setup of the feature prediction models, it is important to discuss potential dependencies between the four morphological features, which are illustrated in figure 3.4. Gender, as innate feature, is not dependent on any other feature, but has influence on the strong/weak inflection feature, as has number. Their influence on strong/weak inflection is rather straightforward; it applies to the settings enumerated in section 3.2.2 and is furthermore restricted to a *local context* within the NP/PP. Similarly, number is determined by the source side and independent of the other features.

In contrast, there is an influence of number on grammatical case which is rather indirect and complex. Verb and subject must agree in number, and the number information on the head of noun phrases allows in some cases to identify the subject, and hence the case, in the feature prediction step. Theoretically, the subject can be unambiguously identified if it is the only phrase in a sentence/clause that matches in number with the

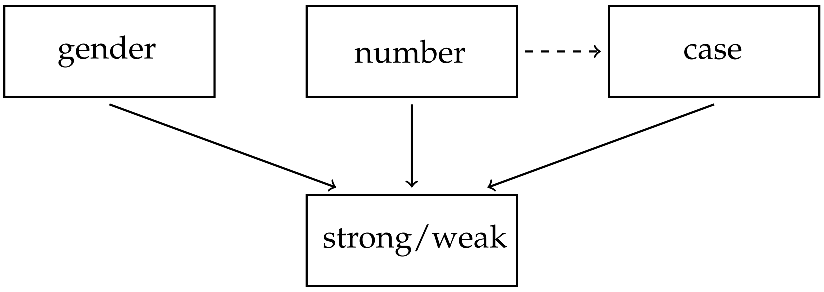

FIGURE 3.4: Relationship and dependencies between inflectional features.

finite verb, as illustrated in example (5) where the first NP is the subject, while in (6), only the second NP can be the subject.

	NP$_{\mathbf{Sg}}$	VVFIN$_{\mathbf{Sg}}$	NP$_{Pl}$	NP$_{Pl}$
(5)	[der Minister]$_{\mathbf{Nom}}$	schickte	[seinen Kollegen]$_{Dat}$	[die Berichte]$_{Acc}$
	the minister	*sent*	*to his colleagues*	*the reports*

	NP$_{Pl}$	VVFIN$_{\mathbf{Sg}}$	NP$_{\mathbf{Sg}}$	NP$_{Pl}$
(6)	[seinen Kollegen]$_{Dat}$	schickte	[der Minister]$_{\mathbf{Nom}}$	[die Berichte]$_{Acc}$
	to his colleagues	*sent*	*the minister*	*the reports*

The interaction between number and case can happen at two levels: the translation and the prediction step. Here, first the translation system indirectly benefits from the number information in combination with an inflected finite verb, as this provides information to select a verb form matching the number of the subject – remember that English finite verbs are not always marked for number (the *-s* for 3rd person singular is only found in present tense, but not past tense, for example). Then, in the feature prediction step, it can help to identify the subject NP. Practically however, the influence of subject-verb agreement in combination with nouns on case prediction is limited for several reasons. First, sentences with constellations as shown above are not always given. And second, the relevant constituents, i.e. verb and subject noun phrase, have to be adjacent, at least to a certain degree. Due to the flexibility of German clause ordering, a verb and its arguments can often be separated by large gaps, as illustrated by the subordinated clause in (7).

	KOUS	NP$_{\mathbf{Sg}}$	[...]	NP$_{Pl}$	NP$_{Pl}$	VVFIN$_{\mathbf{Sg}}$
(7)	dass	der Minister	[...]	seinen engsten Kollegen	viele neue Berichte	schickte
	that	*the minister*	*[...]*	*to his closest colleagues*	*many new reports*	*sent*

Here, the subject NP is at the beginning of the clause, with the finite verb at its very end. In between are the direct and indirect objects, and in the slot represented by *[...]* can be any number of adjuncts. Thus, there can easily be gaps of 8 or more words between verb and subject. The relationship between number and case goes beyond the local noun phrase context and needs to consider a *global context* spanning the entire sentence or clause. Thus, even though the features number and case are not independent, number can only have an influence on grammatical case in some constellations, and even then its practical influence in an actual model can be annihilated if there is too large a gap between the verb and its subject.

When predicting the inflectional features, the question arises to what extent the dependencies between the features need to be considered. An important practical factor is the emerging search space and its tractability. The search space of the prediction task corresponds to all possible prediction outcomes; here, it can be defined as the product of the values of the four features:[14]

- **Number** *Sg, Pl, **
- **Case** *Acc, Dat, Gen, Nom, **
- **Gender** *Masc, Fem, Neut, **
- **Strong/weak** *St, Wk*

This leads to $3 \times 5 \times 4 \times 2 = 120$ possible labels to be predicted. As this is intractable, each feature is predicted separately, with a search space of maximally 5 labels (case). Thus, in the prediction, number, gender and case are considered as independent, whereas strong/weak inflection takes as input the predicted values of the other features. For the prediction of case, the number information on nouns/pronouns is accessible in the form of stem markup. This should be sufficient, as for the identification based on subject-verb agreement, the number annotated to the head of the phrase, i.e. the noun, already provides the required information.

Joint versus individual feature prediction

Fraser et al. (2012) present a comparison between a *joint prediction model* modeling all four features and four *individual prediction models*. These model variants are trained

[14] The special value * stands for undefined. The parse output is not always complete with regard to morphological features: if the parser fails to determine the value of a feature, it is not annotated in the parse tree and subsequently represented as * in the training data.

as Hidden Markov Models (HMMs) on POS-tags with only limited access to lexical features. The HMMs essentially function as language model, and are considerably less demanding to train, thus being tractable also when modeling all four features jointly. While the joint model is indeed better than the four separate models, the HMM models are inferior to the four separate CRFs, which have access to both lexical and POS features. Thus, a joint feature prediction is generally preferable, but the advantage of the CRF in comparison to the HMMs outweighs the benefit of joint prediction. The experiments in Fraser et al. (2012) also show that only the system variant using features predicted by the four individual CRFs significantly outperform the baseline, whereas system variants relying on HMM-based predictions fail to improve over the baseline.

3.3.2 Feature prediction models

For the feature prediction, a model for each feature is trained using the Wapiti toolkit[15] (Lavergne et al., 2010). Wapiti is a toolkit to train and label sequences with discriminative models. It provides implemenations for maximum entropy Markov models and linear chain conditional random fields (CRFs), which are used for the prediction task in the inflection step. Linear chain CRFs are a type of undirected probabilistic modeling methods typically used for structured predictions, and they can take into account context information from the input sequence. In NLP applications, linear chain CRFs are often used to predict a sequence of labels for an input sequence, often for applications such as POS-tagging. In the inflection prediction step, the input sequence is the translation output in the underspecified stemmed representation and the output is a label sequence of the morphological feature values.

The training data for the CRFs combines lexical information in form of stems with linguistic information, i.e. the annotation of the context of the modeled feature. To train the four individual models, there is a set of *common feature functions* that is used in each model (such as stems and POS tags), and *individual feature functions* that take into account the context of only that linguistic feature that is modeled. The output of each model consists of sequences of this feature's values; the outputs of the four models are subsequently combined to provide the full set of inflectional features.

When training and applying the feature prediction models, the stem markup discussed in section 3.2.3 comes into effect by setting a particular value that is then propagated over the rest of the phrase. For example, with the *number* and *gender* of a noun

[15] https://wapiti.limsi.fr/

Stem	**POS**	**Force**	**PrevCase**	**Gender**	**Case**	**Number**	**St/Wk**	**NextCase**	**Label**
0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	
steht	VVFIN	0	X	NG	NC	NN	–	X	-
die<+ART><Def>	ART	F	X	NG	NC	NN	–	X	Nom
Bank<+NN><Fem><Sg>	NN	F	X	Fem	NC	Sg	–	X	Nom
vor	APPR-Dat	0	X	NG	NC	NN	–	Dat	-
eine<+ART><Indef>	ART	F	X	NG	NC	NN	–	X	Dat
ernst<+ADJ><Pos>	ADJA	F	X	NG	NC	NN	–	X	Dat
intern<+ADJ><Pos>	ADJA	F	X	NG	NC	NN	–	X	Dat
Kampf<+NN><Masc><Sg>	NN	F	X	Masc	NC	Sg	–	X	Dat
gegen	APPR-Acc	0	X	NG	NC	NN	–	Acc	-
ein<VPART>fahren<V> <PPast><SUFF><+ADJ><Pos>	ADJA	F	X	NG	NC	NN	–	X	Acc
bürokratisch<+ADJ><Pos>	ADJA	F	X	NG	NC	NN	–	X	Acc
Interesse<+NN><Neut><Pl>	NN	F	X	Neut	NC	Pl	–	X	Acc

TABLE 3.3: Training data example for the feature prediction models with labels for the prediction of *case*.

given, the linear chain CRF learns that the preceding adjectives and articles typically also have the same values for these features. Similarly, the *grammatical case* of prepositions is given and distributed over the rest of the phrase. To model *strong/weak inflection*, the output of the other features is added to the input, and the CRF learns the respective inflection patterns. This is a difference to the CRF prediction in Fraser et al. (2012), where strong/weak inflection is modeled as independent feature.

In the training data, there are no explicit phrase boundaries, but the model learns "phrase patterns" by having access to part-of-speech tags.

Table 3.3 shows a part of the training data with labels for the feature *case*. With the exception of the training data for *strong/weak inflection* (where the values for the other features are added), the "left side" of the data, i.e. stems and linguistic features (columns 0-8 in table 3.3), is the same, and only the "right side", i.e. the label, is different for each model. In the training process, only the feature columns for the linguistic feature modeled are addressed.

Before going into detail about the setup of the feature functions, the linguistic features contained in the training data, as shown in table 3.3, are presented:

Stem, POS The stem+POS sequence output by the SMT system is separated into a sequence of stems and POS-tags. While the stems provide exact lexical context, the

POS-tags provide more coarse-grained structural information. It is assumed that they are particularly important for the CRF to learn phrase-boundaries.

Force This feature indicates whether the CRF should predict an actual value relevant for nominal inflection, or a pseudo-label (represented with "–" in table 3.3). There are three possible values for this feature: words marked with *F* (*force prediction*) are associated with "real" labels, whereas words marked with *0* are associated with the pseudo-label. A third value (*U*) suppresses the prediction of the undefined value *: training examples with the label * are marked with *U*, and by only having the features *F* and *0* in the input data, the undefined value * is effectively prevented as predicted label.

PrevCase This feature represents "previous case" for phrases preceding postponed prepositions (APPO, for example *der Presse zufolge*$_{APPO-Dat}$: 'the press according-to'). Postponed prepositions occur at the end of the phrase, as opposed to "regular" prepositions at the beginning. Words of the category APPO are annotated with the respective value of case.

Gender For nouns, the gender value given in the markup is annotated.

Case The stem-markup only contains case annotation for personal pronouns, which is annotated in this feature column for the values *nominative* and *non-nominative*.

Number For nouns, the number value given in the markup is annotated.

Strong/weak These values are only defined for the strong/weak model and consist of the concatenated labels of the three other models.

NextCase This feature is applied for "regular" prepositions and denotes the case of the PP headed by the annotated preposition (cf. annotation for APPR in table 3.3).

Label The labels for the for models consist of the set of labels as listed in the enumeration on page 38, with the additional pseudo-label "–" for words where features for nominal inflection are not applicable, such as verbs and prepositions.

The example in table 3.3 illustrates how the grammatical case set in the stem-markup is reflected by sequences of the same value spanning a phrase. It is important to note that the features number and gender are straightforward to predict, as they are defined by the stem markup in all contexts, i.e. on the head of each noun phrase or prepositional

Unigram Stems	**Unigram Tags**	**Unigram Case**
stem$_{i-5}$	pos$_{i-7}$	case$_{i-5}$
$\vdots$	$\vdots$	$\vdots$
stem$_i$	pos$_i$	case$_i$
$\vdots$	$\vdots$	$\vdots$
stem$_{i+5}$	pos$_{i+7}$	case$_{i+5}$

2-gram Stems	**3-gram Stems**
stem$_{i-1}$ stem$_i$	stem$_{i-2}$ ... stem$_i$
stem$_i$ stem$_{i+1}$	

2-gram Tags	**3-gram Tags**	**4-gram Tags**	**5-gram Tags**
pos$_{i-1}$ pos$_i$	pos$_{i-2}$ ... pos$_i$	pos$_{i-3}$ ... pos$_i$	pos$_{i-4}$ ... pos$_i$
pos$_i$ pos$_{i+1}$	pos$_{i-1}$ ... pos$_{i+1}$		

TABLE 3.4: CRF training pattern: individual feature function (unigram case) and common feature functions (n-grams stems and pos tags)

phrase. In contrast, case is only defined for prepositional phrases via stem-markup on the preposition. For noun phrases, the feature case is predicted solely based on the context accessible to the CRF – this makes *case* the most difficult feature to predict.

When predicting features for SMT output, the sequences obtained by the four individual models are simply concatenated to feature combinations such as *Neut.Acc.Pl.Wk*, which are then transferred to the component generating inflected forms (cf. section 3.3.3).

The context window and inflectional features to be addressed in the CRF are defined via *feature functions* referring to the respective feature columns in the training data. Table 3.4 shows the set of the most important common and individual feature functions used to train the predictions models. As common feature functions are considered all entries referring to stems (exact lexical information) and POS-tags (coarse information), whereas the individual feature functions refer to the feature that is modeled, such as *case* in the example. For stems and POS-tags, the model looks at a window of 5 (stems) or 7 (POS-tags) words to the left and right. Additionally, there are n-grams (i.e. concatenations of the items in the defined window) for POS-tags and to a lesser extent also for stems. These n-grams primarily take into account the left side and/or the immediate context to the right. An individual feature function refers to the linguistic feature modeled in the respective CRF and takes into account a window of 5 positions to the left and to the right.

3.3.3 Generation of inflected forms

The re-inflection of the stemmed SMT output is basically a lookup of the inflected form for a combination of a stem and a set of morphological features. In a first step, a table containing the full inflectional paradigm for all stems occurring in the translation output is built. Conceptually, this is a straightforward procedure by generating all stem-feature-value permutations and running them through SMOR. However, SMOR's peculiarities require some adaptations to this process because the feature strong/weak inflection is only used as explicit feature if it generates two distinct forms. Furthermore, an extra value for gender, `<NoGend>`, can be used if the forms for the different values of gender are equal, as is often the case in plural.[16] Thus, the stem permutations need to contain the extra values $\emptyset$ for no explicit use of *strong/weak* and `<NoGend>` for plural forms. For invalid stems, SMOR simply returns nothing.

Table 3.5 shows all valid stems and their inflected forms for the noun stems `Park<NN>Automat<+NN>` ('parking ticket machine') and `Test<NN>Flug<+NN>` ('test flight') and the adjective stem `innenpolitisch<+ADJ><Pos>` ('concerning internal politics'). For most nouns (unless, for example, derived from an adjective as in *Abgeordneter*: 'deputy'), the feature *strong/weak* is not relevant. In contrast, it needs to be specified for most adjective forms. Similarly, the feature value `<NoGend>` is not used for nouns, with the exception of nouns derived from adjectives, but is often applied for adjective plural forms. The mapping from stems+features to inflected forms is unique in the sense that a stem-feature combination can typically only be mapped to one surface form[17], but a surface form can be generated by several stems.

The re-inflection step is based on the mapping of stem-feature pairs to surface forms: for each stem in the translation output, the inflected form is simply looked up. Words that are not part of nominal/prepositional phrases are already translated as surface forms; their tag annotation is just removed in the re-inflection process. When searching in the table for an inflected form for a stem-feature combination, the first lookup is based on the entire feature set. If no stem could be identified, the strong/weak annotation is removed, and the actual gender value replaced by the generic `<NoGend>` if applicable. As the use of strong/weak or the generic gender can depend on the stem (such as for nouns derived from adjectives), this procedure allows to consistently use the "classic

[16] There is, however, no value such as `<NoCase>`, even though the plural form *Parkautomaten* in table 3.5 works for all values of case.

[17] Assuming that over-generated, incorrect forms are filtered out. However, some inflected forms allow for alternative spellings, such as *Programms* vs.*Programmes*.

Park<NN>Automat<+NN><Masc><Acc><Sg>	Parkautomaten
Park<NN>Automat<+NN><Masc><Acc><Pl>	Parkautomaten
Park<NN>Automat<+NN><Masc><Nom><Sg>	Parkautomat
Park<NN>Automat<+NN><Masc><Nom><Pl>	Parkautomaten
Park<NN>Automat<+NN><Masc><Dat><Sg>	Parkautomaten
Park<NN>Automat<+NN><Masc><Dat><Pl>	Parkautomaten
Park<NN>Automat<+NN><Masc><Gen><Sg>	Parkautomaten
Park<NN>Automat<+NN><Masc><Gen><Pl>	Parkautomaten
Test<NN>Flug<+NN><Masc><Acc><Sg>	Testflug
Test<NN>Flug<+NN><Masc><Acc><Pl>	Testflüge
Test<NN>Flug<+NN><Masc><Nom><Sg>	Testflug
Test<NN>Flug<+NN><Masc><Nom><Pl>	Testflüge
Test<NN>Flug<+NN><Masc><Dat><Sg>	Testflug
Test<NN>Flug<+NN><Masc><Dat><Pl>	Testflügen
Test<NN>Flug<+NN><Masc><Gen><Sg>	Testflugs
Test<NN>Flug<+NN><Masc><Gen><Pl>	Testflüge
innenpolitisch<+ADJ><Pos><Fem><Dat><Sg><St>	innenpolitischer
innenpolitisch<+ADJ><Pos><Fem><Gen><Sg><St>	innenpolitischer
innenpolitisch<+ADJ><Pos><Fem><Gen><Sg><Wk>	innenpolitischen
innenpolitisch<+ADJ><Pos><Fem><Acc><Sg>	innenpolitische
innenpolitisch<+ADJ><Pos><Fem><Nom><Sg>	innenpolitische
innenpolitisch<+ADJ><Pos><Masc><Nom><Sg><St>	innenpolitischer
innenpolitisch<+ADJ><Pos><Masc><Dat><Sg><St>	innenpolitischem
innenpolitisch<+ADJ><Pos><Masc><Nom><Sg><Wk>	innenpolitische
innenpolitisch<+ADJ><Pos><Masc><Acc><Sg>	innenpolitischen
innenpolitisch<+ADJ><Pos><Masc><Gen><Sg>	innenpolitischen
innenpolitisch<+ADJ><Pos><Neut><Acc><Sg><St>	innenpolitisches
innenpolitisch<+ADJ><Pos><Neut><Nom><Sg><St>	innenpolitisches
innenpolitisch<+ADJ><Pos><Neut><Dat><Sg><St>	innenpolitischem
innenpolitisch<+ADJ><Pos><Neut><Acc><Sg><Wk>	innenpolitische
innenpolitisch<+ADJ><Pos><Neut><Nom><Sg><Wk>	innenpolitische
innenpolitisch<+ADJ><Pos><Neut><Gen><Sg>	innenpolitischen
innenpolitisch<+ADJ><Pos><NoGend><Dat><Sg><Wk>	innenpolitischen
innenpolitisch<+ADJ><Pos><NoGend><Acc><Pl><St>	innenpolitische
innenpolitisch<+ADJ><Pos><NoGend><Nom><Pl><St>	innenpolitische
innenpolitisch<+ADJ><Pos><NoGend><Gen><Pl><St>	innenpolitischer
innenpolitisch<+ADJ><Pos><NoGend><Acc><Pl><Wk>	innenpolitischen
innenpolitisch<+ADJ><Pos><NoGend><Nom><Pl><Wk>	innenpolitischen
innenpolitisch<+ADJ><Pos><NoGend><Gen><Pl><Wk>	innenpolitischen
innenpolitisch<+ADJ><Pos><NoGend><Dat><Pl>	innenpolitischen

TABLE 3.5: Example for the lookup table containing all valid stem-feature combinations as used by SMOR on the left side and inflected forms on the right side. For nouns, the *strong/weak* feature is often not relevant. Adjective forms iterate through all values for gender, including the value <NoGend>. The use of *strong/weak* depends on the feature constellation. *Parkautomat*: 'parking ticket machine'; *Testflug*:'test flight'; *innenpolitisch*: 'concerning internal politics'.

stems	stem markup	predicted features	inflected form	gloss
`die<+ART><Def>[ARTdef]`		Masc.Nom.Sg.St	der	*the*
`Beginn<+NN>[NN]`	Masc.Sg	Masc.Nom.Sg.Wk	Beginn	*beginning*
`die<+ART><Def>[ARTdef]`		Masc.Gen.Sg.St	des	*of-the*
`bauen<V><SUFF><+NN>[NN]`	Masc.Sg	Masc.Gen.Sg.Wk	Baus	*construction*
`die<+ART><Def>[ARTdef]`		Fem.Gen.Sg.St	der	*of-the*
`Gas<NN>Leitung<+NN>[NN]`		Fem.Gen.Sg.Wk	Gasleitung	*gas pipeline*
`South[NE]`		–	South	*south*
`Stream[NE]`		–	Stream	*stream*
`in[APPR-in-Dat]`	Dat	Dat	in	*in*
`Bulgarien[NE]`		–	Bulgarien	*Bulgaria*
`markiert[VVFIN]`		–	markiert	*marks*
`die<+ART><Def>[ARTdef]`		Masc.Acc.Sg.St	den	*the*
`Start<+NN>[NN]`	Masc.Sg	Masc.Acc.Sg.Wk	Start	*start*
`eines[PIS]`		–	eines	*of-one*
`die<+ART><Def>[ARTdef]`		Neut.Gen.Pl.St	der	*ot-the*
`groß<+ADJ><Sup>[ADJA]`		Neut.Gen.Pl.Wk	größten	*largest*
`Projekt<+NN>[NN]`	Neut.Pl	Neut.Gen.Pl.Wk	Projekte	*projects*
`in[APPR-in-Dat]`	Dat	Dat	in } im	*in-the*
`die<+ART><Def>[ARTdef]`		Masc.Dat.Sg.St	dem	
`Energie<NN>Bereich<+NN>[NN]`	Masc.Sg	Masc.Dat.Sg.Wk	Energiebereich	*energy sector*
`,[$,]`		–	,	,
`sagte[VVFIN]`		–	sagte	*said*
`die<+ART><Def>[ARTdef]`		Masc.Nom.Sg.St	der	*the*
`Chef<+NN>[NN]`	Masc.Sg	Masc.Nom.Sg.Wk	Chef	*chief*
`von[APPR-von-Dat]`	Dat	Dat	von	*of*
`Gazprom[NE]`		–	Gazprom	*Gazprom*
`.[$.]`		–	.	.

TABLE 3.6: Post-processing steps for the feature prediction and generation of inflected forms for the stemmed translation output of the input sentence *the start of construction of the south stream gas pipeline in bulgaria marks the launch of one of europe's largest energy projects, gazprom's chief said.*

feature values" in the feature prediction models, even though they might be redundant in some constellations. By reducing/modifying the features to match SMOR's set of feature values, the representation used for prediction and generation can easily be joined.

Table 3.6 shows the different stages in the post-processing step for re-inflecting the translation output. Starting with the feature prediction, the complete set of inflection features is computed using the stem markup as input to set a subset of the values (cf. columns 2 and 3 in table 3.6). Based on stems and predicted features, the generated inflected forms are selected. In a last step, portmanteau prepositions are merged based on a pre-defined set of preposition+article pairs that allow for merging: in the PP *in* *dem Energiebereich* ('in the energy sector'), the preposition and the article are merged to form the final PP *im Energiebereich*. In the current implementation, proper nouns (with the tag NE) are not inflected, but translated as they are.[18] The sentence in table 3.6 is well translated, despite being rather complex and long. In particular at the beginning, there are three consecutive noun phrases, of which the first (*der Beginn*: 'the start') is the subject (nominative) of the clause, followed by two modifying phrases in genitive case – the values for case have to be predicted entirely based on the context. For example, it is generally likely that a sentence starts with a subject phrase. Furthermore, it is likely that the phrase *der Beginn* ('the beginning') is preceding a modifying genitive phrase (corresponding to '[the start] [of something]'). Similarly, the noun phrase following the verb *markiert* is correctly identified as direct object (accusative) – it occurs in a position where a direct object reasonably can be expected.

3.4 Experiments and results

In this section, the results for the presented inflection prediction system are discussed and compared to a baseline system trained on standard data.

3.4.1 Experimental settings

All systems are built with the Moses phrase-based framework. The English–German parallel corpus consists of 4.592.139 sentences (WMT'15 Shared Task[19]) aligned with GIZA++ (Och et al., 2003). A 5-gram target-side language model is built on over 45 million German sentences obtained from concatenating the parallel data and the News'14 set (45.954.154 for the surface system, and 45.717.097 for the inflection prediction model[20]). To obtain the stemmed representation, the German data was parsed with BitPar (Schmid, 2004) and analyzed with SMOR (Schmid et al., 2004). The morphological features number, gender, case and strong/weak are predicted with four CRFs trained with the Wapiti toolkit (Lavergne et al., 2010) on the target-side of the parallel data. The

[18] While proper nouns sometimes can be inflected, e.g. marked for genitive, it was decided to not include proper nouns in the set of words that are stemmed and re-inflected – the main reason is that it cannot be guaranteed that the required form can be generated: SMOR is able to handle some proper names, e.g. *Bulgarien* or $\textit{Bulgariens}_{Gen}$, but not, for example $\textit{Obamas}_{Gen}$.

[19] `http://www.statmt.org/wmt15/translation-task.html`

[20] The size difference is due to minor pre-processing differences and parse-fails for the inflection prediction system.

POS	correct	total	Percentage of correct predictions per tag
ADJA	2720	3163	0.8599
ARTdef	5290	5794	0.9130
ARTindef	1122	1242	0.9034
NN	11001	11526	0.9545
PDAT	344	369	0.9322
PIAT	509	540	0.9426
PPER	2115	2157	0.9805
PPOSAT	601	681	0.8825
PRELAT	12	25	0.4800
PRELS	515	731	0.7045
PRF	397	398	0.9975
PWAT	17	18	0.9444
total	**26644**	**24643**	**0.9249**

TABLE 3.7: Prediction results on clean data:
tuning set (newstest'13), containing 3000 sentences.

system is tuned and tested on news data, using the WMT'13 set (3000 sentences) as tuning data, and WMT'14 (3003 sentences) and WMT'15 (2169 sentences) as test data.

3.4.2 Prediction on clean data

An important factor in the inflection prediction pipeline is the quality of the prediction models. However, it is impossible to measure the accuracy of the prediction models on actual translation output, as the translated sentences obviously differ from the reference translations, which makes an automatic evaluation of the prediction accuracy impossible. The prediction models are thus applied on clean data (i.e. the reference sentence), and the generated inflections are compared with the original surface forms.

Table 3.7 shows the amount of correctly predicted and generated inflections on the development set (3000 sentences). In total, 92.5 % of all predicted forms are correct. Looking at the different tags, it becomes evident that most categories are predicted with an accuracy of over 90 %, with the exception of a few outliers.

Most important for the modeling of nominal phrases are determiners, adjectives and nouns as they represent the majority of the modeled elements. Between these, the generation of nouns works best, whereas the results for determiners and adjectives are somewhat worse. This can be explained by the fact that for nouns, many feature combinations correspond to the same surface form (for example, strong/weak inflection

	tuning-1		tuning-2	
	News'14	**News'15**	**News'14**	**News'15**
Surface System	19.17	20.86	19.03	20.80
Inflection Prediction	19.35	21.21*	19.32*	21.16*

TABLE 3.8: Results (case-insensitive BLEU) for inflection predictions versus surface systems: News'14 (3,003 sentences) and News'15 (2,169 sentences). * denotes a significant improvement over the respective baseline with pairwise bootstrap resampling with sample size 1,000 and a p-value of 0.05.

or accusative/nominative form), whereas the surface forms of articles and adjectives tend to differ more. Thus, measuring the prediction quality by comparing inflected forms is slightly more "forgiving" for nouns than it is for adjectives and articles.

The prediction of pronouns (PPER) is rather reliable; this can be attributed to the annotation of nominative to subject pronouns. On the other hand, the prediction of relative pronouns (PRELAT, PRELS) leads to the worst results. A reason might be that pronouns substitute syntactic functions such as object and subject, and depend on other constituents while not contributing actual lexical information – in short, they can take on any role, but do not provide much information to support the prediction. In terms of occurrence frequency, they do not play an important role, and thus do not have too much of a negative influence.

3.4.3 Translation experiments

Table 3.8 compares the surface system and the inflection prediction system, showing the results for two different runs of parameter tuning[21]. The translation quality is measured by the automatic metric BLEU (Papineni et al., 2002). While all inflection prediction systems are better than the respective system operating on surface forms, the difference is not significant for News'14 in the first tuning run. The second tuning run yields slightly worse results for all test sets; here, however, the inflection prediction system for News'14 is significantly better than the surface system. Summarizing, one can draw the conclusion that the inflection prediction systems result in better translation quality, even though there is not always a significant improvement in BLEU.

Generally, the impact of the inflection prediction approach diminishes with increased training data size (cf. Burlot et al. (2016) for an analysis of an equivalent approach for

[21] As parameter tuning is non-deterministic, the resulting parameter set can be more or less optimal, leading to slight fluctuations in the BLEU scores.

English–Czech translation.) However, the inflection prediction system tends to be better than an equivalent surface system; this has also been confirmed by Cap et al. (2014), who report an improvement over a surface system when applying essentially the same approach for inflected prediction.

Another factor is the evaluation metric: BLEU is a precision-oriented metric that compares n-grams in the MT output with n-grams in the reference translation. As the phenomena modeled by the inflection prediction approach, i.e. consistent inflection within a phrase, often are rather subtle changes on the morphological level – though important for a human reader – they may not be fully captured by an evaluation in terms of BLEU. The assumption that BLEU has a tendency to under-estimate the quality of the inflection prediction system is also supported by the outcome of the manual evaluation of the WMT Shared Task 2015 (system description in Cap et al. (2015a); overall system evaluation in Bojar et al. (2015)): the submitted system employing inflection prediction (combined with source-side pre-ordering) was ranked in the same cluster (i.e. a group of systems found to be of equivalent quality in the manual evaluation) as systems with considerably higher BLEU scores.

3.4.4 Generation of novel word forms

The inflection prediction system is able to generate novel word forms based on the seen stems in the training data, and an important evaluation question is whether the system actually does create unseen word forms. Table 3.9 shows the newly generated words for the two test sets. They consist mostly of compound nouns, but there are also some adjectives. For some of the words generated for test set News'14, the newly generated inflected form also match at sentence-level with the reference translation, indicating that the inflection is correct in the sentence. Of course, this underestimates the amount of useful novel inflected forms, as a translation can be valid even if it does not occur in the reference translation – this is illustrated by the following two examples.

(8) EN as a **placatory** gesture , the prime minister insisted , ...

BL als **abwiegelnden** geste , der ministerpräsident darauf bestanden , ...
as placatory gesture , the prime-minister on-it insisted , ...

INFL als **abwiegelnde** Geste , der Ministerpräsident darauf bestanden , ...
as placatory gesture , the prime-minister on-it insisted , ...

REF als **beschwichtigende** Geste bestand der Premierminister darauf , ...

Newly generated words in News’14
zerknittertes, theatralischem, vorhergesagtes, NSA-Mitarbeiter, geschmeidigste, Bibelversen, **NSA-Mitarbeiter**, Konzerngewinns, Netto-Bestellungen, Flugplatzbetreibers, Pfarreirat, Pfarreirat, Pfarreirat, Herbstfestivals, **37-jähriger**, Einzelwettbewerben, schlafloses, Zaungastes, Unabhängigkeitsplatzes, Absteigers, **Herzfehlers**, **ausbeuterischem**, eingewickelter, 36-jährigem, Toilettensitzes, **Regierungsmitglieds**, Forschungsbeitrags, *sexye, *Pässe-Kontrollen, *Pässe-Kontrollen, Erzählgedicht, *Festhalt, Bordprogramms, Vorbehaltsverkäufer, Vorbehaltsverkäufer, Vorbehaltsverkäufer, tasmanischem, aufschiebbare, aufgeschlitzte, erschöpftem, Speerspitz, Lebensgebiets, Bleistifts, *sexyen, wettbewerborientieren, semi-professioneller, Operationssaals, Revierkampfes, Aufräumens, Online-Marktplatzes.

Newly generated words in News’15
einwilligungsfähiges, Anti-Abtreibungsgesetz, gartenbaulicher, *Schlösserpark, schlafendem, verrohrtes, 32-jähriger, verdrehtem, rosaer, faserigem, eingeatmetem, aufsichtsrechtlichem, gerittenen, Alt-Hippie, herabsetzendes, verifiziertes, oscar-prämierter, stacheliges, Festländern, Lynchens, Co-Vorsitzendem, Raunens, Goldschatzes, Schießclubs, Wachbeamte, Trainingsgeländes, *Residenten-Übel, Zuckens, ausgerenkten, Werbegeschäfts, Frusts, abwiegelnde, Kreuzfahrt-Veranstalter, gemeldetem, Verlobtem, *Superstare

TABLE 3.9: Novel words in the two test sets (i.e. the words do neither occur in the surface training data, nor in the English input sentence). Words in bold-face match with the reference translation at sentence-level. Words marked with “*” are invalid words, i.e. incorrectly generated by SMOR.

As the surface training data does not contain the form *abwiegelnde*, the baseline translation (BL) in (8), has to use the available inflected form *abwiegelnden*, which is the only translation option for the English word *placatory*. In the inflection prediction system (INFL), the word is represented by the stem `ab<VPART>wiegeln<V><PPres><SUFF><+ADJ><Pos>[ADJA]`, from which the correct inflection can be derived in the post-processing step. As the corresponding adjective in the reference translation is different, this improvement is not credited by BLEU.

(9) EN she compared several german sayings with the corresponding **bible verses** and explained the meaning .

BL sie verglich mehrere deutsche redensarten mit den entsprechenden **bibel verse**
she compared several German sayings with the corresponding bible verse
und erklärte die bedeutung .
and explained the meaning .

INFL sie verglich mehrere deutsche Redensarten mit den entsprechenden **Bibelversen**
she compared several German sayings with the corresponding bible-verses
und erklärte die Bedeutung .
and explained the meaning .

REF sie verglich mehrere Sprichwörter mit den entsprechenden **Bibelstellen** und erklärte die Bedeutung .

Compounding is a productive word formation process in German, and thus challenging for machine translation. While there is no explicit compound handling in the presented inflection prediction approach, the ability to generate novel inflected forms from stems is particularly relevant for compounds. This is illustrated in (9), where the compound *Bibelversen* has been correctly generated by the inflection prediction system (INFL). The baseline (BL) opted to translate *bibel verse* separately, but incorrectly inflected. While this is understandable, the correctly inflected compound in the inflection prediction output is preferable. Again, with a different word chosen in the reference translation, this improvement goes unnoticed by BLEU.

Of course, not all novel inflections listed in table 3.9 are correct – some are invalid German words (e.g. *Pässe-Kontrolle* – *Pass-Kontrolle* is correct), and some do not fit into the sentence context. An amusing example is the word creation found in example (10):

(10) EN ... durch ihre Rolle in der **sexyen** Fernsehserie Baywatch ...
... through her role in the sexy television-series baywatch ...

While the inflection of *sexy* is technically correct, this adjective, as a word borrowed from English, remains uninflected – here, SMOR was a bit overzealous!

3.4.5 Discussion and examples

The role inflectional errors play in understanding a translation depends on the type and severity: while agreement violations within a phrase disrupt the fluency of a sentence and are perceived as annoying, they do not necessarily disturb the understanding of the meaning as long as relevant and content-bearing features, such as grammatical case or number, are still represented. Inflectional errors that concern the representation of grammatical case tend to be more problematic as they mix up semantic roles by assigning the wrong syntactic function.

In the following, some example outputs of the inflection prediction system are discussed in comparison to the output of the system trained on standard surface forms. The sentences are chosen to represent the range of inflectional errors, from mildly annoying to hardly understandable.

(11) EN in particular , the actresses play a major role in the sometimes rather dubious staging .

SURF insbesondere die schauspielerinnen spielen eine große rolle in der manchmal
in-particular the actresses play a major role in the sometimes
etwas **fragwürdige** inszenierung .
rather dubious staging .

INFL insbesondere die Schauspielerinnen spielen eine große Rolle in der manchmal
in-particular the actresses play a major role in the sometimes
etwas **fragwürdigen** Inszenierung .
rather dubious staging .

The translations for the sentence in (11) are nearly equal: the baseline translation is correct, except for the inflection of the adjective *fragwürdig* ('dubious'). Both *fragwürdig* and *inszenierung* are rather low-frequency words, and while the correct surface form *fragwürdigen* does occur in the parallel training data with a similar frequency than *fragwürdige*, the parallel data does not contain a bigram of the two words. However, in the language model data, there are two occurrences of *fragwürdige inszenierung* (which is a valid n-gram in another context), but none of *fragwürdigen inszenierung*. This might have lead to a preference for the incorrect inflection selected by the baseline system. In the inflection prediction system, the language model only has the stemmed sequence which does not trigger an unfortunate decision due to random matching. Having access to the relevant features, the correct inflection can be generated in the post-processing step. However, despite the wrong form, the baseline sentence is perfectly understandable.

(12) EN an additional radar sensor checks whether the green phase for the pedestrian can be ended .

SURF **eine zusätzliche** abbildenden **sensoren** überprüft , ob **die grünen phase** für
an additional imaging sensor checks , whether the green phase for
die fußgängerzone kann beendet werden .
the pedestrian-zone can ended be .

INFL **ein zusätzlicher Radar Sensor** überprüft , ob **die grüne Phase** für die
an additional radar sensor checks , whether the green phase for the

Fußgängerzone kann beendet werden .
pedestrian-zone can ended be .

The translations for the input sentence in (12) obtained with the baseline system and the inflection prediction system differ in two NPs: the first NP *eine zusätzliche abbildenden sensoren* is incorrectly inflected due to agreement violations, but also lexically wrong as it also contains the adjective *abbildenden* ('imaging'), while at the same time lacking a translation for *radar*. Also, the second NP in the baseline output is not consistently inflected. Nevertheless, the baseline translation manages to transport most of the meaning of the source sentence, with the exception of the incorrect translation of the modifier *radar*. The inflection prediction system outputs essentially correct translations for both NPs; a minor point of criticism is the realization of *Radar Sensor* as two separate words instead of one compound (*Radarsensor*), which however has no strong negative influence on the understanding of the sentence.

(13) EN those drivers will soon pay the mileage fees instead of gas taxes to the state .
SURF **den fahrern wird** bald die gebühren zahlen , statt gas steuern an den staat .
to-the drivers will$_{Sg}$ soon pay the fees , instead gas taxes to the state .
INFL **die Fahrer werden** bald die Gebühren zahlen , statt Gas Steuern an den Staat .
the drivers will$_{Pl}$ soon pay the fees , instead gas taxes to the state .

In example (13), the inflection of the first NP in the baseline translation is consistent within the phrase (i.e. no agreement violations), but the grammatical case (dative) does not match the syntactic function of the phrase (subject, i.e. the correct case should be nominative). With the phrase *the drivers* translated as a dative object (i.e. as receivers of the payment), the sequence *den fahrern wird* (with the verb in singular), is technically correct with regard to agreement constraints, but does not represent a correct translation – rather, it comes close to meaning the contrary.

This example demonstrates the relation between subject NP and verb number: the realization of *those drivers* as dative-plural phrase does not restrict the selection of the verb number, as both a singular and plural indirect object are possible. It is important to note that the English verb itself (*will*) does not indicate, on its own, the number (as opposed to the regular *-s* suffix that marks English singular verbs in present tense). The realization as dative NP makes the understanding of the baseline translation more difficult, even though one might be able to derive the intended meaning based on the context. In the inflection prediction system, the intermediate stemmed representation only has the feature number, but no grammatical case yet. The constellation "plural-NP at the beginning of a sentence followed by a verb" might trigger a translation of a plural verb form. In the inflection step, the constellation "Plural NP at the beginning of a sentence followed by a verb in plural form" might trigger the prediction of nominative.

Problems in the inflection prediction approach While the previous examples demonstrated different types of improvement over the baseline system, the inflection prediction system also has limits, mainly with regard to the prediction of case. In particular, several adjacent noun phrases next to one another can pose a problem to the case prediction, as illustrated by the following two examples. The task to assign the correct value of case to a cluster of noun phrases is not trivial – due to the flexible word ordering in German, direct/indirect objects and the subject can occur in nearly every position in the sentence, potentially followed by one or more genitive modifiers.

(14) EN in order to give the guests at the anniversary event an insight into the initiatives mentioned.

INFL um [den Gästen]$_{Dat}$ **[die Jubiläumsveranstaltung]$_{\textbf{Acc}}$** [einen Einblick]$_{Acc}$
to the guests the anniversary-event an insight

[in die erwähnten Initiativen]$_{PP}$
into the initiatives mentioned

REF um [den Gästen]$_{Dat}$ **[der Jubiläumsveranstaltung]$_{\textbf{Gen}}$** [einen Einblick]$_{Acc}$
to the guests of-the anniversary-event an insight

[in die erwähnten Initiativen]$_{PP}$
into the initiatives mentioned

In the sentence in (14), there should be a complex indirect object (the dative NP *den Gästen* preceding the genitive NP *der Jubiläumsveranstaltung*), followed by the direct object *einen Einblick*. Instead, the case prediction (accusative for the phrase supposed to be genitive) leads to an unnatural and semantically flawed inflection. The fact that the English verb *to give* has not been translated adds to the problem – there are no clues about a subcategorization frame, or a possibility to narrow down the selection of possible values of case.

(15) EN the magnetic activity of a star , the interaction of its magnetic field and the emitted particle radiation play an important role .

SURF die magnetische aktivität eines sterns , die interaktion **seiner magnetfeld** und
the magnetic activity of-a star , the interaction of-his magnet-field and
der emittierten teilchenstrahlung eine wichtige rolle spielen .
the emitted particle-radiation an important role play .

INFL die magnetische Aktivität des Sterns , der Interaktion **des magnetischen**
the magnetic activity of-a star , of-the interaction of-the magnetic
Feldes und der Teilchenstrahlung wird eine wichtige Rolle spielen .
field and of-the particle-radiation will an important role play .

Similarly, the translation shown in example (15) also illustrates this problematic aspect of the inflection prediction system: while the output of the inflection prediction system is somewhat more fluent than the baseline output, the assignment of case is not correct: *der Interaktion des magnetischen Feldes* ('of-the interaction of the magnetic field') is inflected as genitive phrase, but is actually part of the subject (*the interaction of its magnetic field ... play an important role*). The source sentence contains multiple complex NPs from a domain that is likely not very well covered in the traning data – as a result, both translation variants are difficult to understand. In order to assign the correct values for case, the prediction model needs more information about the source-side function of a phrase, as well as subcategorization requirements of the verb to determine which NPs are subcategorized arguments and what role they represent, and which NPs are rather (genitive) modifiers or other adjuncts.

The case prediction model as presented in section 3.3.2 has only limited access to such information, and thus has only a limited ability for predicting case. The following chapter extends the basic prediction model by integrating external information in combination with source-side features to better handle the prediction of case.

3.5 Related work

There is a large body of research describing approaches to handle complex morphology in machine translation, both for the directions of translating *out of* a morphologically rich language, as well as *into* a morphologically rich language. This section presents related work for both translation directions, with a focus on translating into morphologically complex languages. The section concludes with a short overview of possible applications of the presented approach.

3.5.1 Morphologically rich source-language

The main problem when translating out of morphologically complex languages is the rich source-side vocabulary that is typically not well covered by the translation system. In addition to potentially uncovered inflected forms, much of the information encoded in the source-side language may not be strictly necessary for the translation process. That is not to say that inflectional information in the source language is per-se unimportant: for example, the instrumental case occurring in many Slavic languages denotes that a noun is used as instrument in order to achieve something (e.g. *he writes the letter with a pen*). To correctly translate the meaning, relevant case information does need to be taken into account and transferred into the target-side equivalent, for example a paraphrase such as a *by* or *with* PP. Similarly, syntactic functions such as subject or object are expressed by the position in the sentence. On the other hand, features such as gender and strong/weak inflection are arbitrary and specific to the target side.

The basic concept of translating from morphologically rich source languages consists in reducing the morphological complexity with the primary goal to reduce the vocabulary size to obtain better coverage. Information expressed through inflection is either discarded, or provided through alternative forms of representation. Generally, this is achieved by reducing an inflected *surface form* to its *base form*, optionally enriched with separated *inflectional suffixes* or *morphological tags*.

Nießen et al. (2001) are among the first to handle statistical translation from highly inflected languages. They present a method to better exploit bilingual data by using the relation between different derivatives of the same lemma. They introduce "equivalence classes", inflectional variants of the same base-form sharing similar translations, in order to ignore information not relevant for translation. They report improved translation results in a VerbMobil setting for German–English translation.

Popović et al. (2004) investigate possibilities to improve the translation quality for translating from the morphologically rich languages Spanish, Catalan and Serbian into the morphologically simple English by making use of morphologically enriched word stems. For example, inflected verb forms are substituted by the stem and a morphological annotation representing person, tense and mood. A variant consists in representing an inflected form by its stem and the inflectional suffix, for example *estaremos* ('we will be') → $\textit{esta}_{\text{STEM}}$ $\textit{+remos}_{\text{1PERS.PL.FUT}}$. Their method leads to reduced translation error rates for all three language pairs.

Similarly, Goldwater et al. (2005) modify the source-side of a Czech–English SMT system by reducing inflected forms to lemmas, and introducing "pseudo words" encoding morphological information, such as 1PERS. Such pseudo-words have the additional benefit of corresponding to the English function words, e.g. prepositions, and thus increasing the correspondence between source and target language. They conclude that morphological analysis as a basis for a modified representation can improve translation quality for highly inflected languages, even though the constellation of modeled features remains to be determined for the language pair.

Tiedemann et al. (2015) adopt a similar strategy to deal with the complex inflection of Finnish in a Finnish–English SMT system. As Finnish morphological markers (e.g. grammatical case) often correspond to separate English function words (e.g. prepositions), they experiment with adding "inflection pseudo-tokens": this has the effect of providing a more balanced structure for word alignment, and to reduce the amount of inflected surface forms by representing a part of the inflection in separate tokens. Chapters 6 and 7 will introduce a similar notation to model subcategorization across complement types.

Arabic contains a large set of morphological features, combining both concatenative and templatic morphology, which presents many challenges for machine translation due to the large amount of inflected forms and the information encoded in the features. Sadat et al. (2006) study different word-level pre-processing schemes to identify an optimal input representation for Arabic translation. A combination of various pre-processing schemes leads to improved translation quality for Arabic–English translation.

3.5.2 Morphologically rich target-language

Translating into a morphologically rich language is generally considered as more difficult than the reverse direction. The general concept for translating *from* a morphologically complex language consisted in reducing the source-side vocabulary by simplifying inflected forms to a stem or lemma, while keeping relevant morphological information as morphemes or as tags with morphological annotation. In contrast, a translation system translating *into* a morphologically rich language must output fully inflected forms, which might not even occur in the parallel training data, while fulfilling agreement conditions and other target-side requirements.

There are many approaches dealing with translating into morphologically rich languages, which can be roughly divided into

- methods that integrate linguistic information into the translation model, often with the objective to enforce consistent morphology;
- methods employing morphological segmentation to build translation systems operating on sub-word level;

- word-level systems that model target-side morphology, predominantly by explicit modeling of inflectional features on an intermediate representation with a subsequent generation of inflected forms. The inflection prediction system presented in this thesis belongs to this category.

With some strategies to handle rich target-side morphology, it is possible to generate target-side words that are not contained in the parallel training data, either through recombination of sub-word units, or through the use of an external morphological generation component.

Linguistic information in the translation model

A major problem when translating into morphologically rich languages is inconsistent morphology, for example agreement violations within noun or prepositional phrases, or between a verb and its arguments. Many approaches thus aim at integrating linguistic information into the translation system in order to enforce consistent morphology. Koehn et al. (2007a) propose a "factored translation model" as an extension to the standard phrase-based system to enable the use of rich information, represented as "factors", throughout the translation process. Standard statistical translation systems treat words as simple tokens, and do not take into account different surface forms of the same lemma. Instead, a factored system sees words rather as a vector containing information such as lemma, POS-tag and morphological information. This representation allows to translate lemmas and morphological information separately, and thus to pool evidence of different forms of the same lemma for better generalization (*translation step*). Then, the factors are combined on the output side to generate surface forms by a mapping from output factors to surface forms on word level (*generation step*). Integrating POS-information and morphological annotation as additional information in form of factors led to improved translation quality; however, the generation of inflected forms from lemmas was found to be not effective, due to problems caused by the separate modeling of lemmas and tags during translation.

Many methods to enrich translation models with (morphological) information are based on a factored translation model, as the framework is very general and allows for an easy integration of rich information. For example, Avramidis et al. (2008) enrich the source side of English–Czech and English–Greek systems in order to provide morphological attributes that are required by the target side, but missing on the source side as explicit morphological attribute. The features to model noun-case agreement and verb-person-conjugation are contained in the syntactic structure of the source side and can be derived from a parse output. The features are added as source-side factors. Exploring alternative-path strategies to handle sparse data (i.e. backoff to less factors), leads to moderate improvements for English–Greek and English–Czech translation. Similarly, Loáiciga et al. (2014) integrate rich source-side information to model verb tenses in an English–French factored phrase-based translation system. Aligned English–French verb phrases are annotated with tense-related features such as syntactic structure, semantic roles and temporal ordering and neighboring verb word forms (the "chain of verbs" as appearing in the sentence). These annotated pairs are used to train a "tense predictor" that is applied to the source side to provide the respective features to be annotated in the test set. This strategy lead to an improvement in BLEU; an error analysis revealed that the enriched system improves significantly at translating French tenses.

Williams et al. (2011) address the problem of morphologically inconsistent output by integrating unification-based constraints into a string-to-tree SMT system for English–German. The decoder works on surface forms while checking linguistic unification constraints at the same time in order to enforce inflectional consistency. A second advantage of this approach consists in an early elimination of morphologically inconsistent hypotheses from the search space. The linguistic annotation for the constraints is derived from a statistical parser and morphological analyses. The constraints cover agreement within NPs/PPs, subject-verb agreement and probabilistic constraints for case in NPs. This method leads to a small improvement in BLEU.

Morpheme-based translation

Morpheme-based translation approaches are mainly applied to languages with a highly complex morphology, such as Turkish or Finnish, which are among the most difficult to translate language pairs of the languages covered by the Europarl corpus (Koehn, 2005). The general concept of sub-word level translation consists in segmenting surface forms into smaller entities (morphemes) for the translation step, which are then recombined into valid target-language words after translation.

Oflazer et al. (2007) investigate different representational granularities for sub-lexical representation of Turkish, which is an agglutinative language with highly productive derivational and inflectional morphology. The surface words are segmented and represented by lexical morphemes (instead of surface morphemes), to abstract over suffixes that appear to be different due to phenomena such as vowel harmony. This representation considerably reduces the vocabulary size, but it also entails the risk of erroneous analyses. One problem of the fine-grained segmentation is that some morphological markers have no equivalent on the source-side (English), and thus remain unaligned. In a variant of the representation of Turkish, some markers are thus attached to the root, whereas others are represented as separate tokens, for example markers for grammatical case, which often correspond to English prepositions. After decoding, Turkish surface forms are formed by concatenating the segments, and the 1000-best translations are re-ranked using a language model. The proposed method with a selectively segmented representation of Turkish leads to an improved BLEU score over a baseline built on regular surface forms. In a similar approach, Virpioja et al. (2007) use morpheme-like units obtained through unsupervised morphology learning for the Scandinavian languages Swedish, Danish and Finnish. Danish and Swedish are closely related languages, whereas Finnish is considerably different and has a particularly complex morphology. In contrast to Oflazer et al. (2007), both source language and target language are segmented into morphemes with a flexible and language-independent segmentation method. Despite the considerably reduced vocabulary size, there is no improvement in terms of BLEU for any language direction. However, they show that that the out-of-vocabulary rates are improved. Badr et al. (2008) explore a segmentation for an English–Arabic SMT system, which they extend with factors containing POS tags in addition to the surface word (source side) and surface words, stems and POS tags enriched with the segmented clitics (target side). They also emphasize problems occurring at recombining the segmented SMT output, and combine rule-based and statistical strategies to obtain valid surface forms.

A general problem with morpheme-based translation is the inflated sentence length that comes with the segmented representation. As translation models usually operate on a limited local context set by the phrase length and the order of the language model, this is problematic and cannot be sufficiently overcome by increasing the phrase length or by using higher order language models. In contrast to previous work, where morphological modeling is typically performed as pre-/post-processing steps only, Luong et al. (2010) propose to use a hybrid morpheme-word representation that respects word boundaries. This restriction is realized through a modified phrase extraction routine that only extracts morpheme-token phrases that span a sequence of whole words. Furthermore, the MERT optimization is carried out on word-level (and not on morpheme sequences), in order to avoid that a morpheme-based BLEU score results in suboptimal weights due to the inflated sentence length. Joint scoring using language models based on both word and morpheme sequences aims at alleviating the problem of data sparseness, while at the same time considering a larger context with a language model built on word-level. Evaluated on an English–Finnish translation system, the morphologically enhanced translation system outperforms a baseline system. BLEU computed on morphemes suggests that the improvement might be even larger, but is not reflected by BLEU on regular inflected forms; this assumption is also supported by a manual evaluation favoring the morpheme-based system.

Clifton et al. (2011) develop techniques for handling morphological complexity in general by means of unsupervised morphological methods, with Finnish as target language. In contrast to Luong et al. (2010), who restrict their system to sequences of whole words, and thus eliminate morphologically productive phrases, Clifton et al. (2011) allow phrase pairs with "dangling morphemes" to create new words. To recombine the output of the morpheme-based system, they apply morphology prediction as a post-processing step. Their proposed system improves over a baseline built on surface forms, and also outperforms the results obtained by Luong et al. (2010).

Morphology generation

While morpheme-based translation approaches are promising for highly productive agglutinative languages, they might be less well suited for fusional languages exhibiting a greater degree of syncretism, where individual morphemes cannot always be distinguished. An alternative to morpheme-based translation consists in translation on word-level with a component to model target-side inflectional morphology. The general concept consists in translating into some form of simplified representation of the target language with a post-processing step to generate morphology. This type of approach also corresponds to the inflection prediction system presented in this chapter.

A simple variant is presented by Fraser (2009) who explores a strategy that works without explicit inflectional features by employing two translation steps for English–German translation: first, a simplified representation of the target language is obtained by performing a simple suffix elimination (removal of the typical inflectional suffixes *e*, *en*, *n*, *es*, *s*, *em* and *er*). A translation model built on this representation outputs a first, intermediate translation. A second translation model then translates from simplified to fully inflected German, thus removing morphological complexity from the translation step and considering morphology entirely as a target-side problem. Bojar et al. (2010) apply the same strategy to translate from English to Czech, but enrich their model by marking the prepositions with the case they subcategorize. However, both efforts turn out to be ineffective on large data sets.

Other approaches are based on detailed modeling of linguistic features. Minkov et al. (2007) propose a strategy in which inflected forms are predicted for a sequence of target word stems given the source sentence. They do not yet work with a machine translation system, but present a *reference experiment* where aligned pairs of reference translations are used to evaluate the model without the noise of malformed machine translation output. The experiments are conducted for English–Russian and English–Arabic. They define the operations *morphological analysis*, *stemming* and *inflection* in addition to assuming lexicons for the source and target language that support the three operations. A sentence, assumed to be the output of a translation system, can then be *stemmed* and for every stem, an *inflection* is determined from the set of possible inflections for the given stem. To predict the inflection, a probabilistic model conditioned on the sequence of previous predictions is employed. As features, both monolingual and bilingual information is used, which can be further split into lexical, morphological and syntactic features. Lexical features refer to surface words, morphological features are language-dependent and contain information about relevant inflectional features such as number, gender and case, while syntactic features are derived from the parse structure and contain, for example, part-of-speech tags and dependency information, as well as information about determiner (which are relevant for Arabic inflection), or whether a word appears within a compound on the source side. As baseline for the inflection experiment, a random inflection from the set of inflections defined for a stem is used, as well as a word trigram language model considering some local context. For both languages, the prediction using the rich feature set far outperform the two baselines; variants using all monolingual and bilingual features lead to the best results.

In a follow-up paper, Toutanova et al. (2008) successfully apply the method for inflection prediction proposed by Minkov et al. (2007) to English–Russian and English–Arabic SMT systems, using a phrase-based Moses system and a syntactically informed treelet translation system. For the inflection prediction, they compare re-inflecting the output of a system operating on inflected forms to inflecting the output of a stemmed system. For both languages, the tested variants outperform the respective baselines; inflecting the output of a stemmed system was found to be the best strategy.

The idea of a two-step translation process has been adopted for several languages. For example, Fraser et al. (2012) present a system for nominal inflection for for English–German translation, which is the basis for the inflection prediction system presented in this chapter. In contrast to Toutanova et al. (2008), who conflate word formation and inflection, Fraser et al. (2012) considers word formation as a separate problem; in particular, the generation German compounds in the inflection step is addressed (cf. section 3.5.3 for more details on compounds and word formation). Ramm et al. (2016) explore strategies to extend the system with verbal inflection. Furthermore, Tiedemann et al. (2016) present a similar system for translating from English into Finnish, and Burlot et al. (2016) apply the two-step translation process for English–Czech translation.

For English–Arabic translation, Kholy et al. (2012) compare the effect of modeling morphological features as part of the core translation versus morphological generation using a prediction component. This question cannot be answered generally, but is feature-dependent: in their setup, determiners are best handled as part of the core translation, whereas morphological prediction works best for the inflectional features gender and number.

Chahuneau et al. (2013) present a different approach to the generation of inflected forms. Their method consists in first learning a discriminative model to predict target-language inflection given source-side features. Then, using this model, new translation options are created and are added as "synthetic phrases" to the inventory of already existing translation options. The phrase table-augmentation provides the translation model with multiple entries, and the final decision is left to the translation model during decoding. In order to obtain optimal target-language inflections, a rich set of source-side features is used, including part-of-speech information, dependency analyses, and assigning tokens to Brown clusters trained on large monolingual data. The source-side features are bundled with the set of morphological features of the target language, which are represented in form of feature-value pairs such as *tense=past*. Morphological modeling can be handled through supervised techniques (such as an already existing morphological analysis tool) or through unsupervised techniques that do not require any pre-existing resources and are thus applicable to a broader range of languages. Systems augmented with synthetic phrases were tested for the translation directions English–Russian (using supervised and unsupervised morphological modeling), English–Hebrew and English–Swahili (both using only unsupervised morphological modeling). For all translation experiments, they report significant improvements over a baseline. In the case of English–Russian, the variant with supervised morphological modeling leads to better results than the unsupervised system. An extension to using synthetic phrases has been proposed by Huck et al. (2017), who use a discriminative model relying on rich source- and target-side information to support the integration of synthetic phrases for an English–Czech translation system. Furthermore, they define a "tag template" that prevents the generation component from over-generating phrases that that are incompatible with original phrase, for example by preserving the status of negation or tense. They report clear improvements for small- to medium-sized training data, but also a tendency for improvements to diminish with increasing training data size.

Weller-Di Marco et al. (2016) use synthetic phrases for the generation of prepositions and case markers to model subcategorization across complement types, which is discussed in chapter 7. In comparison to Chahuneau et al. (2013) and Huck et al. (2017), the generation concerns closed-class function words, instead of inflectional variants of open-class content words.

The strategies of modeling target-side inflection in a two-step translation setup with a post-processing component or by enriching the phrase-table with synthetic phrases can be considered as *pipeline strategy* (two-step inflection prediction), or as a *joint model* (synthetic phrases) that integrates the generated inflected forms directly into the translation system without a need for post-processing. Both strategies can overcome the bottleneck of data sparsity by being able to generate unseen inflected forms, as well as providing a better basis for the selection of inflected forms by integrating rich source- and target-side features.

Generally, advantages can be named for both strategies: for example, translating on inflected forms optimized for the source phrase to be translated provides more information than the stemmed representation, allowing the translation hypotheses to be scored against an inflected language model which might have more discriminatory power than a stemmed one (such an observation has been reported by Toutanova et al. (2008)). On the other hand, the more general representation of abstract word stems can be beneficial in the case of data sparsity on the inflectional level. Consider, for example, the sentence in (11), where the correctly inflected form of the adjective did not occur in combination with the noun in the language model, thus triggering an incorrect inflection based on another, incorrect bigram observed in the language model[22]. Another difference between the inflection prediction model and the synthetic-phrases model is the focus of features used to determine inflection: the inflected forms used to augment the translation model in the synthetic phrases system are generated conditioned on *source-side features* in combination with strong language models to capture target-side context, whereas the inflection prediction system rather uses *target-side features* for the

[22] Chahuneau et al. (2013) address the problem of uncovered words by using an additional class-based language model; however, this seems to be relevant on a semantic rather than on an inflectional level.

generation of inflected forms. Given that many inflectional phenomena are target-side specific, and often independent from the source side, it seems plausible to emphasize target-side features for the modeling of target-language morphology. However, for some target-side decisions, the source-side needs to be consulted – for example, the values of grammatical case on the target side often correlate with the syntactic function of the corresponding phrase in the source-sentence. Ideally, target-side inflection is thus conditioned on both source- and target-side features – this constitutes an extension to the inflection prediction model presented so far that will be addressed in the next chapter.

Summarizing, it is difficult to say which of the strategies, joint-model or pipeline-model, is superior, as both have strong and weak points for modeling inflection. With regard to the problem of *word formation* though, it seems that the pipeline-strategy provides a broader range of options. While this aspect is not explicitly addressed in Chahuneau et al. (2013), it seems generally possible to model word formation as long as the involved words occur within one phrase. With the pipeline inflection prediction approach, word formation can be modeled also beyond phrase-boundaries, as the inflection decision is made as last step. For example, the decision to merge a determiner and a preposition into a portmanteau preposition depends on the realization of the target phrase – even if the parts of the PP were translated by multiple phrases, the portmanteau preposition can be realized as required by the context, whereas the synthetic-phrases system cannot produce the sequence [*. . . ins auto . . .*] ('in-the car') based on the phrases [*. . . in*][*das auto . . .*], as the generation step is limited to phrases. While portmanteau prepositions might be considered just a minor technical nuisance, the productivity of (German) compounds poses a serious challenge to machine translation – a widely used approach to handle compounding in SMT consists in *building new compounds* by concatenating non-complex words. To build a new compound, the involved components *must* appear within different phrases – otherwise, the compound would be known already. Section 3.5.3 illustrates how productive word formation can be handled with the inflection prediction-system.

3.5.3 Word-formation and terminology

Compounding is a highly productive word formation process in many languages and represents a challenging problem to machine translation: compounds not covered by the parallel training data remain untranslated. However, the parts of a compound often occur in the training data, and can be used to translate a compound on the source-side, or as "building blocks" to construct new compounds when translating into a compounding language.

As in the previous sections, the strategy for translating *from* a compounding language consists in making the source side simpler by means of compound splitting, whereas translating *into* a compounding language is more complicated as it also requires the generation of (inflected) compounds.

Koehn et al. (2003) are the first to handle compounds in SMT. Translating from German into English, they compare empirical methods for compound splitting to make the parts of the compounds accessible to the translation system. Compound splitting on the source side makes the parts of a compound accessible to the translation model, and as a result, more compounds can be translated. They use a frequency-based approach to compound splitting that consists in splitting a compound into a combination of substrings attested in the training data. Their experiments suggest that an "eager" splitting method leads to the best translation results, as phrase-based SMT tends to forgive over-splitting by simply translating a split unit as a phrase. Yang et al. (2006) employ a different approach to compound handling in SMT by using a backoff model to translate unseen words. For unknown words in the test data, phrase-table entries with a more general representation are generated, by first considering a stemmed form, and then backing off to a split variant of the word. While their method is not targeted specifically at compounds, but at unknown words in highly inflected languages in general, compounds make up a large part of the words treated by their approach. They report an improvement for the translation directions Finnish–English and German–English.

For translating into a compounding language, Stymne et al. (2008) employ a factored model containing special "compound POS tags" to mark compound modifiers and compound heads, based on which compounds can be merged in a post-processing step. Stymne et al. (2011a) extend their previous approach with a CRF sequence model to find better merging points. Fraser (2009) adds inflectional morphology to compound handling for an English–German SMT system by combining two translation models: first, the English input is translated into a stemmed and split German intermediate representation. On this output, stem sequences corresponding to compounds are merged, including the insertion of transitional elements. Then, a second German-to-German translation system translates the stemmed representation into inflected German.

Cap et al. (2014) present an approach that combines the generation of compounds with inflectional morphology by making use of the rich annotation provided by SMOR. Their approach is implemented in an inflection prediction system based on Fraser et al. (2012): the translation model is built on a stemmed representation in which compounds are split based on SMOR's analysis. The post-processing step first merges compounds using rich source- and target-side features to identify optimal merging points for sequences corresponding to compounds, and then applies the inflection step on the merged and stemmed representation. While this approach does not lead to an improvement in terms of BLEU, a manual evaluation reveals that the translation quality of compounds is improved.

The problem of unknown words is amplified when translating data of special domains, in particular when morphologically complex languages are involved. Modeling target-side inflection in SMT has been considered useful in domain adaptation tasks. For example, Hálek et al. (2011) integrate translations of named entities mined from Wikipedia into an English–Czech translation system. Inflectional morphology is approximated by providing all inflected forms in combination with the estimated translation probability. Formiga et al. (2012) model verb inflection in a domain-adaptation scenario for English–Spanish translation. Wu et al. (2008) use dictionary entries to adapt an English–French translation system to a new domain, but do not apply morphological modeling to ensure consistent inflection. Weller et al. (2014a) propose a morphology-aware strategy to integrate bilingual terminology obtained from comparable corpora into English–French SMT systems, comparing methods to integrate term-translation pairs into general language translation systems. The morphology-aware system used in this paper is a re-implementation of the inflection prediction system described in this chapter, adapted to the language pair English–French.

3.6 Transferability to other languages and resources

The process of inflection prediction in SMT as presented in this chapter is tailored to the language pair English–German with the particular resources BitPar and SMOR. However, the setup is organized in a way that the linguistic resources can be exchanged. This section briefly outlines the roles of the linguistic resources used, and how individual components can be substituted when adapting to another language.

Annotation of morphological features The basis to create the stemmed representation for the translation system and the training data for the feature prediction models is data annotated with the relevant morphological features. In the system setup as described in the previous section, the output of the constituent parser BitPar is used. However, morphological tags can also be obtained by any other parser or morphological tagger. For this work, BitPar was selected as its lexical resources are tied to SMOR.

Morphological analysis and generation A crucial part of the inflection prediction approach is the morphological analysis and generation: while many steps in the presented system setup are implemented with a view to SMOR (such as mapping 'standard features' to 'SMOR-internal features'), most of this is handled externally in a way that a specific look-up table for both *surface-form → stem* and *stem+features → inflected forms* are built (see table 3.5 for an example of the table to look up inflected forms for a stem+feature combination). If such tables are created based on alternative resources, they can be integrated in the inflection prediction approach with only minor modifications to the current implementation.

Set of inflectional features Adapting the inflection prediction approach to another language likely also changes the set of features to be modeled. The core of the inflection prediction system is independent from a fixed set of inflectional features. Rather, the set of part-of-speech tags that are subject to inflection prediction and the expected stem-markup per tag are defined in a configuration file, and can thus also – theoretically – be defined for other languages.

English–French inflection-prediction Experiments to model another language pair have been made with English–French inflection prediction (cf. Weller et al. (2013a) and Cap et al. (2015a) for system descriptions of submissions to the WMT Translation Shared Task). In contrast to German inflection, French nominal morphology only has the features number and gender, which makes the modeling an easier task. While the resulting inflection prediction system is able to improve over a surface-form baseline trained on a small parallel corpus, the improvement disappears when training the systems on larger data sets. It has to be noted, though, that the English–French parallel data is considerably larger than the English–German parallel data. The large training data size coupled with the easier prediction task might be the reason why the inflection prediction approach does not work as well for translating into French.

3.7 Summary

This chapter presented a morphology-aware translation system that employs a two-step approach to separate the translation process from the generation of target-side morphology: first, a translation system trained on an underspecified representation enriched with selected inflectional features produces an intermediate translation output. To inflect the stemmed translation output, the set of inflectional features is predicted and the external morphological tool SMOR generates inflected forms for the stem-feature pairs in a post-processing step.

Modeling target-side inflection as post-processing allows to address three main problems when translating into a morphologically complex language: with the underspecified representation, the model can generalize better at training time; inflectional features are predicted in sentence context, providing a sound basis to generate correct inflection; and the use of a morphological resource enables the system to generate inflected forms independently from their occurrence in the parallel training data.

In the presented experiment, modeling target-side morphology leads to improved results over a baseline system operating on surface forms, even though the difference in BLEU is not always significant. Looking at translation examples suggests that there is evidence for the system to benefit in all three categories: there are novel inflections, even though they seem to have a comparatively small effect, at least when measured in BLEU. On the other hand, the ability to generate the *correct* inflection is likely to contribute more to the overall improvement – a generated form may not be novel at word-level, but each inflection is tailored to a specific and potentially new context. Furthermore, the generalization effect is not restricted to estimate better translation statistics, but also turns out to be useful at phrase-level in the language model.

The prediction of inflectional features works well on a *local level* by establishing feature agreement within phrases. However, the prediction models so far are restricted to local context only. This is not problematic when predicting features that are already set by the stem markup (gender and number), or that can be derived locally from the phrase (strong/weak inflection). However, a locally restricted model does not provide a sufficient basis for the prediction of grammatical case, which depends on many factors on source and target side, and often cannot be derived from a local context, but needs to take into account information from the entire sentence. Thus, improving the prediction of case by extending the prediction model with rich source- and target-side features is the main focus of the following chapter.

In addition to the straightforward benefit of improving the translation quality, the abstract representation of the target-side data holds many possibilities for a focused modeling of inflectional features: Besides handling target-side features more efficiently in the feature prediction model in the generation step, the abstract target-language representation allows to integrate different types of externally obtained (linguistic) information, in particular with regard to grammatical case, which will be explored in the later chapters of this thesis.

Chapter 4

Improving the Prediction of Case Using Rich Features on Source- and Target-Side

While the inflection prediction system improves over a baseline trained on standard surface form data, the feature prediction step is restricted to local target-side context. This restriction is problematic for the prediction of grammatical case. Case indicates the syntactic function of a phrase in the sentence (such as subject or object), and in comparison to the other inflectional features, thus also carries a semantic dimension. To model grammatical case, both source-side and target-side features need to be consulted: by looking at the source-side, the syntactic function of a phrase in the input sentence can be transferred to the target-side and realized accordingly when inflecting the stemmed translation output. Taking into account target-side subcategorization information at the same time ensures that target-side requirements are met. In this chapter, source-side features are combined with subcategorization information obtained from an external database with the objective to improve the prediction of case. While there is no significant improvement in BLEU, a manual evaluation confirms that the enriched case prediction model leads to improvements. The settings and experiments presented in this chapter are published in Weller et al. (2013b).

4.1 Motivation

Translating into morphologically rich languages is challenging for two main reasons: the large inventory of inflected forms creates a *data sparsity problem* and insufficient access to information necessary to ensure a correct target-side inflection leads to the *problem of selecting the correct inflection*.

The inflection prediction system presented in the previous chapter addresses the problem of data sparsity by generating inflected forms, and thus the ability to produce forms not occurring in the parallel training data. A target-side sequence model looking at local sentence context ensures that agreement restrictions are satisfied. The modeling of number, gender and strong/weak inflection of NPs and PPs is straightforward and usually restricted to a local context. In contrast, grammatical case reflects the syntactic function of a phrase and thus also requires global information from the entire sentence, making case a feature that is more difficult to predict. In addition, the syntactic function of a phrase also depends on the meaning expressed in the source sentence and often cannot be decided by looking at the translation alone. This relation to the source side has been ignored so far; only the feature number is (indirectly) transferred from the source side by dividing the translation options into singular translations and plural translations through the annotation in the stem markup. However, this annotation strategy cannot be applied to model case: As the function of an NP is determined by the source sentence, and its realization in terms of grammatical case further depends on the choice of target-side verb and its subcategorization frame, a stem markup prior to translation does not make sense. Marking a noun as *accusative*, for example, would make it unavailable for any other function, and thus counteract the aim of providing a general representation. The modeling of case can thus not be handled by enriched stem markup during translation, but is instead addresses in the prediction step. As the stemmed representation used for translation is underspecified with respect to syntactic functions, it is possible to abstract over individual forms and create inflected target-side forms as needed. However, this also means that there are no "clues" for the distribution of syntactic functions in the stem sequence, and the case prediction model depends on local context information for the assignment of case.

This chapter presents an extension to the inflection prediction step that takes into account richer information on source and target side to improve the prediction of grammatical case:

- **Source-side syntactic information** aims at ensuring that the function of a phrase as expressed in the source sentence is correctly reproduced in the target-sentence. Since the stemmed representation is underspecified with regard to syntactic functions, access to the source-side syntactic features is necessary if a phrase's function cannot be clearly derived from its target-side context.

- **Target-side selectional preferences** take into account target-side restrictions such as those imposed by verbal subcategorization frames. This is of particular importance when the verb and its arguments are not adjacent. In addition, statistics about *genitive modification*, the German equivalent to *noun-of-noun phrases*, aim at modeling modifiers[1] in contrast to subcategorized constituents.

German is a less configurational language than English, where syntactic functions are expressed through the position of the constituent in a sentence. For example, an English subject typically occurs at the beginning of a clause, followed by verb and object, thus adapting a Subject-Verb-Object (SVO) ordering. In contrast, German word order is more flexible and the syntactic function is expressed through grammatical case. This makes case a particularly important feature, as it contributes to the understanding of the sentence. A wrong value of case may thus have a direct impact on the perception of *adequacy*, whereas errors concerning the other features relevant for nominal inflection rather impact the perception of *fluency*. The difficulty of predicting case is furthermore increased by the fact that several, potentially complex noun phrases (cf. the example in figure 3.1), can occur as a long sequence where it is not immediately clear whether a phrase is subcategorized, or a modifier of a subcategorized phrase. The types of confusion of syntactic functions can be roughly categorized into

- mistaking an argument subcategorized by the verb for a non-subcategorized phrase or vice-versa: this mostly concerns two adjacent noun phrases, where the second NP can either be a modifier (in genitive case) of the first NP, or both NPs can be subcategorized;
- confusing the the types of arguments: direct object vs. indirect object vs. subject.

The examples below, produced with the inflection-prediction system presented in section 3.4.3, illustrate these error types in case prediction.

(1) in rare instances of low visibility , the crew will instruct passengers to turn off their devices during landing .

in seltenen Fällen von geringer Sichtbarkeit , die Crew **der Passagiere**$_{Gen}$ anweisen ,
in rare cases of low visibility , the crew ***of-the passengers*** *instruct ,*

ihre Geräte während der Landung ausschalten .
their devices during the landing turn-off .

[1] While there can be many phrases that are *not* subcategorized, the confusion between NPs modifying a preceding NP vs. two adjacent, subcategorized NPs, such as $[\mathrm{NP}]_{Dir.OBJ}$ $[\mathrm{NP}]_{Ind.OBJ}$, has been found to be particularly challenging, and is thus addressed explicitly in the prediction model.

The sentence in (1) is sufficiently well translated in terms of stems, and a correct inflection would have resulted in a well understandable, although not perfect translation. However, the inflection of the noun phrase *der Passagiere* ('of-the passengers') as a genitive modifier of *die Crew* ('the crew') instead of a direct object makes the understanding of the translation more difficult than necessary. The phrase *die Crew der Passagiere* is well-formed with regard to agreement constraints, but makes no sense in this constellation. One could assume that *die Crew* often precedes a genitive modifier, such as the *die Crew der Lufthansa*, whereas *Passagiere* makes a generally plausible genitive modifier, such as *Rechte der Passgiere* ('rights of-the passengers'). Clues for the correct inflection, *[die Crew]*$_{SUBJ}$ *[die Passagiere]*$_{Dir.OBJ}$, can be found both on the source and on the target side: for both English phrases *the crew* and *the passengers*, the syntactic function is preserved in the translation. On the target-side, the subcategorization frame of the verb *anweisen* ('instruct') requires a subject and a direct object which need to be filled by the constituents *Crew* and *Passagiere*. The proposed inflection of *Passagiere* as genitive modifier contradicts both the source-side structure and the target-side restrictions. Looking at the source-side syntactic constellation in combination with target-side subcategorizational requirements should thus provide the relevant clues for case prediction.

(2) kenya media drew the ire of authorities by broadcasting security camera footage of troops who were dispatched to the scene of the attack purportedly robbing the upmarket mall .

Kenia Medien zog den Zorn **die Behörden**$_{Acc}$ die Sicherheit Kamera Aufnahmen von
Kenya media drew the ire ***the authorities*** *the security camera footing of*

Truppen , die zum Schauplatz des Angriffs angeblich raubt die vornehme Mall .
troops , who to-the scene of-the attack purportedly robs the upmarket mall .

The sentence in (2) is less straightforward as the previous example in (1), as its translation on the level of stems is already flawed. The main difficulty to predict case in this sentence consists in assigning the respective function to each phrase in the sequence of three noun phrases *[den Zorn] [die Behörden] [die Sicherheit Kamera Aufnahmen]*, of which the last one is particularly complex and should contain only one compound (*Sicherheitskameraufnahmen*) instead of three individual nouns. Here, the phrase *die Behörden* ('the authorities') is inflected as direct object in accusative case, instead of being realized as a genitive modifier of the object phrase *Zorn* ('ire'). While the prediction of case is difficult due to the complexity of the sentence, looking at the source side and consulting target-side preferences can again help for a better prediction of case: first, the source-side construction (*the ire of authorities*) corresponds to the correct target-side realization as a genitive phrase. Furthermore, the genitive phrase *Zorn der Behörden* is more plausible than two direct objects next to each other.

(3) the role-play ended with the words " a restless spirit finally found rest and returned home forever " .

das Rollenspiel endete mit den Worten " **einem unruhigem Geist**$_{Dat}$ endlich Ruhe
*the role-play ended with the words " **to-a restless spirit** finally rest*

gefunden und nach Haus zurückgekehrt " .
found and to home returned " .

The sentence in (3) shows a confusion between indirect object (predicted) and subject (correct). While the sentence is acceptably translated at the level of stems, the inflection of *einem unruhigen Geist* ('to-a restless spirit') considerably changes the translation for worse. While the sequence *[NP]*$_{Ind.OBJ}$ *[etwas Ruhe]*$_{Dir.OBJ}$ *[VERB]* is possible (and not too far off from the source-side meaning) with a verb such as *gönnen* ('to allow: to allow NP some rest'), the verb *finden* (to 'find') typically does not occur with an indirect object[2], but instead subcategorizes a subject and a direct object. Additionally, the syntactic function on the target-side corresponds to the one on the source side.

The previous examples showed problems at the case prediction step, but also illustrated how access to the source side and information about target-side subcategorization frames and selectional preferences can help to improve the prediction of case.

4.2 Combining source-side and target-side information for the prediction of grammatical case

This section demonstrates that features obtained from both the source and the target side cannot only be helpful, but contribute complementary information, and often are indispensable to sort out syntactic functions in a stemmed sentence.

4.2.1 Information on source-side syntactic functions

In many sentences, a human reader can figure out the semantic role of each constituent based on general world knowledge and common sense. For example, it is generally assumed that a *cake* is *eaten* by a *person* rather than the other way round: there is a strong

[2] A construction with an additional indirect object in the role of a beneficient such as *Ich habe dir das Buch gefunden* ('I found you the book': 'I found the book for you') is possible, but somewhat unusual.

preference for a *person* to be the *agentive* subject[3] when co-occurring with the verb *to eat*. Similarly, a word such as *cake* is highly likely to be the object (in this case the *patient*) of this verb. In some (stemmed) sentences, however, there is not enough context to determine the semantic roles without looking at the source side, as illustrated in (4):

(4) ... `weil[KOUS] die[ARTdef] Minister<+NN><Masc><Sg>[NN] die[ARTdef]`
... *because the minister the*

`{Merkel}<TRNC>Regierung<+NN><Fem><Sg>[NN] dazu[PROAV] aufruft[VVFIN]` ...
Merkel-government to-this urges ...

Looking at the stemmed sequence, it is straightforward to assume that *the minister* is the subject. However, it cannot be decided based on the target-side context alone whether the phrase *the Merkel-government* is a direct object (inflected in accusative case: *the minister urges the Merkel-government*), or a modifier to the preceding phrase (inflected in genitive case: *the minister of the Merkel-government*), resulting in the two inflected variants shown in (5) and (6).

(5) ... weil der Minister [**die Merkel-Regierung**]$_{\textbf{Dir.OBJ}}$ dazu aufruft ...
... *because the minister to-the Merkel-government to-this urges* ...
'... because the minister urges the Merkel-government to ...'

(6) ... weil der Minister [**der Merkel-Regierung**]$_{\textbf{Gen.MOD}}$ dazu aufruft ...
... *because the minister of-the Merkel-government to-this urges* ...
'... because the minister of the Merkel-government urges to ...'

In this constellation, the subcategorization frame of the verb *auffordern* provides no help to assign syntactic functions – here, the direct object is optional, and the phrase *the Merkel-government* is plausible as both modifier and direct object. In such cases, only source-side information can help to predict the correct value of grammatical case: transferring the syntactic function of *Merkel-government* to the target-side and using it as additional feature to enrich the case prediction model. A general advantage in using source-side features is the fact that the source side is intact and can be parsed. This is not the case for the translation output which is likely to be ungrammatical and thus challenging to parse.

[3] While syntactic functions and semantic roles are not the same, they are closely related. For the modeling of grammatical case described in this chapter, a clear distinction between the two is not necessary; and the new features integrated into the prediction model entail both syntactically (source-side functions) and semantically (lexical co-occurrence) motivated information.

While the English side provides *some* clues, for example to distinguish between argument and modifier, it is not sufficient as a basis for case prediction on its own: German has four cases, whereas English does not explicitly mark case (save for some pronouns, or the occasional genitive marker *'s*), but rather uses the position in the sentence or prepositions to express semantic roles.

Assuming that the translation step keeps the general sentence structure intact during translation, it can be expected that the **subject** in the source sentence corresponds to the target-sentence. Similarly, a German **genitive modifier** (cf. example (6)) is often expressed as a *noun-of-noun* phrase in English, and thus provides a sound basis to identify such adjunct phrases.

In contrast, the English side offers only sparse clues to distinguish between German **direct** and **indirect objects**, as English does not differentiate between dative and accusative. As a rule of thumb, German direct objects (accusative case) often correspond to English direct objects. For German indirect objects (dative case), it less straightforward: for ditransitive verbs, such as *to give*, the indirect object can be expressed by a second noun phrase: *Bob gave Alice a present*. However, it is also possible to realize the second object in form of a prepositional phrase: *Bob gave a present to Alice*. Furthermore, many verbs subcategorizing an indirect object in German use a direct object on the English side – thus, the English sentence does not help much to differentiate between accusative and dative, as illustrated in example (7):

(7)
er sieht ihn$_{Acc}$ 'he sees him'
er glaubt ihm$_{Dat}$ 'he believes him'

Such subcategorizational requirements are independent from the source side, and can only be solved with target-side information.

4.2.2 Target-side subcategorization frames

The need for explicit target-side subcategorizational information is motivated by several reasons: first, German verbs can subcategorize a direct and/or indirect object, and the *subcategorization frame* of a verb outlines the requirements of the respective verb. Second, in addition to the "bare" subcategorization frame of a verb, additional *information at the syntax-semantic interface* may be necessary in order to correctly assign phrases to their roles in the sentence without violating selectional preferences. Furthermore, the verb and its subcategorized elements may be separated by large gaps such that a sequence-based model does not have access to all relevant parts; extending the prediction model with richer features helps to "bridge" such gaps and to connect the relevant pieces of information through the sentence. This is illustrated by the following set of examples: The sentences (9), (10) and (11) all contain the stem sequence shown in (8), yet they need different inflections depending on the choice of the verb.

(8) `[ART Mitarbeiter] [ART Bericht] [ART Kollegen] <Verb>`
ART employee ART report ART colleague <Verb>

(9) [Der Mitarbeiter]$_{SUBJ}$ hat [den Bericht]$_{\mathbf{Dir.OBJ}}$ [dem Kollegen]$_{\mathbf{Ind.OBJ}}$ gegeben.
The employee has the report to-the colleague given.

(10) [Der Mitarbeiter]$_{SUBJ}$ hat [dem Bericht]$_{\mathbf{Ind.OBJ}}$ [des Kollegen]$_{\mathbf{Gen.MOD}}$ zugestimmt.
The employee has on-the report of-the colleague agreed.

(11) [Der Mitarbeiter]$_{SUBJ}$ hat [den Bericht]$_{\mathbf{Dir.OBJ}}$ [des Kollegen]$_{\mathbf{Gen.MOD}}$ gelesen.
The employee has the report of-the colleague read.

Looking at the subcategorization frames of the verbs *geben* ('to give'), *zustimmen* ('to agree') and *lesen* ('to read'), a likely way for inflecting the sentences becomes obvious: *geben* (9) has a strong preference for a ditransitive subcatgorization frame containing an agentive subject (nominative case), a benefactive (dative case) and a patient (accusative case) – thus, all three noun phrases in the sentence are needed to fill the three slots. In contrast, *zustimmen* (10) has a strong preference to select only an agentive subject, and an indirect object theme (dative case). Thus, only two NPs can receive case from the verb, whereas the third NP is instead a genitive modifier of the dative object. Similarly, the verb *lesen* (11) only requires two arguments, a subject and a direct object (accusative case), leaving the third NP to be a genitive modifier. The general subcategorization frame thus provides information about the number of arguments (intransitive, transitive or ditransitive), as well as the type of object (direct/indirect).

The following examples illustrate that information about the bare subcategorization frame is not always sufficient for the prediction of case, but that the prediction model needs to be enriched with information about the syntax-semantic interface.

(12) [Der Mitarbeiter]$_{\mathbf{SUBJ}}$ hat [dem Kollegen]$_{\mathbf{Ind.OBJ}}$ [den Bericht]$_{\mathbf{Dir.OBJ}}$ gegeben.
The employee has to-the colleague the report given.

(13) [Der Mitarbeiter]$_{\mathbf{SUBJ}}$ hat [den Bericht]$_{\mathbf{Dir.OBJ}}$ [dem Kollegen]$_{\mathbf{Ind.OBJ}}$ gegeben.
The employee has the report to-the colleague given.

(14) [Dem Kollegen]$_{\textbf{Ind.OBJ}}$ hat [der Mitarbeiter]$_{\textbf{SUBJ}}$ [den Bericht]$_{\textbf{Dir.OBJ}}$ gegeben.
To-the colleague has the employee the report given.

(15) [Den Bericht]$_{\textbf{Dir.OBJ}}$ hat [der Mitarbeiter]$_{\textbf{SUBJ}}$ [dem Kollegen]$_{\textbf{Ind.OBJ}}$ gegeben.
The report has the employee to-the colleague given.

The sentences (12)–(15) represent variations of the sentence *the employee has given the report to the colleague*. While in (14) and (15) the first constituent is stressed, all variants are natural. As German word order is flexible, such permutations occur regularly, and as a result, the position of a noun phrase is no reliable indicator for the semantic role of the phrase, even though subjects tend to appear at the beginning of a clause. In each sentence, the verb and the participating noun phrases are identical. Without information in the stemmed output, the prediction of case relies on local context and simple ordering statistics such as that a subject is more likely to appear at the beginning of a clause rather than at the end. Here, semantically motivated information that goes beyond mere syntactic requirements can help to derive an appropriate assignment of case. For example, both noun phrases *Mitarbeiter* and *Kollege* would satisfy the agentive subject role of the verb *geben* better than *Bericht*, which is rather fitting for the role of the patient. However, given that both *Mitarbeiter* and *Kollege* are appropriate for the roles of subject and recipient in this setting, semantic information alone is not sufficient, but needs to be combined with source-side information to offer the full set of features necessary for case prediction.

Finally, the sentence in (16) demonstrates a further motivation for enriching the case prediction model, that is based on a rather practical reason: unlike the other nominal features which are locally constrained within one noun phrase, grammatical case requires global information from the entire sentence. As a result, long gaps between the verb and its arguments further add to the difficulty of case prediction.

(16) Er hat dem **Bericht** des Kollegen nach kurzem Nachdenken vorbehaltlos
He has to-the report of-the colleague after short thinking unconditionally

zugestimmt
agreed

Here, there are six words between the indirect object *Bericht* and the verb *zustimmen*, and even more between subject and verb. Thus, while the subject at the beginning of the sentence is likely to be predicted correctly by the model, the decision between direct and indirect object for *Bericht* is not possible without access to the verb.

4.2.3 Combining source-side and target-side information

This chapter aims at improving the modeling of case in an inflection-prediction system by combining two strategies to integrate information at the syntax-semantic interface.

First, the syntactic functions of source-side noun phrases are projected to the respective target-side noun phrases in the translation output, to allow the prediction model to make an informed decision when several phrases are plausible candidates for a particular argument role. Second, an external knowledge base containing information about German verb subcategorization frames and noun-noun modification provide a basis to comply with target-side restrictions. The subcategorization information models association strength between verb-noun pairs and noun-noun tuples. Such a database can thus tell the prediction model that, for example, the verb *zustimmen* is likely to subcategorize only an indirect object in addition to the subject, by having evidence for many objects in dative case, while no (or only very few) objects in accusative case are attested in the database. This functionality corresponds to the *subcat frame prediction*. Furthermore, the verb-noun co-occurrences listed in the database are informative on a semantic level: the association strength between *Bericht* as direct object with the verb *lesen* can be expected to be higher than *Bericht* as a subject, and thus triggers a prediction as direct object. Similarly, noun-noun$_{Gen}$ pairs provide a basis to identify genitive modifiers, and thus discourage the generation of absurd modifier constellations as *Crew der Passagiere* ('crew of-the passengers'), as observed in example (1).

The two strategies are complementary, as the two types of information aim at different angles of the case prediction problem. Combining these two strategies approaches a simplified level of semantic role definition, but is based only on co-occurrence data extracted from dependency-parsed corpora.

4.3 Enriching the case prediction model

This section explains how source-side information is projected onto the target-side, and how the database containing the verb-noun tuples and the noun-noun$_{Gen}$ modifier pairs are obtained and integrated into the case prediction model.

4.3.1 Hierarchical translation model

This experiment is based on a hierarchical translation system (Chiang, 2005), whereas the previous chapter used a phrase-based translation system. This is motivated by the fact that the output of the translation system is a tree containing structural information spanning over the entire sentence, and allows to identify, for example, non-adjacent verb-object pairs if they are spanned by a VP in the tree structure.

Instead of translating flat phrases, a hierarchical system extracts translation rules which allow the decoder to provide a tree spanning over the translated sentence (Galley et al., 2004). The rules are extracted from parsed data, and must respect the phrase boundaries set by the parser, which restricts the number of extractable rules. The presented system uses a *string-to-tree* setting where only the target side is parsed, while the source side consists of flat text – such as system setup is less restrictive, and the tree structure is only really relevant for the translation output. The extracted translation rules are of the following form, with generic X-nodes on the source side and linguistically motivated nodes derived from a parse tree on the target side:

$\texttt{[X]}_1\ \texttt{allows}\ \texttt{[X]}_2 \rightarrow \texttt{[NP]}_1\ \texttt{[NP]}_2\ \texttt{erlaubt}$

$\texttt{[X]}_1\ \texttt{allows}\ \texttt{[X]}_2 \rightarrow \texttt{[NP]}_1\ \texttt{erlaubt}\ \texttt{[NP]}_2$

The example shows how rules can stretch over larger spans by representing a noun phrase by a non-terminal symbol. Furthermore, the two rules illustrate how different word orderings are more efficiently covered; cf. also section 2.2 for more details.

Huang et al. (2006) showed that a rich annotation of the nodes used in translation improves translation quality, in particular annotating a PP node with the lemma of the preposition it contains. Following their suggestion, the training data is annotated with case on NPs and case+preposition on PPs. Figure 4.1 depicts a target-side sentence taken from the training data for the hierarchical model.

4.3.2 Subcategorization information

This section presents the extraction of subcategorization information and its integration into the case prediction model.

Extracting different types of subcategorization information

Verbs play a central role for the structure and meaning of sentences and discourse. Verb information, such as subcategorization frames and statistics about preferred arguments, can provide an interface for the syntactic and semantic level for generating the correct inflection in a translated sentence. While subcategorization statistics can be derived from text corpora, they are not directly accessibly but need to be extracted based on non-trivial methods, ideally relying on a parsed representation.

```
<tree_label="top">
  <tree_label="s">
    <tree_label="np_nom">
      <tree_label="art">
        die{+ART}{Def}{ARTdef}                 der           the
      </tree>
      <tree_label="nn">
        Klima{NN}Wandel{+NN}{Masc}{Sg}{NN}     Klimawandel   climate change
      </tree>
    </tree>
    <tree_label="vvfin">
      unterscheidet{VVFIN}                     unterscheidet differentiates
    </tree>
    <tree_label="prf">
      sie{+PPRO}{Refl}{3}{o}{Sg}{PRF}          sich          (Refl.Pron.)
    </tree>
    <tree_label="pp_appr_in_dat">
      <tree_label="appr_in">
        in{APPR-in-Dat}                        in            in
      </tree>
      <tree_label="piat">
        zweierlei{+INDEF}{Pro}{Invar}{PIAT}    zweierlei     two
      </tree>
      <tree_label="nn">
        Hinsicht{+NN}{Fem}{Sg}{NN}             Hinsicht      respects
      </tree>
    </tree>
    <tree_label="pp_appr_von_dat">
      <tree_label="appr_von">
        von{APPR-von-Dat}                      von           from
      </tree>
      <tree_label="adja">
        ander{+ADJ}{Pos}{ADJA}                 anderen       other
      </tree>
      <tree_label="nn">
        Umwelt{NN}Problem{+NN}{Neut}{Pl}{NN}   Umwelt        environmental
      </tree>                                  problemen     problems
    </tree>
  </tree>
  <tree_label="punc.">
    .{$.}                                      .             .
  </tree>
</tree>
```

FIGURE 4.1: Example of target-side structure in the training data for the hierarchical translation model. English sentence: 'Two factors differentiate global climate change from other environmental problems.'

For example, Scheible et al. (2013) describe the tool *SubCat-Extractor* that obtains subcategorization information from dependency-parsed German corpora. It relies on a set of detailed rules to extract various aspects of verb and complement information. In particular, all complements of a verb are extracted and stored with relevant information, such as lemma, part-of-speech and position in the sentence. Similarly, prepositional phrases occurring in combination with verbs are collected and stored with the respective information. From such a database, verb-object pairs and similar constructions can easily be extracted.

For the case prediction, two types of subcategorization information are applied to model *verb-noun subcategorization case prediction* and *noun-noun modification prediction*. The more general *subcat frame prediction* is considered as implicit in verb-noun information which rely on specific frames.

Verb-noun subcategorization case prediction: Verb-noun tuples referring to specific syntactic functions within the subcategorization frame (subject, direct/indirect object) are integrated with an associated probability for the respective function, based on the observed frequencies of the respective verb-noun$_{Case}$ tuples. In addition to subject and object noun phrases, the subcategorization database includes information about verb-preposition-noun triples, in order to predict the case of prepositional phrases. Since many prepositions subcategorize only one grammatical case, this is only relevant for prepositions that allow for two cases, such as *in*, which can be used locationally (dative case) or directionally (accusative case).

Noun-noun modification prediction: In addition to handling subcategorized noun phrases, it is also important to identify modifying noun phrases that should not be mixed up with noun phrases contained in the subcategorization frame. Typically, such modifiers are genitive noun phrases modifying a preceding noun phrase, for example [*die Mitgliedstaaten*] [*der Europäischen Union*]$_{\textbf{Gen}}$ ('the member states of the European Union'), where the genitive phrase often corresponds to an English *of*-phrase. To this end, noun-noun$_{Gen}$ tuples with their respective corpus frequencies are integrated in order to pass on preferences for a specific function.

Example: subcategorization and modifier prediction

The probabilities for the respective syntactic functions (for verb and noun tuples) are computed based on the occurrence frequencies of the three observed values *Acc*, *Dat* and *Nom*. For the noun-noun$_{Gen}$ modifiers, the co-occurrence frequencies are used as basis for enriching the case prediction model.

noun+verb tuple		**Nom**		**Acc**		**Dat**	
		[freq]	**[%]**	**[freq]**	**[%]**	**[freq]**	**[%]**
Vertrag verbieten	*contract prohibit*	55	91.67	4	6.67	1	1.67
Vetrag besagen	*contract state*	57	98.28	1	1.72	0	0
Vertrag aufsetzen	*draft contract*	0	0	25	100	0	0
Vertrag unterschreiben	*sign contract*	153	11.45	1178	87.95	5	0.6
Vetrag prüfen	*verify contract*	8	10.26	70	89.74	0	0
Vetrag anfechten	*context contract*	7	12.28	49	85.96	1	1.76
Vertrag erfüllen	*fulfill contract*	79	24.31	244	75.08	2	0.61
Vertrag verlängern	*renew contract*	182	21.59	652	77.34	9	1.07
Vertrag entnehmen	*refer to contract*	0	0	1	5.88	16	94.12
Vertrag zustimmen	*approve contract*	6	2.43	9	3.64	232	93.93

TABLE 4.1: Verb-noun tuples for the noun *Vertrag* ('contract') with statistics for the respective functions *subject* (nominative), *direct object* (accusative) and *indirect object* (dative).

noun-noun$_{Gen}$		**freq**
Aufhebung Vertrag	*withdrawal of contract*	163
Vertrag Union	*contract of union*	152
Öffnung Markt	*opening of market*	1140
Markt Zukunft	*market of future*	188
Markt Industrieland	*market of industrial country*	52
Markt Wertpapierbörse	*market of stock exchange*	33

TABLE 4.2: Examples for NP modification as noun-noun$_{Gen}$ structures.

Tables 4.1 and 4.2 show the kind of extracted tuples, and how they represent preferences for particular syntactic functions. For example, the combination *Vertrag+besagen* ('contract + state') almost exclusively occurs with *Vertrag* as subject; this can be attributed to the fact that *besagen* is an intransitive verb and does not subcategorize further arguments. Similarly, the pair *Vertrag+entnehmen* ('refer to contract') has a strong preference for *Vertrag* as indirect object. There can be more variation between subjects and objects, as can be observed in the middle part of table 4.1: while there usually is a strong tendency for accusative in the examples showed, nominative case can also occur for some verbs. This might be partially due to incorrect parse analyses as nominative and accusative often have equal surface forms, but the use of *Vetrag* as subject of the respective verb is often plausible, even though not dominant. For example, for *Vertrag+erfüllen* ('fulfill contract'), both roles are equally plausible, as exemplified in examples (17) and (18), even though there is a general preference for accusative.

(17) *... when the estimated costs to fulfill a contract exceed ...*

(18) *... even when a contract fulfills these form requirements ...*

For the prediction of case, it is not necessarily important to have a clear preference for one value, but the underlying idea is rather to provide the system with a set of several features that point the prediction into the right direction. For example, consider the sequence [ART *Vertrag*] [ART *Union*] ('[ART contract] [ART Union]') in the two following (constructed) sentences:

(19) ... besagt [ART Vertrag] [ART Union] ...
... *states [ART contract] [ART Union] ...*

... besagt [der Vertrag]$_{\textbf{Nom}}$ [der Union]$_{\textbf{Gen}}$ ...
... *states the contract] of-the Union] ...*

(20) ... überträgt [ART Vertrag] [ART Union] [ART Kompetenz] ...
... *transfers [ART contract] [ART Union] [ART competence] ...*

... überträgt [ART Vertrag]$_{\textbf{Nom}}$ [ART Union]$_{\textbf{Dat}}$ [eine Kompetenz]$_{\textbf{Acc}}$ ...
... *transfers the contract] to-the Union] [a competence] ...*

In sentence (19), there is a clear preference for nominative for *Vetrag* in combination with *besagen*, as well as an occurrence of *Union* as genitive modifier to *Vertrag*. This gives the model a good basis to predict grammatical case for the two noun phrases.
In sentence (20), the tuple *Vertrag übertragen* can both plausibly occur in nominative or accusative case. In contrast to the intransitive *besagen*, *übertragen* is bitransitive, thus requiring a direct and indirect object. In combination with the information that *Union* is unlikely as direct object of *übertragen*, whereas *Kompetenz* is preferably used as direct object of the verb *übertragen*, a prediction as illustrated in example (20) can be derived.

Integrating subcategorization knowledge

The task of the case prediction CRF consist in predicting a sequence of labels, with case values for inflected words in NPs and PPs and a "dummy-label" for all other words. Conceptually, the values of inflectional features are set by the head of the phrase (i.e. the noun), and then distributed over the rest of the phrase (such as articles and adjectives). Following this scheme, the subcategorization information is annotated on nouns to be propagated over the phrase during the case prediction.

There are two possibilities to identify the verb-noun tuples and noun-noun pairs for which the subcategorization statistics are looked up: (i) the verb-noun tuples and noun-noun pairs can be identified based on the tree-structure that is part of the output when using a string-to-tree system; (ii) the relevant tuples can be obtained based on the corresponding source-side dependencies projected to the target side. Note that the source side is regular, well-formed text that can be parsed and analyzed without restrictions, whereas the machine translation output is not well-formed (and in stemmed format), and thus cannot be parsed.

The subcategorization features are added to the features used to train CRFs, namely n-grams of stems and POS-tags. When integrating the subcategorization information into the CRF, the tuple probabilities and noun-noun$_{Gen}$ frequencies need to be discretized, as the model interprets numbers as strings. For the probabilities, the space between 0 and 1 is divided into five buckets: B$_{p=0}$, B$_{0<p\leq 0.25}$, B$_{0.25<p\leq 0.5}$, B$_{0.5<p\leq 0.75}$ and B$_{0.75<p\leq 1}$. The frequencies are bucketed to the powers of ten: $f = 0, f = 1, 2 \leq f \leq 10, 11 \leq f \leq 100$, etc. This representation allows for a more fine-grained distinction in the low-to-mid frequency range, whereas pairs in the high-frequency range are mapped into fewer groups. This setup aims at supporting the decision whether a noun-noun pair is a regular noun modifier, or rather a random co-occurrence of two nouns.

Table 4.3 shows an example for integrating the subcategorization features into the training data for the CRF. The noun *Unternehmen* ('companies') at the beginning of the sentence has a strong preference (bucket = 1) for nominative case in combination with the verb *erhalten* ('obtain'), while *Mittel* ('funding') is annotated to be a likely direct object to the verb *erhalten*. The noun *Einführung* ('introduction') occurs within a PP headed by *für* ('for'), a preposition that does only subcategorize accusative case, making an annotation redundant in this context. Lastly, the noun *Technologie* ('technology') is marked as a possible noun-noun$_{Gen}$ construction with a co-occurrence frequency assigned to the bucket $11 \leq f \leq 100$.

In addition to the probabilities and frequencies of the observed pairs, the CRF is also provided with the respective bigrams containing the two elements in the tuple. This is particularly important for verb-noun tuples, as the verb and the noun can often be separated by considerable gaps: in the example shown in table 4.3, there are 8 words between subject and verb, and 5 words between direct object and verb. By also annotating the respective relevant verb (or the governing noun in a potential noun-noun$_{Gen}$ construction) at the noun-level, all relevant target-side information is available in a compact and structured form: verb, noun and the respective associations with the different functions. Bridging the potentially large distances between a verb and its arguments has the additional effect of making the CRF less dependent on the actual observed context windows, which might also help to close the gap between

	Gloss	Stem	Tag	Acc	Dat	Nom	Verb	Gen	N1	Gold
1	*companies*	`Unternehmen<NN>`	*NN*	0.00	0.00	1.00	`erhalten`	–	–	*Nom*
2	*should*	`sollten<VVFIN>`	*VVFIN*	–	–	–	–	–	–	–
3	*financial*	`finanziell<ADJ>`	*ADJ*	–	–	–	–	–	–	*Acc*
4	*funding*	`Mittel<NN>`	*NN*	1.00	0.00	0.00	`erhalten`	–	–	*Acc*
5	*for*	`für APPR<Acc>`	*PRP*	–	–	–	–	–	–	–
6	*the*	`d<ART>`	*ART*	–	–	–	–	–	–	*Acc*
7	*introduction*	`Einführung<NN>`	*NN*	–	–	–	–	–	–	*Acc*
8	*new*	`neu<ADJ>`	*ADJ*	–	–	–	–	–	–	*Gen*
9	*technologies*	`Technologie<NN>`	*NN*	–	–	–	–	100	`Einführung<NN>`	*Gen*
10	*obtain*	`erhalten<VVINF>`	*VVINF*	–	–	–	–	–	–	–

TABLE 4.3: Adding subcategorization information into SMT output. (EN input: 'companies should obtain financial funding for the introduction of new technologies'). On the right, the correct labels are given.

the well-formed training data and the flawed MT output by replacing the target-side context window with structured relevant information.

4.3.3 Source-side information

To predict the grammatical case in a stemmed translation output, information about the syntactic functions on the source-side is essential, as illustrated in section 4.2.1. Between the source and target language, syntax-semantic functions can be isomorphic (such as an English subject having the same function in its German translation), but this is not always the case. (Refer to section 4.2.1 for more details on approximate similarities between English and German arguments.)

To obtain source-side features, the English data is dependency-parsed, using a parser by Choi et al. (2012), and verb-noun triples are identified using dependency structures. The identified verb-noun pairs are then transferred to the target side via word alignment. The projection focuses on English subjects and direct objects, as well as *noun-prep-noun* phrases (which can often be equivalent to German noun-noun$_{Gen}$ constructions). This set of English syntactic functions is generally likely to have a target-side NP or PP. Here, the fact that the case prediction model is applied to machine translation output might be slightly beneficial: while the training corpus contains a considerable amount of parallel sentences with diverging structures, the output produced by the SMT output tends to contain more isomorphic translations, as extensive structural changes are typically costly during the translation process.

Figure 4.2 shows the projection of source-side syntactic functions to the target side. As the annotation is to be applied to the German side, the projection process is based on German nouns in the MT output: for each noun in the stemmed MT output, the English

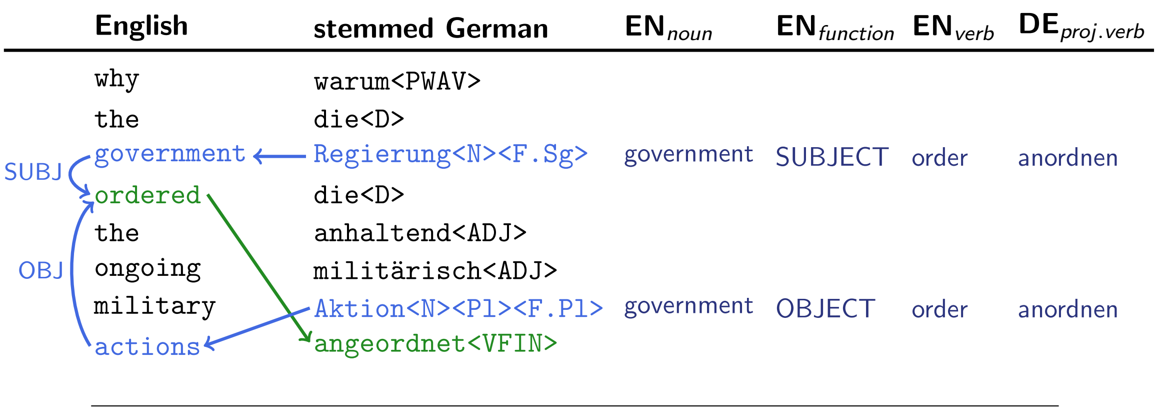

FIGURE 4.2: Projecting source-side syntactic functions to the target-side via word alignment.

equivalent is identified through the word alignment. Using the English dependencies, the governing English verb and the relation to the verb (i.e. subject vs. object) is extracted. Following the alignment from the English verb, the respective German verb can be identified. The feature set *English noun + synt. function + English verb + aligned German verb* is then annotated on noun-level in the CRF. Similarly, candidates for German noun-noun$_{Gen}$ constructions are identified by extracting and projecting source-side *noun-prep-noun* phrases, resulting in the set of *English noun1 + English noun2 + German noun1*, where noun1 corresponds to the governing noun in the phrase, and noun2 is the noun receiving the annotation, i.e. the noun to be inflected in genitive case.

4.4 Experiments and results

The effect of integrating source-side features and subcategorization information into the case prediction model is evaluated in two settings, prediction quality on clean data and performance measured in BLEU on actual translation output. Additionally, a small-scale manual evaluation was performed on a subset of the test set, confirming that there is a small improvement with the enriched prediction model. The experiments presented in this section are also described in Weller et al. (2013b).

4.4.1 Data and experimental setup

The experiment is based on the hierarchical translation system (Chiang, 2005) that is contained in the Moses framework. For word-alignment, GIZA++ (Och et al., 2003) was used with the “setting grow-diag-final-and". The rule-table was computed with the default parameter setting for GHKM extraction (Galley et al., 2004), using the implementation by Williams et al. (2012).

The training data for the SMT system consists of 1,485,059 parallel English and German sentences; the target-side part of the parallel data is used to build a 5-gram language model. The development and test sets contain 1025 and 1026 sentences, respectively. All data (training data and the development/test) were released for the Machine Translation Shared Task at the 2009 ACL Workshop on Machine Translation.

The setup of the inflection prediction system corresponds to the description in chapter 3. To prepare the stemmed data, the German text was parsed with BitPar (Schmid, 2004) and analyzed with SMOR (Schmid et al., 2004). For each inflectional feature (number, case, gender and strong/weak inflection), a CRF to predict the feature values was trained on the German side of the parallel training data using the Wapiti toolkit (Lavergne et al., 2010). Each model has access to stems, POS-tags and the feature to be predicted within a window of up to four positions to each side. To measure the effect on using richer features for predicting case, only the case model is modified, whereas the other prediction models, as well as the translation system, remain unchanged.

The target-side subcategorization tuples were extracted from Europarl and German newspaper data (HGC: “Huge German Corpus”, 200 million words). This data setting provides a combination of resources that matches the training data (Europarl constitutes a large part of the parallel training data) and the domain of the test set (news domain).

For the HGC data set, an already existing set of extracted subcategorization information was used; the extraction method is described in Schulte im Walde (2002), and relies on the parser described in Schmid (2000). To extract subcategorization information from Europarl, the extraction routine from Scheible et al. (2013) was applied to data parsed with Bohnet (2010).

4.4.2 Results on clean data and machine translation output

The case prediction model enriched with source- and target-side features is evaluated in two settings: (i) the prediction accuracy on clean data and (ii) the effect on machine translation output in terms of BLEU (Papineni et al., 2002). Finally, the best system is evaluated manually.

	0	1	2	3	4
	surface system	**simple prediction**	**subcat. features (tuples from EN side)**	**source-side features**	**source-side + subcat. features**
BLEU	13.43	14.02	14.05	14.10	14.17
Clean	–	85.05 %	85.65 %	85.61 %	85.81 %

TABLE 4.4: Results of a baseline surface system (0), the simple prediction system (1) vs. three systems enriched with extra features (2)–(4).

Table 4.4 shows the results for different system settings: System (0) is a baseline system trained on raw data without morphological processing. Systems (1)–(4) rely on the same inflection-prediction system, and the same models to predict number, gender and strong/weak inflection – differing only in the case prediction model. System (1) uses a simple case prediction model, that only has access to stems and POS-tags, corresponding to the standard model as presented in chapter 3. Systems (2), (3) and (4) contrast different ways of enriching the case prediction model, with system (2) using only subcategorization features, system (3) using only projected source-side features, and system (4) combining both source- and target-side features in addition to the standard features, stems and POS-tags. In systems (2) and (4), the tuples based on which the subcategorization information was integrated were identified with projected source-side dependencies[4].

Prediction accuracy

A simple way to asses the performance of the case prediction model is an evaluation of prediction accuracy on clean data instead of machine translation output. This has the advantage that the sentences can be parsed, and the predictions can be measured against the parse labels. For this evaluation, the number of correctly predicted case labels is measured: this is less forgiving than measuring correctly predicted forms, as many words take on the same surface form for different values of case. It also has to be noted that the prediction accuracy computed over *all* NPs/PPs to a certain extent downplays the difference between the prediction models, as this evaluation also includes instances that are very easy to predict, namely for PPs headed by a preposition that subcategorize only one value for case.

The row "Clean" in table 4.4 shows the results for prediction accuracy on the development set for the different case prediction models. All variants of the enriched

[4] For comparison, system (2) obtained a BLEU score of 14.00 when using the tree structures produced by the decoder as a basis to integrate target-side information.

models are better than the model using only standard features: the models using only source- or target-side features perform equally, and the model relying on the combined features obtains the best result. This shows that the additional features can improve the prediction of case.

A general problem in this feature prediction scenario is the mismatch between well-formed training data and the non-fluent machine translation output. Similar problems are also mentioned, for example, by Stymne et al. (2011b) and Kholy et al. (2012), who propose to first translate and annotate the training data, and to then add this new, artificial training data to the original training data. In addition to creating the practical problem of translating the training data, there is also the problem of reliably transferring the morphological annotation. Instead, by experimenting with the removal of features prone to over-fitting, the model is designed to be robust; generally, it can be observed that more complex models are less robust with regard to the mismatch between test and training data (i.e. improvement for the task of clean data prediction, but no effect on the translation output). In particular, high-order n-grams of POS-tags or stems (i.e. sequences of tags or stems) were found to have an overly high influence. Restricting the n-gram order of stems and tags also gives more impact to the new features.

Translation quality in BLEU

Table 4.4 shows the results for the different system variants in case-insensitive BLEU. The inflection prediction systems are significantly[5] better than the baseline surface system. However, there is not much difference in terms of BLEU between the outputs using the different case prediction models.

An important factor in this evaluation is the fact that the underlying machine translation output is the same for all tested prediction variants – to a certain extent, it is thus not possible to obtain a large difference in BLEU. Furthermore, changing the case of an NP or PP does not necessarily affect the inflection of all components in that phrase. For example, the determiner of an NP can be the only indicator for case:

(21) Er sieht [**den** alten Mann]$_{\textbf{Acc}}$
He sees the old man

(22) Er folgt [**dem** alten Mann]$_{\textbf{Dat}}$
He follows the old man

[5] Using pairwise bootstrap resampling with sample size 1,000 and a p-value of 0.05.

		enriched preferred	simple preferred	equal
(a)	**person 1**	23	11	12
	person 2	21	8	17
	person 3	26	11	9
(b)	**person 1**	23	5	18
	person 2	21	11	14
	person 3	29	8	9
(c)	**agreement**	17	2	6

TABLE 4.5: Manual evaluation of 46 sentences: without (a) and with (b) access to EN input, and the annotators' agreement in the second part (c).

Thus, it is not surprising that BLEU cannot reflect a potential improvement obtained through better prediction of case. Another problem with BLEU as an evaluation metric is its tendency to reward fluency rather than adequacy due to being a precision-oriented measure, as has been observed, for example, by Wu et al. (2009a) and Liu et al. (2010). Aiming at improving adequacy through better modeling of case is thus a factor that makes BLEU a non-optimal evaluation metric.

4.4.3 Manual evaluation of the best system

In order to better understand the effect of enriching the prediction model, a subset of the test set was manually evaluated. To this end, the output of the best prediction model (the variant combining both source- and target-side features) was compared to the output of the prediction model operating only on standard features.

To obtain a test set for the manual evaluation, the set of different sentences between the simple system and the enriched system (144 of 1026), was restricted to sentences with a source-side length between 8 and 25 words, resulting in 46 sentences to be evaluated. The objective of this length restriction is to provide sentences that are not too difficult to annotate, as longer sentences tend to be disfluent and thus harder to rate than shorter sentences.

The evaluation consists of two parts: first, the participants should rate the two inflected sentences (better/worse), but are not shown the English source sentence. This first setting aims at assessing the *fluency* of the produced sentences. The second part aims at measuring *adequacy*: with access to the English sentence, the participants should decide which variant better reproduced the content of the input sentence.

	P1 - P2	P1 - P3	P2 - P3
κ	0.6184	0.4467	0.3596

TABLE 4.6: Pairwise inter-annotator agreement between the participants.

Table 4.5 shows the evaluation obtained with 3 German native speakers. Generally, all participants prefer the enriched system over the simple system in both evaluation settings (i.e. with/without access to the source sentence). The row "agreement" lists the number of sentences agreed upon by all three annotators.

Table 4.6 shows the pairwise inter-annotator agreement for the task of deciding between three possible labels (*enriched preferred*, *simple preferred*, *none preferred*). While there is a substantial agreement between participants P1 and P2, the agreement with participant P3 is considerably lower. Furthermore, there is a general tendency to agree on sentences with the label *enriched better*, whereas sentences of the categories *simple preferred* or *none preferred* are often disagreed upon. This is illustrated by the number of sentences where all annotators agree, cf. line (c) in table 4.5: only two sentences were labeled as *simple preferred* by all three annotators. This distribution also illustrates how difficult it is to rate the output of machine translation output.

4.4.4 Examples

This section shows some examples to illustrate how using a combination of source- and target-side features can help to improve the inflection of the MT output with regard to the realization of case, but also points out the limits of the approach.

(23) EN hundreds of policemen were on alert , and [**a helicopter**]$_{\textbf{Subj}}$ circled the area with searchlights .

BL Hunderte von Polizisten auf Trab , und [**einen Helikopter**]$_{\textbf{Acc}}$ eingekreist das Gebiet mit searchlights .
hundreds of policemen on alert , and a helicopter circled the area with searchlights .

V4 Hunderte von Polizisten auf Trab , und [**ein Helikopter**]$_{\textbf{Nom}}$ eingekreist das Gebiet mit searchlights .
hundreds of policemen on alert , and a helicopter circled the area with searchlights .

In the inflection obtained with the simple prediction model in example (23), the subject phrase *a helicopter* has been inflected as direct object, whereas the enriched prediction model (system V4) correctly predicted nominative case for this NP.

(24) EN while 38 percent put **[their trust]**$_{\textbf{Obj}}$ in viktor orbán .

BL während 38 % **[ihres Vertrauens]**$_{\textbf{Gen}}$ schenken in Viktor Orbán .
while 38 % of-their trust give in Viktor Orbán .

V4 während 38 % **[ihr Vertrauen]**$_{\textbf{Acc}}$ schenken in Viktor Orbán .
while 38 % their trust give in Viktor Orbán .

The sentence in example (24) shows a typical error of the simple inflection prediction system – the confusion between object and genitive modifier. The structure of the sentence has the subject (*38 %*) next to the direct object (*ihr Vertrauen*: 'their trust'). However, a noun phrase preceded by *38 %* is generally a good candidate for being inflected as a genitive modifier. The enriched prediction model can benefit from the verb-noun pair *Vertrauen schenken* ('give trust'), thus correctly inflecting the phrase as direct object. Interestingly, only two of the three annotators prefer the inflection obtained by the enriched system, while the third one was undecided. A factor that might have played a role in his/her decision might be the error concerning the preposition *in* later in the sentence: the combination *Vertrauen schenken* ('give trust') requires an additional indirect object, i.e. a noun phrase with the recipient of the trust – thus, a prepositional phrase headed by *in* is incorrect here and might negatively impact the perception of the sentence. An alternative translation variant is *Vertrauen setzen in* ('put trust in'), which is structurally equivalent to the construction in the English input sentence, and requires a prepositional phrase for the recipient.

(25) EN more than $ 100 billion will enter **[the monetary markets]**$_{\textbf{Obj}}$ by means of public sales .

BL mehr als 100 Milliarden Dollar werden durch öffentlichen Verkauf
more than 100 billion dollars will by public sale

[der Geldmärkte]$_{\textbf{Gen}}$ treten .
of-the money-markets enter .

V4 mehr als 100 Milliarden Dollar werden durch öffentlichen Verkauf
more than 100 billion dollars will by public sale

[die Geldmärkte]$_{\textbf{Acc}}$ treten .
the money-markets enter .

A similar problem is also present in example (25): there are two possibilities to translate *enter the money market*, the first being structurally equivalent with the English construction (*[den Geldmarkt]$_{\textit{Dir.OBJ}}$ betreten*), or with a second variant that makes use of a prepositional phrase (*[auf den Geldmarkt]$_{\textit{PP.Acc}}$ treten*: 'to step into the money market'); between these two, the second one is preferable. The stemmed translation output contains an unfortunate mix of both variants: the verb *treten*, but no preposition *auf* in the phrase containing *Geldmärkte* – consequently, the sentence cannot be inflected correctly. While the inflection of *money markets* as genitive phrase obtained with the simple prediction model (meaning *by the public sales of the money markets*) is without doubt wrong, the inflection obtained with the enriched system is not entirely useful either, as the construction of the sentence is incorrect to begin with. This problem is also reflected in the evaluation – only one annotator rates the inflection obtained with the enriched system to be better, whereas the two others give the label *none preferred*.

With the currently presented approach for inflection prediction, the problem of missing/wrong/superfluous prepositions cannot yet be handled: chapters 6 and 7 discuss strategies to handle subcategorization across complement types, modeling both NPs and PPs in one step.

4.5 Related work

This section of related work is separated into three parts: predicting case markers in an English–Japanese translation task, general research on using rich context information features, and the integration of semantic roles into machine translation.

4.5.1 Prediction of case markers in Japanese

A paper directly related to the work discussed in this chapter is the prediction of Japanese case markers in a simulated English–Japanese translation setting (Suzuki et al., 2006). They compare two different settings: using only monolingual Japanese features, and using a combination of source-side and target-side features. Japanese case markers indicate grammatical functions (such as subject, object and location), and are difficult to generate during translation. In many cases, they do not have a direct equivalent on the English source-side, but are rather expressed through sentence ordering or otherwise. Even though German and Japanese are different languages, the general task of predicting syntactic functions on the target-side is essentially the same.

Nominal particles in Japanese are post-positions attached to nouns; they can be categorized into three categories: *case particles*, where one particle may represent different functions (e.g. the particle *ni* can indicate an object or a location); *conjunction particles* equivalent to the English phrases 'and' and 'or'; and *focus particles*. Particles from the last two categories are not modeled, expect for the special topic particle *wa*, that is one of the most frequent particles in Japanese. In total, 10 case markers and the special particle *wa* are modeled, as well as the concatenation of some case markers and *wa*, resulting in a set of 18 case particles to choose from in the prediction task, which consists in first identifying phrases requiring a case marker (binary yes/no decision); and then the prediction of one or more case particles in a phrase ("*bunsetsu*").

In the monolingual prediction task, a classifier is trained on a rich feature set including POS/lemma/word information from a window around to the target noun, and further features such as sibling information, negation, voice and dependency information in form of parent-child pairs. The classifier clearly outperforms baseline variants selecting either the majority particle or the particle that maximizes the probability in a language model. A joint prediction model that has access to the labels of other predictions leads to the best results.

In the bilingual prediction task, source-side dependency information is made available through projecting source-side information to the target-side using word alignment. Adding source-side information leads to an improved prediction result.

Toutanova et al. (2007) extend the work of Suzuki et al. (2006) and integrate the case classifier into an actual translation system. As method of integration, they choose to add the classifier as feature functions into the n-best reranking component, as this allows to efficiently make use of long-distance information that would be difficult to handle during decoding. This approach leads to improvements in BLEU, in particular when expanding the n-best list with further case variations. The improvement is also confirmed in a manual evaluation.

The general task of predicting case in a translation scenario as presented in Toutanova et al. (2007) is very similar to the approach discussed in this chapter. However, there are also differences between this work and Toutanova et al. (2007)'s method, namely the integration of the classifier into the system's re-ranking component (Toutanova et al. (2007)), whereas the prediction task in this chapter is carried out as a post-processing step. Another difference is that the case prediction in this work is integrated into a system that models full nominal inflection, which is not necessary in Japanese, where the case marker can just be attached to the noun.

4.5.2 Integration of rich context information

There is a lot of research on integrating rich source-side information into machine translation, though not always with the purpose of modeling case.

Avramidis et al. (2008) enrich the source-side of English–Czech and English–Greek systems in order to provide morphological attributes that are required by the target side, but missing on the source-side as explicit morphological attribute. The features to model noun-case agreement and verb-person-conjugation are contained in the syntactic structure of the source-side and can be derived from a parse output. The features are added as source-side factors. Exploring alternative-path strategies to handle sparse data (i.e. backoff to less factors), leads to moderate improvements for English–Greek and English–Czech translation. Similarly, Yeniterzi et al. (2010) integrate rich source-side context in form of "complex structural tags" as source-side factors into an English–Turkish translation system, leading to improved BLEU scores. A related concept is presented by Daiber et al. (2015) who predict target-side morphological features to enrich the source-side of a translation system, leading to improved translation quality.

A different approach to include source-side information is presented by Jeong et al. (2010) who integrate a discriminative model into string-to-tree *treelet* systems to translate into the morphologically rich languages Bulgarian, Czech and Korean. The rich context information considered during the translation allows the model to select translations according to the respective syntactic and morphological context, represented by a rich set of local context features, dependency-based and morphology-based features. They obtained improvements for all language pairs. Another variant of integrating a discriminative model into machine translation is proposed by Tamchyna et al. (2014); an extension of the model also takes into account more target-side context from the (partially) produced translation during decoding (Tamchyna et al., 2016).

Subotin (2011) proposes an exponential translation model to translate into morphologically complex languages. Working on English–Czech translation, he integrates the features number, gender and case into the hiero structures in a way that they can be passed on over long distances. The modeling of number is straightforward in his setting, as it can easily be obtained from the source side. Case is modeled based on source-side features only; they include the syntactic function of the aligned phrase (subject, object or nominal predicate), the preposition governing the aligned phrase, as well as features indicating the presence or absence of numerals and possessive markers. Additionally, lemmas of governing verbs over a frequency threshold are considered as lexical features.

Gender is a target-side specific feature, and cannot be obtained from the source side. For words that take their gender value from a governing noun, such as adjectives and verbs, a special tracker feature is used to link the gender to the noun's gender. In the decoding process, the feature values are communicated per "rule messager" if a non-local feature lies outside the span of a rule. This setting also allows to generate target inflections not occurring in the parallel training data, for which an external morphological generator is used. This generation process is subject to several restrictions enforcing agreement and coherence. For example, the number of a noun or verb cannot be changed, or the negation status and the degree of comparison of adjectives has to match with the original form. This method leads to improvements in BLEU for systems trained on a small as well as on a large data set.

4.5.3 Integration of semantic roles

This section gives a short overview of the integration of semantic roles into SMT. While this work does not, strictly speaking, integrate semantic roles into SMT, the combined information of selectional preferences in the subcategorization data base and the projected source-side features aim at approximating semantic role information.

Assuming that semantic roles are preserved between source and target language, there has been a growing interest in integrating semantic role information into machine translation; Wu et al. (2009a) evaluate the impact of semantic roles in machine translation. Working with parallel sentences from PropBank (Palmer et al., 2005), i.e. data with gold standard annotations of syntactic and semantic roles, they evaluate the accuracy of semantic roles in translations obtained with an SMT system, confirming their hypothesis that incorrect translation of semantic roles is indeed responsible for many errors in machine translation. Comparing semantic and syntactic functions, they come to the conclusion that semantic roles are more consistent between languages than syntactic information (at least for the evaluated language pair English–Chinese). A last experiment aims at enforcing semantic consistency between the input sentence and the translation by re-arranging the translation such that the semantic role labels obtained by parsing match with those of the input sentence. This approach leads to an improvement in both BLEU and METEOR. In Wu et al. (2009b), the idea of re-ordering translation hypotheses to increase the match of semantic role labels with those in the input sentence is described in more detail.

Liu et al. (2008) add semantic-role information to tree-to-string transducer templates, but report mixed results for this approach. In a later work, Liu et al. (2010) add semantic roles projected from the source-side into a tree-to-string transducer system, proposing two feature types: a reordering feature for "skeleton-level permutation" to ensure that source and target semantic roles are aligned, and a feature to penalize missing semantic roles in the translation. The semantic features, in combination with tree-to-string templates, are trained in a conditional log linear model and improve the output of an English–Chinese translation system.

Gao et al. (2011) build a hierarchical translation system with rules containing information about the semantic role structure on the target side. During translation, the rule setup ensures that only hypotheses with complete semantic structures can be produced. For experiments on Chinese–English translation, they report an improvement in BLEU.

Bazrafshan et al. (2013) enrich the non-terminal symbols of a string-to-tree translation system with semantic role information (method 1). Furthermore, the rule extraction process is modified such that rules represent the complete predicate-argument structure of a verb (method 2). Such a rule has the smallest tree fragment containing all arguments of a predicate and the predicate itself. Semantically enhanced rules are combined with regular rules by an added binary feature. In their experiments, they find that method 2 leads to improved translation results, whereas method 1 is not able to outperform the baseline. This is attributed to the fact that in method 1, the system has only partial information about the semantic structure, whereas the rules in method 2 can learn the complete semantic structure.

4.6 Summary

This chapter presented an extension to the inflection prediction system to improve the prediction of grammatical case through the use of source-side features and an external subcategorization database. As case also carries a semantic dimension, access to source-side features (namely the function of a phrase in the source-sentence) is indispensable in order to ensure that the semantic role of a source sentence is appropriately reproduced in the corresponding target-side phrase. Target-side subcategorizational requirements, such as differentiating direct/indirect objects, are addressed by subcategorization information in the form of verb-noun$_{Case}$ tuples with the associated probabilities. Furthermore, the distinction of modifiers in contrast to subcategorized arguments is handled through combining the subcategorization information with statistics about noun-noun$_{Gen}$ constructions. Combining projected source-side features with target-side subcategorization information brings together two different types of information that are both necessary for the prediction of case.

An evaluation of the enriched prediction model showed minor improvements, both in terms of prediction accuracy on the development set, and in terms of BLEU on machine translation output. However, as the case prediction only constitutes a post-processing step, not much improvement in terms of BLEU can be expected as there are no lexical or structural changes. A manual evaluation contrasting the inflections obtained with the simple and the enriched prediction model confirmed a positive effect of the enriched prediction model with regard to fluency and adequacy.

A closer look at the evaluated examples showed that the case prediction model, basic or enriched, comes to its limits when the sequence of stems needs to be modified on order to allow for a successful inflection. In particular, this applies to missing and superfluous prepositions that are either expected as part of the subcategorization frame, or are not included in the subcategorization frame – thus, prepositions, which are generally considered challenging in machine translation, also need to be taken into account when modeling subcategorization. While the next chapter focuses on integrating selectional preferences to improve the translation of prepositions, chapters 6 and 7 discuss strategies to model subcategorization across complement types.

Chapter 5

Noun Class Information to Model Selectional Preferences of Prepositions in Statistical Machine Translation

The previous chapter used externally obtained subcategorization information to improve the assignment of syntactic functions in the inflection prediction approach, implemented as a post-processing step as part of the feature prediction with a focus on modeling NPs. However, the translation of prepositions is also a challenging problem in machine translation: prepositions are highly ambiguous, and the choice of prepositions depends on many factors, most prominently the governing noun/verb and the noun within the prepositional phrase. In this chapter, noun class information, obtained from a lexical resource or through clustering, is used to model selectional preferences of prepositions during translation. The annotation of noun class information into the parse trees used for the training of a hierarchical machine translation system aims at obtaining more precise translation rules: by grouping training instances according to their annotation, the resulting translation rules do now allow for any PP to be applied in its context, but restricts the selection to a specific semantic class. While the presented approach does not lead to improved translation results, it gives valuable insights to the many problems arising when translating prepositions: for example, structural differences (such as NP vs. PP) between source- and target-side can be challenging for SMT systems. The experiments described in this chapter are published in Weller et al. (2014b).

5.1 Motivation

The translation of prepositions is a challenging problem in machine translation. While prepositions are a closed-class set of high-frequency words, the task of selecting a correct preposition for a given context is surprisingly difficult and depends on several factors. A preposition can play different roles in a sentence: some prepositions convey a meaning, for example *on the moon* or *under water*, whereas others are merely functional, namely in a subcategorized context, such as *to believe in something* or *to consist of something*. Many prepositions also occur in contexts where they are neither just functional, nor just transporting a particular meaning. For successfully translating prepositions, the translation system must chose a preposition that suitably reproduces the source-side content, while fulfilling target-side requirements at the same time. Typical target-side factors that determine the choice of preposition are the subcategorization frame of a *governing verb or noun*, in combination with *selectional preferences* (often based on the noun in the PP) that can impose additional restraints.

An example for selectional preferences affecting the choice of preposition can be observed for the translation of *to learn from* into German: depending on the content of the PP, the phrase needs to be translated using different prepositions, as illustrated in example (1).

(1) to learn from PERSON → lernen von PERSON
to learn from ABSTRACT → lernen aus ABSTRACT

Thus, when one *learns from a person*, the preposition in the German translation should be *von*, the typical literal translation of *from*, whereas the translation *learning from an abstract entity*, such as *past*, requires the preposition *aus* ('out-of').

Such selectional preferences are difficult to learn for a standard machine translation system, as this requires a generalization over subcategorized elements in the context of a particular target-side verb or noun in combination with the respective source-side meaning in the case of a not entirely functional preposition.

To model selectional preferences for prepositions in machine translation, this chapter proposes a method that integrates *noun class information* in form of tree-labels into a string-to-tree translation system. The annotation of noun class information into the target-side parse trees aims at obtaining more precise translation rules that group the translation options for prepositions with respect to the annotated noun classes. As a result, the noun class annotation restricts a translation rule to PPs of a specific semantic class, as illustrated by example (2).

(2) learn [X_{PP}] → [PP-von-PERSON] lernen
learn [X_{PP}] → [PP-aus-ABSTRACT] lernen

Reflecting the selection restriction from example (1), the corresponding translation rule is separated into a rule producing the preposition *von* for the semantic class of *persons*, and another rule generating the preposition *aus* for words of the class *abstract*.

As basis for the annotation of noun class information, a resource-based method is compared with two variants of a corpus-based approach:

- nominal concepts induced from the lexical semantic taxonomy *GermaNet*
- k-Means cluster analyses relying on standard distributional window co-occurrence
- k-Means cluster analyses relying on syntactic features

The use of a lexical resource such as GermaNet (Hamp et al., 1997) provides a conceptually refined form of target-language information for a given set of nouns. In contrast, noun clustering relying on large target-language corpora allows for a better generalization over complex contexts (either in "raw" form or using dependency information) that goes beyond the potential of the context of a translation rule or a language model.

While the annotation of noun classes adds semantically fine-grained information, it also leads to a loss of generalization in contexts where such fine-grained information is not relevant. This is compensated for by making generic rules accessible to the annotated system. Additionally, new translation rules that cannot be derived from the parallel data are generated to further enrich the translation system.

The experiments described in section 5.4 will show that the proposed annotation of noun classes does not lead to an improvement in comparison to a baseline without annotation. However, the experiments provide insights into enriching an SMT system with noun class information with regard to (i) the method of integration, and (ii) the type of resources used to derive the noun class information. Integrating noun class information via tree labeling turns out to be to inflexible, as it is impossible to find a generally applicable level of semantic information. At the level of resources, the comparatively simple clustering method based on window information outperforms both noun class information obtained through clustering based on syntactic context and the classes as defined in the high-quality lexical resource GermaNet.

Furthermore, a closer look at the evaluated examples shows that prepositions tend to be prone to being mis-translated when there are structural differences between source and target language, such as a PP on the source side being equivalent to a noun phrase on the target side or vice-versa. Another problem are prepositions in a subcategorized context requiring an "unpopular" translation that is different from the predominant literal translation – here, it seems difficult for the system to override the default translation.

This chapter addresses the translation of prepositions at a semantic level by integrating noun class information with the aim to enforce selectional preferences during translation. The two following chapters, 6 and 7, also include a syntactic dimension by proposing a strategy to handle the structural differences of NPs vs. PPs. To address the problem of default translations in special contexts, the method to handle the translation of prepositions departs from the standard type-based translation probabilities to techniques that allow to take into account token-level requirements. These techniques are extended, in chapter 7, to the generation of synthetic phrase-table entries, that are conceptually similar to those in this chapter, but rely on a more refined setup.

Thus, while the experiments in this chapter are not successful in that they do not improve the translation quality, the observations made in the evaluation establish the basis for the two following chapters.

5.2 Obtaining noun class information

The annotation of noun class information aims at refining hierarchical translation rules such that they reflect selectional preferences of prepositions that are tied to the noun in the prepositional phrase. This section covers the three approaches used to obtain noun class information: as first variant, classes from the lexical resource GermaNet are mapped to the nouns in the training data. The second and third variant rely on noun clustering, either based on window information or on dependency structures. These three variants, based on a *hand-crafted lexical resource* or on *structured/unstructured context* in large corpora, are conceptually disjunct and provide a systematic assessment of selectional preferences.

5.2.1 Pre-processing

Before computing noun classes, the training data undergoes two steps of pre-processing in order to ensure a consistent representation of noun class information.

Regular nouns versus named entities

A problem when annotating noun classes is the fact that a word type can be tagged as both *noun* (tag = NN) and *named entity* (tag = NE), which leads to a potentially inconsistent annotation. Thus, the first part of pre-processing consists in resolving inconsistent parsing decisions for word types analyzed as both noun or named entity: only words recognized as nouns by the high-coverage morphological analyzer SMOR (Schmid et al., 2004) are considered common nouns. As finite-state based tool, SMOR relies on a finite lexicon, but can analyze and recognize all compounds and derived forms created on the basis of its vocabulary. Words not recognized as nouns by SMOR are considered named entities; they are further classified into *organization, location, person* and *rest* using Faruqui et al. (2010).

This pre-processing step ensures a consistent labeling of all nouns with the same lemma. Furthermore, by excluding “non-nouns” such as types or parse-errors (that are often infrequent and thus likely to deteriorate clustering), only “true" nouns that can be expected to have a sound basis for feature extraction are considered for clustering. Similarly, GermaNet coverage is restricted to common nouns, and filtering non-nouns consistently in a pre-processing step leads to a cleaner annotation.

Reduction to head nouns

The second part of the pre-processing consists in compound handling in order to generalize the clustering over nouns with the same head noun. German noun compounding is very productive and can lead to sparsity and coverage problems: a high number of infrequent compounds is likely to have a negative effect on the clustering performance. The clustering is thus applied to head nouns, and compounds are added into the class of their head noun.

The compounds are split using the linguistically informed splitter by Fritzinger et al. (2010), which combines morphological analyses (obtained from SMOR) with corpus statistics to find the best splitting option for a given compound. The compound splitting is applied on the type-level and does not take into account the compositionality of a compound (e.g. *Hexenschuss*: ’lumbago’, lit: ’witch shot’). While the reduction of compounds to the head noun might introduce some noise for a small number of non-compositional compounds, this is likely outweighed by the gain in generalization obtained for the large number of compositional compounds.

5.2.2 GermaNet

GermaNet (Hamp et al., 1997; Kunze, 2000) is a lexical resource for German similar to the English WordNet (Fellbaum, 1998). It is a lexical-semantic taxonomy that groups words of the same concepts into synsets.

To obtain word classes for the annotation in the SMT training data, the GermaNet class of the (head) noun is looked up. GermaNet is graph-structured, and the different hierarchical levels determine the degree of generalization, i.e. coarse-grained to fine-grained. For the annotation of the trees in the training data, noun classes from the levels 2, 3, 4 and 5 (counting from the top level) were used. In some cases, a unique assignment is not possible, as words can belong to different synsets in GermaNet. In such cases, the word is put into the class with the lowest GermaNet-internal ID number. Furthermore, nouns not found in GermNet are assigned to a special "rest class" (16.357 of 211.360 after compound processing).

5.2.3 Clustering

For clustering, the standard k-Means algorithm contained in *R* (R Core Team, 2014) is used, relying on features extracted from the target-side of the parallel data and an additional large web corpus (Faaß et al., 2013), resulting in ca. 45 million sentences total, cf. section 5.4.1 for more details on the experimental setting.

Low-frequency nouns ($f \leq 5$ in the combined corpora) are excluded from clustering, and added to the cluster with the nearest centroid in a post-clustering step. In combination with the pre-processing steps focusing on clustering only "true" nouns and head nouns, this step aims at providing a solid base for the clustering step by removing nouns that are potentially problematic.
For the clustering, two sets of features are contrasted:

- *Bag-of-word features*: content words in a window of 10 words to each side of the noun
- *Syntactically motivated features* referring to subcategorization criteria:

Noun		**GermaNet (level 5)**
Minister	*minister*	human being
Kanzler	*chancellor*	human being
Mehrheit	*majority*	group
Opposition	*opposition*	configuration
Enthebung	*dismissal*	ending/stop

TABLE 5.1: A selection of nouns from a cluster (window-features) representing the domain *politics* and the corresponding GermaNet classes.

P preposition governing the target noun

VO verb governing the target noun

VPN verb governing the target noun in a prepositional phrase

NPN noun governing the target noun in a prepositional phrase

The k-means algorithm allows to set the number of clusters – however, it is difficult to determine a number of clusters that provides a good representation of nouns and an optimal level of abstraction for the SMT system at the same time. To assess the effect of the granularity represented by the clusters, the annotation of the parse trees compares clusterings with sizes between 10 and 300 clusters.

A general effect of the clustering is the classification into "topic-like" clusters that rather contain nouns belonging into a certain domain, but that do not necessarily represent a generalization over specific noun types. Table 5.1 contrasts the grouping obtained with a clustering relying on window features and the classification according to GermaNet: the *politics*-related cluster contains both *persons* (*minister, chancellor*) and abstract concepts such as *dismissal* and *opposition*, whereas the classes assigned by GermaNet rather represent a generalization over specific noun types by grouping the nouns *chancellor* and *minister* as *human beings* and the remaining terms *majority*, *opposition* and *dismissal* into an individual group each.

The use of syntactic features in contrast to window-based features aims at better capturing selectional preferences to obtain classes that provide more salient information to model prepositions in SMT. See for example Erk et al. (2010) or Schulte im Walde (2006) for more information on modeling selectional preferences.

5.3 Using noun class information in SMT

This section presents how the parse trees used to train a string-to-tree system are annotated with noun classes. These basic annotated systems are further extended with

```
<tree ="s">
  <tree="adjd"> wirtschaftlich</tree>                    economically
  <tree="vafin-haben"> hat </tree>                       has
  <tree="np-LOC">
    <tree="ne-LOC"> malaysia </tree>                     Malaysia
  </tree>
  <tree="vp">
    <tree="pp-von-167">
      <tree="prep-von-167"> von </tree>                  from
      <tree="pposat"> seinen </tree>                     its
      <tree="nn-167"> nachbarn </tree>                   neighbours
    </tree>
    <tree="vvpp"> gelernt </tree>                        learned
  </tree>
</tree>
```

economically, Malaysia has **learned from its neighbours***.*

```
<tree ="s">
  <tree="kous"> dass </tree>                             that
  <tree="np-180">
    <tree="art"> die </tree>                             the
    <tree="nn-180"> amerikaner </tree>                   Americans
  </tree>
  <tree="vp">
    <tree="pp-aus-291">
      <tree="prep-aus-291"> aus </tree>                  from
      <tree="art"> der </tree>                           the
      <tree="nn-291"> vergangenheit </tree>              past
    </tree>
    <tree="vvpp"> gelernt </tree>                        learned
  </tree>
  <tree="vafin-haben"> hätten </tree>                    have
</tree>
```

that the Americans had **learned from their past***.*

FIGURE 5.1: Example for trees annotated with noun class information contrasting the different translations for *learn from* into *lernen von* and *lernen aus*. The representation of the data in the example is simplified for better readability, as the system is trained on the stemmed representation needed for the inflection prediction system.

non-annotated baseline rules and synthetic PP rules.

The translation system applies the inflection prediction process (following Fraser et al. (2012); cf. chapter 2 for more details). While the modeling of target-side morphology is not immediately necessary for the integration of noun class information, it allows for an easier handling of portmanteau prepositions, which are split in a pre-processing step and merged in a post-processing step. Thus, during translation, prepositions occurring as portmanteaus are represented the same way as non-portmanteau prepositions, which simplifies the step of annotating noun class information.

5.3.1 Annotating translation rules with noun classes

The annotation of noun class information is applied to both the preposition node (`prep`) and its father node (`pp`), as well as to noun- and NP-nodes. A richer annotation of the parent-node (such as the PP-node) in hierarchical MT has been shown to be beneficial (Huang et al., 2006), as it helps to transport context information from a higher level (such as the PP-level) to the terminals (the preposition).

Figure 5.1 shows the annotated target-language parse trees, with noun classes being represented by indices. The example illustrates how the annotation serves to create two translation variants for the English phrase `learned [from NOUN]`$_{PP}$, namely `VP → pp-von-167 gelernt` and `VP → pp-aus-291 gelernt`. These translation variants reflect the selectional preference for *lernen PREP*, as the translation rule with the preposition *von* can only be filled with words of the class 167 (*person*), whereas the rule generating the preposition *aus* only accepts nouns of the class 291 (*abstract concept*).

5.3.2 Adding non-annotated rules

While the integration of noun class information leads to fine-grained rules that represent selectional preferences, the annotation of NP and PP nodes might also lead to overly specific rules. To overcome a potential loss in rule generalization, non-annotated rules are added to the set of enriched translation rules. In the experimental section (section 5.4), three variants of adding non-annotated rules are compared:

- **BL** Addition of non-annotated baseline rules.

- **BL+cutoff** Rules derived from source-target pairs with an occurrence frequency of f $\leq$ 5 are likely not representative for selectional preferences and are thus removed, keeping only the non-annotated rules for such cases.

preposition-noun-verb		**freq**
aus nn-166 lernen	*to learn from nn-166*	38
für nn-166 lernen	*to learn for nn-166*	5
in nn-166 lernen	*to learn in nn-166*	30
mit nn-166 lernen	*to learn with nn-166*	5
von nn-166 lernen	*to learn from nn-166*	80
über nn-166 lernen	*to learn about nn-166*	80

TABLE 5.2: Subcategorization information induced from large monolingual corpora.

- **BL+subst** Only baseline rules with a higher translation probability than the respective annotated rules are kept. This leads to a replacement of rules representing no clear selectional preference such as `to buy nn1/nn2/nn3/...` with the general rule `to buy nn`.

5.3.3 Generating new PP rules

Another problem is the general coverage of translation rules, independent of the annotation with noun class information: in particular for rules spanning longer sequences, not all potentially necessary rules can be derived from the parallel data. A second extension of the translation system thus focuses on the addition of synthetic rules to supplement the set of rules extracted from the training data. New rules containing prepositions are created by duplicating annotated PP rules in which the existing prepositions are substituted. This aims at providing the full set of *functional* prepositions, which typically convey only minimal meaning and thus can be substituted by different prepositions without a meaning change. The set of functional prepositions is estimated by identifying subcategorized prepositions using a subcategorization lexicon (Eckle, 1999). The resulting set comprises 17 prepositions: *an, auf, aus, bei, durch, für, in, mit, nach, über, um, unter, von, vor, wegen, zu, zwischen*.

The rules are generated to reflect selectional preferences by computing the translation probabilities based on co-occurrence frequencies of the respective noun-prep-noun and verb-prep-noun tuples observed in a monolingual corpus. This subcategorization information is extracted from the target-side part of the parallel data and a large web corpus (the same data used for the feature extraction for the clustering, cf. section 5.4.1).

Table 5.2 shows frequency information for the preposition-noun-verb tuple *aus-nn-166-lernen*, which serves as a basis to generate the rule variants in table 5.2, where

source-side:learn [X]$_{PP}$, [X]$_{S}$

original rule (target-side)	**prob**
vp → [pp-von-166] lernen , [s]	1

new PP rules (target-side)	**prob**
vp → [pp-aus-166] lernen , [s]	0.159
vp → [pp-für-166] lernen , [s]	0.021
vp → [pp-in-166] lernen , [s]	0.126
vp → [pp-mit-166] lernen , [s]	0.021
vp → [pp-von-166] lernen , [s]	0.336
vp → [pp-über-166] lernen , [s]	0.336

Table 5.3: Generated variants for a base PP translation rule.

the target side of the original annotated rule is expanded into six rules containing prepositions observed with the verb *lernen* and nouns of the annotated class.

The generation of new rules is only applied on the level of PP nodes and PREP nodes; the rest of the rule, such as terminal symbols or the source-side, remain the same. To keep the number of rules manageable, only rules with a frequency of f $\geq$ 5 were used as base rules from which variants are generated. Furthermore, only rules with a translation probability of p $\geq$ 0.001 are kept. In the experiments in section 5.4, two variants containing new rules are presented:

- **New Rules** Addition of generated translation rules, as described above
- **BL+new** Addition on non-annotated baseline rules in combination with newly generated rules

5.4 Experiments and results

This section compares the variants of annotating different types of noun class information into the parse trees used to train a string-to-tree inflection prediction SMT system. Following an overview of the results in terms of BLEU and a small manual evaluation, the difficulty of translating prepositions is discussed based on selected examples.

5.4.1 Data

The system is trained on 1.5 million parallel sentences (Europarl and news domain) from the 2009 WMT Translation Shared Task, with a 5-gram language model trained on

System	BLEU
Baseline	13.95
GermaNet-2 (25)	13.93
GermaNet-3 (79)	13.77
GermaNet-4 (175)	13.67
GermaNet-5 (392)	13.67

System	BLEU
Window10	14.01
Window50	14.18
Window75	13.69
Window100	14.13
Window300	13.71

Syntactic features	P	VO	VPN	NPN
100 classes	13.85	13.85	13.79	13.71
50 classes	13.84	14.06	14.06	13.91

TABLE 5.4: Results in case-insensitive BLEU for different annotation variants: GermaNet and clusterings based on window information or syntactic features; the scores are averaged over two tuning runs. The numbers in brackets for GermaNet indicate the number of classes and the numbers 2,3,4,5 denote the respective level.

System	BL	BL	BL	new	BL
		cutoff	subst	rules	new
Window50	13.95	13.99	14.04	14.11	13.98
Window75	14.16	13.96	14.07	13.66	14.01
Window100	14.01	13.94	13.96	14.14	14.02

TABLE 5.5: Results for systems extended with non-annotated rules and new PP rules.

the target-side data. The development and test sets consist of 1025/1026 sentence of news data, also from the 2009 WMT Shared Task.

The German side of the data is parsed with BitPar (Schmid, 2004). To build a string-to-tree system, GHKM extraction (Galley et al., 2004) with standard parameters (using the implementation by Williams et al. (2012) that comes as part of the Moses framework) is used. For generating inflected forms from the stemmed translation output, the morphological tool SMOR (Schmid et al., 2004) is used, in combination with morphological features predicted with the Wapiti toolkit (Lavergne et al., 2010), cf. section 3.3 for more details on the process of inflection-prediction.

The context vectors for the clustering are calculated on the large web corpus *SdeWaC* (44 million sentences, Faaß et al. (2013)) combined with the target-side of the parallel data. This monolingual corpus is also the basis for extracting the tuples to estimate the translation probabilities for the generated translation rules.

5.4.2 Results

Table 5.4 shows the results for systems enriched with noun class information, both based on cluster analyses and GermaNet, in comparison to a baseline system without noun class annotation. Unfortunately, none of the enriched systems is significantly better than the unannotated baseline. However, the cluster analyses based on window-features seem to obtain better results than systems annotated with classes obtained through GermaNet or cluster analyses based on syntactic features. This outcome is intuitively plausible: on the one hand, GermaNet tends to suffer from coverage problems and is often unduly fine-grained (for example, the word *chancellor* is assigned to 2 classes at level 5, namely *organism* and *living being*, which is a distinction that exceeds the classification requirements in this application.) On the other hand, the syntactic features are more sparse than the window-based features. This is due to the simple fact that it is always possible to extract a context vector of content words within a window for a given noun, while the extraction of syntactic features is more restrictive and limited to constellations in which the respective feature is defined. The window-based clusters thus seem to provide the most robust representation of selectional preferences. The number of annotated classes has no strong influence, even though there is a tendency for less classes leading to better results.

Three window-based systems (window-50/75/100) are additionally extended with non-annotated rules ("BL", "BL-cutoff" and "BL-subst") and newly generated PP rules ("new rules"), as well as a combination of non-annotated and new rules ("BL+new"). The results are displayed in table 5.5. None of the system variants leads to an improvement in comparison to the respective original system or the baseline, even though there is a moderate improvement in some settings for system *windows-75*, the worst system in table 5.4. Furthermore, none of the studied variants seems to perform better or worse than the other variants.

Comparing the baseline and the enriched systems shows that the enriched systems use on average more and shorter translation rules than the baseline system, and also make more use of *glue rules*. Glue rules can be considered as a sort of last resort for the system during translation, as they allow to "glue together" unrelated translation rules. For example, the systems window-50/70 use an average of 11.99/11.62 glue rules per sentences, whereas the baseline only uses 7.10 glue rules. Similarly, the average length of the target-side of a translation rule decreases from 2.19 (baseline) to 1.91/1.92 for the window systems. At the same time, the average sentence length is stable over these systems, varying between 25.3 and 25.5 words. Assuming that a lower amount of glue rules in combination with longer translation rules is preferable, this outcome can be considered as an indicator of a general problem with the enriched rules: as the longer and more specific rules in the enriched system often are not applicable anymore, they need to be replaced by a combination of shorter rules, resulting in a loss of the context provided by a single longer rule.

These results contradict the initial objective of providing new, generalized information to model selectional preferences through annotating noun class information. Rather, introducing noun classes by means of parse tree annotation comes at the cost of losing basic context information as rules spanning over larger chunks are often not available anymore. On the other hand, it has to be noted that the introduction of noun classes, at least those derived through window-based clustering, do not harm the performance of the translation system and even lead to small improvements in some settings, despite causing the system to rely on considerably smaller translation units. Thus, it seems that the introduced information *is* useful to compensate the general loss of context due to smaller rules, but at the same time leads to noise through overly specific rules that ultimately prevents the system to really gain from the new information.

5.4.3 Examples of improved translations

This section shows some examples for improvements obtained with noun class annotation in the system window-50, the overall best system.

(3) EN more than $ 100 billion will **enter the monetary markets** by means of public sales

BL mehr als 100 Milliarden Dollar wird **die Geldmärkte** durch öffentlichen Verkauf
more than 100 billion dollar will the money-markets by public sale
gelangen
get

W50 mehr als 100 Milliarden Dollar **auf die Geldmärkte gelangen** wird durch den
more than 100 billion dollar on the money-markets get will by the
öffentlichen Verkauf
public sale

In sentence (3), the translation of *enter → gelangen* requires the preposition *auf* ('to get on the money market'), which is correctly produced by the enriched system. Alternatively, it is also possible to translate the phrase *enter the money markets* without a preposition, for example with the verb *erreichen* in combination with a direct object, keeping the syntactic structure as in the English input sentence ('reach the money markets').

(4) EN the charge that she **concentrated** too much **on foreign affairs** , ...

BL der Vorwurf , dass sie ∅ **auswärtige Angelegenheiten** zu stark **konzentriert** ist
the charge , that she ∅ foreign affairs too much concentrated is

W50 der Vorwurf , dass sie zu sehr **auf die auswärtigen Angelegenheiten konzentriert**
the charge , that she too much on the foreign affairs concentrates

I sentence (4), the preposition *auf* for the translation of *concentrate <u>on</u>* is missing in the baseline, whereas is output by the enriched system. Furthermore, the ordering of the words in the enriched system is slightly better than in the baseline output, even though the auxiliary is missing in the output of the window-50 system.

(5) EN one of the local residents even classified the quarrels with eastern european immigrants as a fight **for** survival.

BL eine der Anwohner selbst ein Kampf **für** das Überleben der Streitigkeiten mit
one of-the residents even a fight for the survival of-the quarrels with
osteuropäischen Migranten eingestuft .
the eastern-european immigrants classified

W50 eine der Anwohner sogar eingestuft die Streitigkeiten mit osteuropäischen
one of-the residents even classified the quarrels with eastern-european
Einwanderern wie ein Kampf **ums** Überleben.
immigrants as a fight for survival

In sentence (5), the translation of the phrase *fight for survival* is understandable in the baseline translation, but the preposition *für*, the predominant literal translation of *for*, is undeniably a poor choice in this context. The variant generated by the enriched system, the portmanteau preposition *ums* (*um+das*), leads to a more fluent translation.

5.5 Translation quality of prepositions

This section presents a more in-depth evaluation of the translation quality of prepositions, which also serves the purpose of illustrating typical problems when translating prepositions. As BLEU is not an optimal metric to assess potential, fine-grained improvements on a linguistic level, an additional manual evaluation of the translation

	for	**on**	**in**	**at**
Total prepositions	59	48	110	30
Correct in baseline	40	24	81	15
Correct in window-50	42	25	110	30

TABLE 5.6: Comparison of the number of correctly translated prepositions in the baseline output versus the best enriched system.

quality of prepositions is carried out. In this evaluation, sentences containing the prepositions *for, on, in* or *at* translated differently in the baseline and the window-50 system are compared. The sentences to be evaluated are restricted to a length of 5 - 20 words. The test set also includes sentences where the translation of the English preposition as a "null" preposition is possible or required, as illustrated in examples (6) and (7).

(6) that lead to a knock-on **fall in exports** to western europe

das führt zu einem erheblichen **Rückgang der Exporte** nach Westeuropa
that lead to a considerable fall the$_{Gen}$ *exports to western europe*

In sentence (6), the preposition in the phrase *fall in exports* needs to be translated as a German genitive phrase, as in the translation in the example. Alternatively, a translation containing a preposition is possible, for example *Rückgang an Exporten*. Translating English prepositional phrases of the form *noun-prep-noun* as a German genitive noun phrase is rather common; this is mostly the case for *noun-of-noun* phrases.

(7) ... has again **commented on** the problem of global warming

hat erneut ∅ **das Problem** der globalen Erwärmung **kommentiert**
has again ∅ the problem of-the global warming commented

In contrast, the verb in the English sentence (7) subcategorizes the preposition *on* that is not part of the subcategorization frame of the German translation *kommentieren*, which just requires a direct object. Alternatively, choosing a different verb as translation for *comment*, for example *sich äußern*, leads to a construction requiring the preposition *zu*.

The fact that often several translation variants are possible, mostly depending on the verb, makes it difficult to compare the system's output to the reference translation containing possibly a different verb in combination with the associated preposition. Thus, instead of matching the translated prepositions with the reference translation, a preposition is considered to be translated correctly if the produced PP or NP (with a "null" preposition) is an acceptable translation given its context in the sentence. The results in table 5.6 show that there is only little difference between the baseline and the system enriched with noun classes, even though the enriched system is slightly better.

In general, it is impossible to observe a systematic behaviour or a "pattern" of prepositions that are handled in a specific way in the baseline or the enriched system. However, there seems to be a type of prepositions that is particularly difficult to translate, namely prepositions with a predominant literal meaning occurring in a comparatively infrequent subcategorized context. Due to the predominant literal sense, they are often mistranslated (both in the baseline and the enriched systems), as illustrated by the following example:

(8) EN for example, germany has been **criticized for passivity**

MT beispielsweise hat Deutschland ***für Passivität kritisiert** worden
for-example has Germany *for *passivity criticized been*

REF **wegen Passivität** wurde zum Beispiel Deutschland **kritisiert**
because-of passivity has been for example Germany criticized

The preposition *for* is often used with its literal sense and thus can often be translated literally by the preposition *für*. However, being subcategorized by *criticize*, as in (8), *for* expresses a cause, requiring a translation other than *für*. While a translation with *für* is generally understandable, it is definitely ungrammatical. A correct translation here is the preposition *wegen* ('because of'), as can be seen in the reference translation. The same tendency to being error-prone can be observed for similar constructions such as *detain **for** corruption* → ***wegen** Korruption verhaften* or *to look **for** sth.* → ***nach** etwas suchen*.

These examples provide some interesting insights into the complexity of translating prepositions. Depending on the respective context of a prepositional phrase, different factors such as the relation of being a merely functional preposition versus conveying a meaning in combination with selectional preferences, such as the noun class of the involved noun, seem to play roles of varying importance. As not all factors are equally important in every constellation, the rather inflexible annotation method is not able to capture relevant information as required by the respective context, but always provides the same type of information at the same level granularity.

5.6 Related work

The handling of prepositions is a challenging problems in machine translation, but so far, research focusing on the translation of prepositions has been mostly reported for rule-based systems. Gustavii (2005) builds a classifier on bilingual features and selectional constraints to correct the output of a rule-based translation system for the language pair Swedish–English based on bilingual features. The classification task consists in recognizing contexts where a "default preposition", such as the most frequent translation of the source preposition, should be overridden with a more suitable other preposition. The classifier is built via transformation-based learning, and is applied to simulated MT output that consists of the target-side part of parallel data, in which the modeled prepositions are replaced by the "default translation" of the source preposition. Applying the classifier leads to an increased amount of correct prepositions; the system has however not been tested on actual machine translation output.

Naskar et al. (2006) outline a method to handle prepositions in an English–Bengali translation system. They use WordNet in combination with a bilingual example base covering a set of idiomatic prepositional phrases, but do not report any evaluation.

Huang et al. (2006) present a syntactically motivated strategy to relabel trees in the training data for a hierarchical SMT system translating from Chinese into English. Their annotation strategies lead to considerable improvements in BLEU. Assuming that the syntax trees provided by the Penn Treebank are too general, they propose to relabel the trees such that they capture more grammatical distinctions that are relevant for translation. They distinguish between two forms of annotation: external and internal. Internal annotation contains information about a node and its relatives (such as parent or sibling nodes) that are otherwise not accessible, and consists of lexical and tag information. To integrate lexical information, a lexical item, i.e. the terminal word, is annotated to its parent node and further ancestors in order to make the ancestor less general. In particular, this relabeling strategy is applied to prepositional phrases by annotating the preposition (the terminal word) to the parent node and the PP grandparent node, which results in an improved BLEU score. The same strategy is also applied to other categories, such as determiners, auxiliary verbs and conjunctions. In contrast to the lexical information propagating the terminal word to its ancestors, the tag information uses a non-terminal category to provide more specific information at a higher node, but not as specific as the particular word. This feature mainly aims at improving tense and auxiliary errors, for example by annotating the POS class of a verb at the VP level. Lastly, the external annotation provides more information about sister and parent nodes in order to provide more general context information.

The tree annotation strategies proposed by Huang et al. (2006) are obviously related to the ideas presented in this chapter; and in fact, we adopt their suggestion to annotate prepositions at PP-level. However, there are two main differences to the work presented in this chapter: first, their annotation is mainly syntactically motivated, whereas the annotation of noun classes is more semantically motivated. Furthermore, their annotation is rather a relabeling strategy that shifts already existing information between different parts in the tree, whereas the annotated noun classes are obtained externally.

Husain et al. (2007) model the translation of prepositions from English to the Indian languages Hindi and Telugu, which both have different criteria than English for selecting prepositions. Illustrated by examples, they motivate that both the governor of a preposition (such as the verb) and the governed noun in the prepositional phrase are required in order to determine an appropriate preposition. Their approach is rule-based and consists of extracting context and semantic information as a basis to determine the sense of the source preposition. Context information contains, for example, syntactic, lexical and morphological information. The semantic information is derived from WordNet and similar resources, and serves to identify categories such as person, time and place. The rules to select the correct sense of a preposition are manually constructed and are applied linearly. In their evaluation, they measure the precision of preposition translations according to the default sense (baseline) and the prepositions translated according to the created set of rules. The preposition translation rules outperform the baseline for both language pairs. An error analysis reveals that ambiguous analyses in WordNet are often responsible for a wrong classification of preposition senses, as an arbitrarily picked WordNet analysis can accidentally satisfy the constraints for an inappropriate preposition sense. This finding is similar to the observation in section 5.4.2 that word classes obtained with GermaNet tend to yield slightly worse results than with word classes obtained by clustering. Furthermore, Zollmann et al. (2011) integrate cluster information into syntactic SMT, although not for the translation of prepositions.

Bazrafshan et al. (2013) enrich non-terminals in a string-to-tree system with semantic role information, but find that this approach does not lead to an improvement over the baseline, assumingly because this annotation only provides partial information about the semantic structure of a predicate; cf. section 4.5 for more details.

Nădejde et al. (2016) argue that mistranslated predicate argument structures are often due to the fact that translation rules do not contain the lexical heads of verbs, nouns and prepositions; and in particular do not reflect preferences for potential fillers in the case of ambiguous predicates. To model selectional preferences between predicates and their arguments in a string-to-tree setting, they look at the semantic association between predicates and their arguments on the target-side. Selectional preferences are modeled for verb arguments, including prepositional phrases, as well as for noun phrases. The selectional preference features are trained on dependency triples (dependency relation, predicate, argument) extracted from parsed target-side data, and are integrated into the decoder. They evaluate their method in a German–English translation system, but find that they cannot improve automatic evaluation metrics, even though they report minor improvements, for example for verb translations, observed in a manual evaluation. They assume that general errors in the translation procedure, such as faulty target-side trees generated by the decoder, or incorrectly translated verbs, have a harmful impact on the functioning of the selectional preference features.

5.7 Discussion and summary

This chapter started with the hypothesis that noun class information is useful to model selectional preferences by means of creating semantic-informed and more precise translation rules for prepositions. While the modeling of prepositions is well motivated through the fact that prepositions are generally known to be difficult to translate, and selectional preferences are an important factor for the selection of prepositions, the outcome of the experiments indicates that the proposed annotation of noun classes is not optimal. In the following, different aspects of the annotation strategy are discussed.

Hard versus soft constraints The annotation of semantic classes on NP/PP nodes in the parse trees used to train a string-to-tree system amounts to hard constraints leading to overly specific rules in many contexts. As a compensation, the enriched systems were further extended with non-annotated rules in low-frequency contexts, as well as with PP-translation rules synthesized from existing translation rules with preposition preferences derived from monolingual data. Previous work, for example Marton et al. (2008), comes to the conclusion that soft constraints often work better than hard constraints. It might therefore be preferable to model selectional preferences differently, for example by means of feature functions that reward good choices, rather than inflexible markup in the grammar. This would, however, require extensive changes to the decoder.

Level of generalization Another problem with the presented approach is its implicit prerequisite of an optimal level of generalization to represent information relevant for selectional preferences. While experimenting with different levels of granularity aims at identifying a suitable level of information generalization, it is impossible to find a generally applicable optimal level to represent noun class information. This is in line with semantic research on selectional preferences as verb subcategorization features (Schulte im Walde, 2006; Joanis et al., 2008): across subcategorizing words, it is difficult to determine a generally valid level of generalization in lexical resources. Thus, annotating semantic classes of fixed granularity is not flexible enough to take into account the varying needs of different contexts – it always leads to rules of the same degree of specificity that are not adapted to different contexts.

Resources In the experiments, two types of resources to obtain noun classes were contrasted: semantic classes derived from the lexical resource GermaNet, and noun clustering. None of the variants was found to provide optimal noun class information: WordNets in general are known to be very fine-grained and to contain many ambiguities, which makes it difficult to derive generally applicable noun groups (Navigli, 2006; Palmer et al., 2007).

In contrast, the window-based cluster analyses turned out to represent topics rather than a generalization over specific noun types and are thus do not provide the ideal type of information to model selectional preferences. As opposed to the unstructured information used for the window-based clustering, the syntactic dependencies constitute the type of information needed to determine appropriate prepositions, in particular the governing noun/verb and the noun in the PP. Therefore, a clustering analysis derived from syntactic features can be expected to better capture selectional preferences. However, as syntactic features are inherently more sparse than simple context windows, the clusters obtained with this method also did not lead to improvements, but even failed to reach the performance of the window-based systems.

Outlook This chapter thoroughly explored distributional and resource-based methods to obtain noun class information, but found that none of these variants is optimal. While each of these strategies has advantages, either on the level of quality or coverage, they also suffer from weaknesses, namely insufficient coverage or being too fine-grained (resource-based), or a clustering into topics rather than noun classes (distributional). Problems at the resource-level could be addressed, at least partially, by combining different methods. For example, GermaNet could provide an initial set of noun classes that is then expanded through distributional methods to obtain robust and wide-coverage clustering analyses. Combining the advantages of the considered resources could lead to a more promising strategy to obtain classes representing salient information on selectional preferences and constitutes a challenging task for future work, that is however beyond the scope of this thesis.

While the annotation of noun classes does not lead to improved translation results, it gives valuable insights to the many problems arising when translating prepositions: for example, structural differences (such as NP–PP) between source- and target-side can be challenging for SMT systems. Similarly, prepositions with a strong default translation appearing in a context where the default translation is not valid were found to be problematic. The following chapters 6 and 7 explore variants of an alternative approach that directly addresses these issues. Making more flexible use of source and target-side information in a classifier to *predict* prepositions given the sentence context, appropriate prepositions can be generated on the target-side, either as a post-processing step (chapter 6) or by means of contextually conditioned synthetic phrases (chapter 7).

Chapter 6

Target-Side Generation of Prepositions

The translation of prepositions is a challenging task, as has been illustrated in the previous chapter. To select a correct preposition, many factors on the source and target side need to be considered: for example, content-bearing prepositions depend mainly on the source side, whereas functional prepositions rather depend on target-side subcategorization requirements and selectional preferences. Structural differences between source and target side can pose an additional problem, and might result in redundant or missing prepositions. In this chapter, prepositions are treated as a target-side generation problem. This strategy specifically addresses structural differences between source and target-side: in a pre-processing step prior to training the translation system, all NPs and PPs are transformed into a PP with an underspecified preposition by substituting overt prepositions with placeholders, and by inserting placeholders for "empty" prepositions at the beginning of NPs. After translation, actual prepositions, including the empty prepositions to form NPs, are predicted on the translation output using rich source- and target-side features. The experiments indicate that the removal of prepositions from the translation system leads to a loss in discriminatory power, and that generating prepositions in a post-processing step is not sufficient. However, the placeholder representation provides a sound basis for modeling prepositions, in particular with regard to handling structural differences on the source and the target side. The experiments described in this chapter are published in Weller et al. (2015).

6.1 Motivation

The translation of subcategorized elements, in particular prepositions, is a difficult task in machine translation. This has been shown in numerous evaluations; for example Williams et al. (2015) find that mistranslated prepositions constitute one of the most frequent error types in their English–German translation system. Similarly, Popović et al. (2015) name the translation of prepositions as problematic based on an evaluation of several language pairs, including English–German translation.

Prepositions can be roughly categorized into two types: functional and content-bearing prepositions, with many prepositions falling somewhere between. Functional prepositions tend to convey only little meaning and are mainly determined by their governors, for example verbs (*to believe in sth.*) or nouns (*interest in sth.*). Additionally, the noun in the prepositional phrase also can play a role, for example the differentiation between *person* and *abstract concept* when translating *to learn from → lernen von/aus* in the previous chapter. As subcategorization frames are target-side specific, the realization of functional prepositions mainly depends on target-side restrictions.

In contrast, content-bearing prepositions are mostly determined by the source-side; their meaning must be conveyed in the translation while also meeting target-side constraints, such as illustrated by the following examples.

(1) to be in the box → in der Kiste sein
to be under the box → unter der Kiste sein

The prepositions in (1) express a simple geometric relation, and their translation is straightforward. In particular, the noun in the prepositional phrase can be substituted by other (concrete) nouns, such as *house, car* or *pancake*. However, there are often factors at play that influence the choice of preposition.

(2) to go to the cinema → ins Kino gehen
to go to the beach → an den Strand gehen

While the meaning of the preposition *to* is essentially the same in both sentences in (2), i.e. expressing a directionality, two different prepositions are required in the German translation, depending on the noun in the phrase.

(3) to go to work → zur Arbeit gehen
to go to the authorities → aufs Amt gehen

The prepositions in (3) can be considered as functional in lexicalized expressions – while *zur* ('to-the') in the first sentence still expresses a directional meaning as in the English expression, *aufs* ('on-the') cannot be interpreted literally, but is entirely determined by the target-side context.

As many prepositions cannot be classified as either functional or content-bearing, both target-side and source-side factors contain important information that needs to be considered when determining the realization of a preposition. Oftentimes, the relevant information is not directly accessible in a standard translation system. For example, subcategorization is difficult to capture in the language model or within the translation rule if the verb and its subcategorized elements are not adjacent.

With regard to subcategorization, an additional problem consists in the fact that a prepositional phrase in the source sentence does not necessarily need to be translated by a prepositional phrase, but can also be realized as a nominal phrase.

(4) to call for sth. → $\emptyset$ etw. erfordern
→ nach etwas verlangen

Depending on the verb in the translations given in (4), the subcategorized phrase is either realized as a noun phrase in accusative case (with an "empty" preposition), or as a prepositional phrase with the overt preposition *nach*.

The aspect that a phrase can be realized as either a prepositional phrase or as a noun phrase, depending on the subcategorization frame of the verb, has not been addressed previously: the use of subcategorization information in chapter 3 aimed at improving the prediction of case in the post-processing step of the inflection prediction system, but does not allow for the transformation of a PP into an NP or vice-versa – some of the examples in the evaluation showed that there are indeed problems due to NPs being realized as PPs or the other way round. The use of noun-class information in chapter 4 rather focused on semantic constraints induced by the noun classes, and did not put much emphasis on modeling the relation between the different subcategorized elements.

In this chapter, prepositions are treated as a target-side generation problem which moves the selection of prepositions out of the decoding step into a post-processing component. This allows to select an appropriate preposition when the sentence is fully translated and all relevant information is available, most importantly the verb. Thus, the prediction model can take into account target-side requirements while still having access to the source side.

During translation, an abstract representation is used as a placeholder for prepositions, serving as a basis for the generation of prepositions in the post-processing step. To generate target-side prepositions, all subcategorized elements of a verb are considered and allotted to their respective function – either as NP with an "empty" preposition or as PP with an overt preposition as determined in the prediction step.
In the experiments presented in this chapter, two aspects are of particular interest:

- relevant features to obtain a meaningful abstract representation of prepositions during translation
- the combination of source- and target-side features to predict prepositions on the translation output

While the BLEU scores of some of the experiments are encouraging, they fail to surpass the baseline. However, the abstract representation proposed in this chapter provides many opportunities to model complement types: In chapter 7, it will be used as a basis to generate prepositions (empty and overt) already *during* the translation step, instead of handling the creation of prepositions in a post-processing step *after* the translation, which will finally result in some improvement. The ideas presented in this chapter can thus be considered as a direct "stepping stone" to the approach presented in the following chapter.

6.2 Methodology

The method of generating prepositions on the target-side is integrated into an English–German morphology-aware phrase-based SMT system. As described in chapter 3, it first translates into an underspecified stemmed representation with a component to generate fully inflected forms.

An important factor in the prediction step is the prediction of grammatical case, which corresponds to determining the syntactic function of a phrase. The use of source-side features and subcategorization information presented in chapter 4 (Weller et al., 2013b) aimed at improving the modeling of case, but had the shortcoming of handling NPs and PPs separately without being able to change between the two categories. The method of generating prepositions, both "empty" or overt, based on placeholder prepositions extends the previous setting and effectively allows the generation of NPs and PPs.

6.2.1 Translation and prediction steps

To build the translation model, the target-side data is converted into the underspecified stemmed representation with stem markup containing information about number and gender on nouns (cf. section 3.2.3) in which prepositions are substituted with placeholders ("PREP"). Additionally, "empty" placeholders to mark the beginning of

input	stemmed SMT output	prep	morph. feat.	inflected	gloss
∅ ⟶	`PREP`	**∅-Acc**	–	–	
what	`welch<PWAT>`	Acc	Acc.Fem.Sg.Wk	welche	*what*
role	`Rolle<+NN><Fem><Sg>`	Acc	Acc.Fem.Sg.Wk	Rolle	*role*
∅ ⟶	`PREP`	**∅-Nom**	–	–	
the	`die<+ART><Def>`	Nom	Nom.Masc.Sg.St	der	*the*
giant	`riesig<ADJ>`	Nom	Nom.Masc.Sg.Wk	riesige	*giant*
planet	`Planet<+NN><Masc><Sg>`	Nom	Nom.Masc.Sg.Wk	Planet	*planet*
has	`gespielt<VVPP>`	–	–	gespielt	*played*
played	`hat<VAFIN>`	–	–	hat	*has*
in ⟶	`PREP`	**bei-Dat**	–	bei	*for*
the	`die<+ART><Def>`	Dat	Dat.Fem.Sg.St	der	*the*
development	`Entwicklung <+NN><Fem><Sg>`	Dat	Dat.Fem.Sg.Wk	Entwicklung	*development*
of ⟶	`PREP`	**∅-Gen**	–	–	
the	`die<+ART><Def>`	Gen	Gen.Neut.Sg.St	des	*of-the*
solar system	`Sonnen System<+NN><Neut><Sg>`	Gen	Gen.Neut.Sg.Wk	Sonnensystems	*solar system*

TABLE 6.1: Prediction of prepositions, morphological features and generation of inflected forms for the lemmatized SMT output. German cases: Acc(usative), Nom(inative), Dat(ive), Gen(itive).

a noun phrase are inserted based on constituency-parses (Schmid, 2004). To obtain a symmetric data structure, empty prepositions at the beginning of noun phrases are also added on the source side, based on dependency parses (Choi et al., 2012).

The example in table 6.1 illustrates the translation and prediction steps: the columns “input” and “stemmed SMT output” show the the English input data and the stemmed output of the SMT system, which contains 4 basic placeholders “`PREP`”. In the next step, values for the prepositions, as well as grammatical case are predicted (column “prep”). The first two phrases represent a the direct object (*welche Rolle*: 'what role') and the subject (*der riesige Planet*: 'the giant planet'), and are realized accordingly as noun phrases with empty prepositions in accusative and nominative case. This also corresponds to the English structure, where the respective phrases are noun phrases with the same syntactic function. The next phrase (*bei der Entwicklung*: 'in the development') is realized as prepositional phrase with the preposition *bei*. Note that the “translation” of *in* → *bei* is not the expected default translation of *in*, but required in this context. The placeholder heading the last phrase (*des Sonnensystems*: 'of-the solar system') is predicted as an empty preposition in genitive case, in contrast to the English *of*-prepositional phrase. After predicting prepositions and grammatical case and removing empty prepositions, the MT output corresponds to the standard stemmed representation as presented in chapter 3. To inflect the stemmed sentences, the morphological features are computed (column "morph. feat.") and inflected forms are generated with SMOR (column "inflected").

The component to predict prepositions is thus integrated into the inflection prediction system in a way that the component responsible for the generation of inflected forms does not need to be modified, while at the same time the modeling of prepositions benefits from the abstract representation in two ways: first, it allows for an easy handling of portmanteau prepositions (e.g. *zum* = *zu+dem*: 'to-the') as inflected forms are generated at the very end of the pipeline, and their surface realization can be adapted as required. Second, all subcategorized elements are available in an abstract form and can be allotted to their respective functions (object/subject NPs, PPs) and inflected accordingly. Furthermore, the handling of structural mismatches between the source and target side is simplified, as the insertion of empty placeholders provides a better basis for a clean alignment. Typical structural mismatches concern different subcategorization frames (with/without preposition) in source and target language, such as in (5), where the target-side verb needs to be known in order to decide for the realization of the argument:

(5) to pay attention <u>to</u> sth. → <u>auf</u> etw. achten
→ <u>∅</u> etw. beachten

For prepositions governed by a noun, a typical form of alternation can be exhibited by phrases of the form *noun von/an/... noun* ('noun of noun')[1], which are often roughly equivalent (6), but not necessarily interchangeable in every context.

(6)
Rückgang der Exporte → decrease of-the exports
Rückgang von/an Exporten → decrease of/in exports
Exportrückgang → export decrease

To determine the type and case of each complement, the prediction model can make use of the full source-side and target-side sentence context. Resolving the category of complement *after* translation aims at avoiding problems observed in previous evaluations, where the translation was "stuck" with a combination of verb + incorrect complement type that the system could not recover from.

[1] Additionally, such structures can be expressed as a compound noun.

6.2.2 Prediction models

To predict prepositions, source- and target-side features are combined info a first-order linear chain CRF that provides a flexible framework to make use of the different knowledge sources. The target-side features mainly address functional prepositions, for example through the use of distributional information about subcategorization preferences, whereas source-side features (for example the aligned source-side preposition) tend to be more important to convey the meaning of a content-bearing preposition. However, the features are designed to not require an explicit distinction between the categories of *subcategorized* or *content-bearing*, as the model is optimized on the respectively relevant features for each context during training.

During model training and in the prediction step, relevant information such as the governing noun/verb is presented in a refined form, in contrast to a standard SMT system where relevant information *can* be available (such as the immediate context in a translation rule or the n-gram considered by the language model). Furthermore, even if the required information is observed by an ngram or within a translation rule, it is not represented in its "pure form", but there are variations such as adjectives or determiners that lead to different observed n-grams/phrases, thus preventing the model from generalizing. Additionally, the features used by the prediction model can bridge over large gaps between the verb and its subcategorized elements, and thus allow to capture relations that are out of the scope of translation rules and the language model.

6.3 Abstract representation of prepositions

In addition to providing a means to model complement types by target-side generation of prepositions, another important objective of the reduced representation of prepositions consists in simplifying and generalizing the SMT system. However, the experiments (cf. section 6.5) will show that that replacing prepositions by simple placeholders has a negative effect on the general translation quality. The effect that an overly simplified translation system loses discriminative power has also been observed by Toutanova et al. (2008): their results indicate that keeping morphological information during translation can be preferable to removing it, despite the increased vocabulary. Thus, as extension to the basic approach with plain placeholders ("PREP"), the placeholders are enriched such that they contain more relevant information and (at least partially) represent the content of the preposition while still being abstract. The annotation added to the prepositions is mostly syntactically motivated. The annotation variants are listed below:

Grammatical case The placeholder is enriched by annotating the *case* of the preposition it represents: for overt prepositions, case is often an indicator for its semantic content (such as direction/location), whereas for empty prepositions, case indicates the syntactic function of the NP.

Grammatical case + governor The annotation of case is further extended by marking whether the placeholder is governed by a noun or a verb, in order to provide a rough, more global information indicating whether the phrase is directly subcategorized by a verb or part of a complex noun phrase.

Grammatical case + functional/non-functional This annotation variant takes into account whether a preposition is functional or content-conveying: based on the entries of a subcategorization lexicon (Eckle, 1999), placeholders representing prepositions in a subcategorized context are annotated as such in addition. This annotation aims at providing a more global context, as well as establishing a separation between prepositions used "literally" as a content-conveying preposition vs. "non-literally" in a subcategorized context.

Mixed representation Based on the information annotated previously (functional/non-functional), a system containing both placeholders and regular prepositions is created: assuming that functional prepositions contribute less in terms of meaning, these continue to be represented by placeholders annotated with case and the type of governor (verb/noun). For all non-functional prepositions, the actual preposition with the same annotation (case + type of governor) are kept. This mixed representation of actual and abstract prepositions aims at separating the set of prepositions into those that can (presumably) be translated in a straightforward way, and those that mostly depend on complex target-side restrictions that might not be captured entirely in the translation model. In the prediction step, there will be a further distinction between predicting the values for all prepositions, as opposed to only the placeholder prepositions.

6.4 Predicting prepositions

In this section, the features used for the prediction of the placeholder prepositions are explained and the prediction quality is evaluated on clean data.

6.4.1 Features for predicting prepositions

To predict the values for the placeholder prepositions, source-side and target-side features are combined into a first-order linear-chain CRF. Table 6.2 shows the features used to train the prediction model. In addition to target-side context consisting of adjacent stems and part-of-speech tags (5 words to the left/right side), three feature types are used: (1) source-side features, (2) projected source-side features and (3) target-side subcategorization information, which are explained in more detail below.

Source-side features This set of features is obtained through word alignment and on the basis of dependency-parses on the source side. The source-side features comprise

- the word aligned to the German placeholder preposition: this can be a source-side empty or overt preposition (cf. "prp" in the column "source-side" in table 6.2
- the governing noun or verb (a verb in the example, cf. column "g.verb")
- the governed noun in the English phrase and its syntactic function in relation to the governor (cf. "func,noun")

Projected source-side features For this set of features, the extracted source-side features are projected to the target side using word-alignment, see column "projected source-side". The projection step allows to identify verb-object and verb-subject pairs on the target side. Parsing the SMT output is not an option, as it is not only non-fluent, but also in stemmed representation – deriving inflection-relevant features from parsed data that needs to be inflected prior to parsing would be a chicken-and-egg problem.

Subcategorization information The third feature type, target-side subcategorization information (column "target-side subcat") provides subcategorizational preferences for the observed verb in form of *preposition-case-verb* triples (cf. section 6.5.1 for more details on the extraction method and the corpora used). For each preposition that is a possible value for the placeholders, including the empty preposition, the co-occurrence

stem	**gloss**	**source-side**			**projected source-side**		**target-side**	**label**
		prp	**func,noun**	**g.verb**	**noun**	**g.verb**	**subcat**	
aber	*but*	–	–	–	–	–	–	-
PREP	*PRP*	$\emptyset$	subj, we	endure	wir	leiden	$\emptyset$-Nom:5 $\emptyset$-Acc:0 *unter-Dat*:4	$\emptyset$-Nom
wir	*we*	–	–	–	–	–	–	Nom
leiden	*suffer*	–	–	–	–	–	–	-
...	...	...	...	...	...	...	...	...
auch	*too*	–	–	–	–	–	–	-
PREP	*PRP*	$\emptyset$	obj, effect	endure	Treibhauseffekt	leiden	$\emptyset$-Nom:5 $\emptyset$-Acc:0 *unter-Dat*:4	unter-Dat
die	*the*	–	–	–	–	–	–	Dat
Treibhaus	*greenhouse*	–	–	–	–	–	–	Dat
effekt	*effect*							

TABLE 6.2: Prediction features in the training data. Source-sentence with inserted empty prepositions:"*..., $\emptyset$ we too are having to endure $\emptyset$ the greenhouse effects*". The entries in the column "stems" contain the full stem markup (POS tags + number/gender information on nouns), but are represented in a shortened form for better readability.

frequency of the verb with the respective preposition-case combination is listed. As the CRF interprets numbers as strings, the frequencies are bucketed[2] into values ranging from 0 (no evidence) to 5 (high amount of observations); the example in table 6.2 shows only entries for the values *$\emptyset$-Nom, $\emptyset$-Acc* and *unter-Dat*, omitting the remaining combinations due to space reasons. The subcategorization information tells the model that, for example, *unter etwas leiden* is more probable than $\emptyset$, even though the English sentence contains no preposition (*to endure sth.*).

In addition to triples of the form *preposition-case-verb*, tuples of the form *noun-noun$_{Gen}$* are integrated to help the model decide whether two immediately adjacent phrases should be realized as a complex noun phrase with a genitive modifier, or rather as two adjacent NPs, for example NP$_{Acc}$ NP$_{Dat}$ (direct/indirect object). This type of information is not displayed in table 6.2, but is contained in the subcategorization information.

Feature interaction To train the model, individual features as explained above can be combined into tuples and triples, such as *projected-verb + projected-noun + $\emptyset$-Nom* or *EN-function +$\emptyset$-Nom* , from which the model can derive preferences. Through such feature combinations, the model can learn that an English subject is fairly likely to be realized as a subject (*$\emptyset$-Nom*) in many contexts.

In the example in table 6.2, the model can learn that the second placeholder is unlikely to be realized as a direct object (which is the role of the phrase in the English

[2] The frequencies are bucketed to the powers of ten: [0]: $f = 0$; [1]: $f < 10$; [2]: $10 \leq f < 100$; [3]: $100 \leq f < 1000$; [4]: $1000 \leq f < 10000$, [5]; $f \geq 10000$

sentence), as there is no evidence for an accusative object. Instead, there is a strong preference for the combination *unter+Dat*. Furthermore, the projected noun (*Treibhauseffekt*) is an unlikely subject for *leiden*, and thus rules out the possibility of *∅-Nom*. For the first placeholder preposition, all features point to a realization as subject – the position at the beginning of a clause (marked by the conjunction *aber*) just before the verb, being aligned to the English subject and general evidence for the verb to occur in combination with a nominative phrase. This example shows how the features interact, and in particular that no explicit distinction between subcategorized and content-bearing prepositions is required. Furthermore, it can be seen that the projected source-side features bridge the gap between the verb and the placeholder realized as *unter* (middle part of the sentence omitted in the table).

6.4.2 Evaluation of prediction accuracy

The quality of the prediction model determines to a large degree the success of generating prepositions in SMT output. Thus, before integrating the prediction component into the machine translation pipeline, the quality of the prediction model is evaluated in a clean data setting. Table 6.3 shows the performance of the prediction task on the tuning set (3000 sentences news data). The column "prep+case" lists the accuracy of predicting both the preposition and the case it subcategorizes; the entries in column "prep" show the results for predicting only the preposition. For the prediction, a model relying on source-side features and projected source-side features (line 1 in table 6.3) is contrasted to an extended model that also has access to subcategorization features (line 2 in the table); both features also use stem and POS information of adjacent words.

The subset of the features used in model (1) can be considered as indispensable for the prediction task: source-side information is crucial to reproduce the meaning of content-bearing prepositions, wheres the projected target-side features serve the purpose of indicating target-side restrictions. The addition of subcategorization information in model (2) can be interpreted as generalization going beyond the information given in model (1), as it is obtained from a corpus considerably larger than the data used to train the prediction model. However, the addition of the subcategorization information does not lead to an improvement over model (1).

Table 6.4 lists the prediction results for the set of prepositions to be modeled, ranging from 95 % to 22 %. The realization as empty preposition constitutes by far the majority. Looking at the list of the top-3 predicted prepositions, it can be seen that the realization

	Features	prep+case	prep
1	basic + source	73.58	85.76
2	basic + source + subcat	73.42	85.78

TABLE 6.3: Prediction results on clean data (tuning set: 3000 sentences of the news domain).

prep	acc.	top-3 predicted (freq)
∅	95.17	∅ (10235), in (134), von (95)
in	79.19	in (1123), ∅ (170), von (21)
gegen	78.72	gegen (37), ∅ (4), von (2)
zwischen	77.50	zwischen (31), ∅ (4), von (3)
vor	77.14	vor (81), ∅ (10), bei (3)
von	72.15	von (495), ∅ (86), in (29)
nach	68.70	nach (90), ∅ (22), in (4)
zu	64.67	zu (238), ∅ (60), in (21)
an	61.09	an (179), ∅ (47), in (22)
unter	60.71	unter (34), ∅ (12), von (4)
um	60.56	um (43), ∅ (17), für (3)
mit	59.67	mit (253), ∅ (92), von (18)
auf	59.56	auf (215), ∅ (59), in (32)
aus	55.38	aus (72), ∅ (25), von (19)
für	55.13	für (231), ∅ (106), von (23)
über	47.15	über (58), ∅ (26), von (8)
durch	24.66	∅ (32), durch (18), von (10)
bei	23.00	∅ (60), bei (46), in (37)
wegen	22.22	wegen (4), für (4), ∅ (3)

TABLE 6.4: Prediction results for the different prepositions and the top-3 predicted prepositions.

of an empty preposition instead of an overt preposition tends to be the most frequent error; similarly, the prepositions *von/in* ('of/in'), which occur with a generally high frequency, are often output instead of the correct preposition.

Interestingly, for the most difficult to predict preposition (*wegen*: 'because-of'), the preposition (*für*: 'for') is listed among the most typical errors – this is likely caused by constructions such as *criticize/sue for* (*wegen kritisieren/verklagen*), that were found to be particularly difficult to translate in a previous evaluation (cf. example (8) in chapter 5).

It is difficult to estimate whether the quality of the used prediction models is sufficient for our application. Tsvetkov et al. (2013) cite a study that gives an accuracy of 94-96% for human participants annotating determiners on NPs given a full sentence context and 83-88% when given a context of 4 words. While this task is different from our application, it can be regarded as related to a certain extent. Given these values, we consider our prediction accuracy of roughly 85% to be adequate, even though a large chunk of this value is contributed by the empty preposition.

6.5 Experiments and results

This section compares the results for the different variants of annotating the placeholder prepositions. In addition to measuring the translation quality with BLEU, the quality of the generated prepositions is automatically evaluated for a subset where the relevant elements (verb and noun) match with the reference translation.

6.5.1 Data and resources

The translation system is a standard phrase-based Moses system trained on 4.3 million sentences of parallel English–German data (based on the data released for the WMT'14 translation shared task[3]) with a language model built on 10.3 million sentences. For tuning the translation system and for the clean data prediction evaluation, 3000 sentences of news data (WMT'13) are used; the test set consists of 3003 sentences of the news domain (WMT'14).

To obtain the underspecified stemmed representation, the German data is parsed with BitPar (Schmid, 2004); for the morphological analysis and the generation of inflected forms, the tool SMOR (Schmid et al., 2004) is used. The prediction models for the inflectional features (case, number, gender, strong/weak) are built on the German part of the parallel data using the Wapiti toolkit (Lavergne et al., 2010), using the same setup as in the earlier experiments. The English side of the data is dependency-parsed with Choi et al. (2012) to allow for the extraction of the dependency relations used as features in the preposition prediction model.

The models to predict the prepositions are trained on half of the parallel data, as the larger set of labels to predict leads to a considerable increase in memory usage, making the use of all parallel German data intractable. The prediction model for prepositions is trained as a sequence model using the Wapiti toolkit. Setting the model up as a sequence model, in contrast to a maximum entropy model, allows to take into account decisions made earlier in the sentence, and even though the model considers previous decisions only on bigram-level, it still gives access to the prediction of an immediately adjacent phrase at the beginning of the current phrase. The subcategorization tuples are

[3] `http://www.statmt.org/wmt14/`

extracted from German web data (Faaß et al., 2013) and Europarl, using the extraction routine by Scheible et al. (2013).

6.5.2 Evaluation with BLEU

Table 6.5 shows the results of the baseline variants. The system *Baseline*$_{surface}$ is a baseline system operating on regular surface forms, without modeling target-side morphology. The system *Baseline*$_{infl}$*-(a)* is a standard inflection prediction system without any treatment for prepositions. A variant to the basic inflection prediction system is system *Baseline*$_{infl}$*-(b)*, where all prepositions are first removed from the output and then repredicted, using the two CRF prediction models described in table 6.4, i.e. a model built on basic context features, source-side features and projected source-side features, and a second model relying additionally on target-side subcategorization information. The reprediction step does not lead to much change in terms of BLEU – while this shows on one hand that the prediction step itself is not harmful, it also has to be noted that only modifying existing prepositions as a mere post-processing step is not sufficient. In particular, it is not possible to model empty versus overt prepositions.

Table 6.6 shows the results for the different variants of the placeholder system. Using the basic placeholder (□) representation (S1) leads to a considerable drop of BLEU in comparison to the baseline variants – this result indicates that prepositions are indeed important during the translation step, and carry some (semantic) information, despite often being merely functional or not translatable in a straightforward way. Enriching the basic placeholders with case (S2) leads to an improvement of ca. 0.4, showing that the abstract representation of the placeholders plays a significant role for the general translation quality, even if it is as simple as grammatical case.

Having considerably improved the translation result, the annotation of case is further enriched with the type of governor (verb or noun) in system (S3), but leads to a worse result in contrast to just annotating case. As extension, the status of the placeholder, subcategorized or not subcategorized, is added (S4). This leads to minor improvement, even though the observed difference is very small.

The last variant employs a mixed representation of actual prepositions and placeholders: Assuming that functional prepositions contribute only little in terms of meaning, only functional prepositions are represented by placeholders, whereas non-functional prepositions are kept. For the prediction step, two variants are compared: in (S5a), all prepositions are re-predicted; in (S5b), only the placeholder prepositions are predicted,

System	**Prepositions**	**BLEU**	**Prep-CRF**
Baseline$_{surface}$	–	16.84	–
Baseline$_{infl}$ (a)	–	17.38	–
Baseline$_{infl}$ (b)	re-predict	17.36 17.31	basic + (proj.) source + subcat basic + (proj.) source + subcat

TABLE 6.5: Baseline variants (case-insensitive BLEU).

	Representation of place-holders	**BLEU**	**BLEU**
		basic+(proj.) src	basic+(proj.) src+subcat
S1	□	16.81	16.77
S2	□+Case	17.23	17.23
S3	□+Case+(V|N)	16.91	16.89
S4	□+Case+(V|N)+subcat	17.09	17.08
S5a	□+Case+(V|N): functional prp +Case+(V|N): non-func.	17.12	17.06
S5b	□+Case+(V|N): functional prp +Case+(V|N): non-func.	17.29	17.29

TABLE 6.6: Results for place-holder systems (case-insensitive BLEU).

whereas the translated functional prepositions are kept. While this last experiment reaches the level of the baseline, it cannot outperform it.

The negative outcome of the set of annotation variants with information about the subcategorization status could be explained by the previous assumption that it is difficult, if not impossible, to categorize prepositions into disjunct sets of functional and non-functional – many prepositions are neither clearly functional, nor are they clearly non-functional, but fall somewhere in-between.

While none of the studied variants outperforms the baseline, the results can still be considered encouraging as they illustrate that (i) the representation of prepositions during the translation step is crucial and has a considerable influence on the general MT quality, and that (ii) applying the prediction step to a set of carefully selected prepositions has a positive impact on the translation results.

6.5.3 Evaluation of prepositions

The commonly used evaluation metric BLEU measures the n-gram overlap of the obtained translation with a reference translation. As BLEU rather reflects the general

	BL	**S2**	**S5**
verb$_{MT}$ = verb$_{REF}$	502	469	503
verb$_{MT}$ = verb$_{REF}$, noun$_{MT}$ = noun$_{REF}$	270	260	271

TABLE 6.7: Subsets where governing verb/governed noun are the same in MT output and reference.

	BL	**S2**	**S5a**	**S5b**
verb$_{MT}$ = verb$_{REF}$	245 48.8%	233 49.7%	261 51.9%	250 49.7%
verb$_{MT}$ = verb$_{REF}$, noun$_{MT}$ = noun$_{REF}$	179 66.3%	174 66.9%	188 69.4%	178 65.7%

TABLE 6.8: Percentage of correctly predicted prepositions for the subsets from table 6.7.

translation quality, and is not well suited to analyze specific phenomena in the translation output, we propose a simple method to automatically measure the translation accuracy of prepositions.

The automatic evaluation of prepositions is difficult, as the selection of a preposition is determined by its context, and thus depends on the actual sentence structure, the lexical choice of verbs and nouns, and so on. Thus, it is not possible to just compare the translated or predicted prepositions with those in the reference translation, as the details important for the choice of the preposition are likely to be different. To compare the quality of the translated prepositions in the baseline translation with the predicted ones in the best systems from table 6.6, the evaluation is restricted to instances where the relevant parts, namely the governing verb and the noun governed by the preposition are the same in both the translation and the reference. Looking at only the relevant context, it is possible to automatically evaluate the preposition accuracy.

In this evaluation, PPs governed by nouns (such as *noun-von-noun*: 'noun of noun') are ignored, as they can often also be formed as *noun-noun*$_{genitive}$ constructions, which turned out to be cumbersome for the evaluation. This type of alternation can be considered a regular pattern for complex noun phrases, and while the resulting structures are often more or less equivalent, this is not always the case. Furthermore, the criteria for the selection of prepositions in complex noun phrases can be less strict than the restrictions imposed in a verbal subcategorization frame, which occasionally allows for several correct prepositions, as illustrated in (7):

(7) the amount of → die Menge an/von
→ die Menge ART$_{Gen}$

The set for which to evaluate the preposition accuracy is determined based on the reference translations: For each preposition in the reference sentence that belongs to the group of modeled prepositions (cf. table 6.4 for a complete listing), the governing verb and the governed noun are identified on the basis of dependency parses. In the different translation outputs, the respective equivalent parts (preposition, noun, verb) are identified via word alignments, using the source sentence as a pivot between the reference and the obtained translation. The comparison between the words is made on lemma-basis.

The evaluation is applied to the overall best systems (S5a/b) and (S2); table 6.7 gives an overview of the amount of instances where the reference and the MT output contain matching relevant parts, differentiating between a match of all three elements (preposition, verb, noun) and a match between only verb and preposition. Note that the slightly lower number of preposition-noun-verb triples in system (2) that match with the reference translation is not per-se an indicator for inferior translation quality, as the selection routine does not take into account the possibility of synonyms.

Table 6.8 shows the accuracy for the respective subsets in the different system variants: The percentage of (presumably) correct prepositions is slightly higher than in the baseline system. However, the differences are too small to be truly meaningful. Interestingly, there is a noticeable difference between systems (5a) and (5b), which are derived from the same MT output, and are thus lexically equal. While (5a) obtains comparatively high scores for the accuracy of generated prepositions, it has a lower BLEU score than the variant (5b). This outcome can be taken as an indicator that the realization of prepositions is not necessarily correlated to BLEU scores.

The proposed method of comparing *verb+preposition+noun* triples against a reference translation is useful to assess the effect of modeling prepositions in SMT, but it is important to point out that this form of evaluation only gives partial insights. First, the evaluation is centered around prepositions observed in the reference translation, which excludes sentences with prepositions in the translation, but not in the reference. Furthermore, the reference translation is often structurally different from the input sentence, and consequently is likely to also differ from the obtained translations which mostly tend to preserve the source-sentence's global structure. Similarly, instances with lexical differences are also not considered.

Nevertheless, the restriction to *verb+preposition+noun* triples to represent the relevant context for a comparison constitutes a simple method to automatically evaluate the translation and generation of prepositions in SMT output. While the results are not overwhelming, they can still be considered as encouraging, as they show that it is generally possible to model prepositions in an SMT system, and that the proposed abstract representation of prepositions provides a sound basis for the modeling of prepositions if annotated with sufficiently salient features, for example grammatical case.

6.5.4 Examples

This section compares the output of the baseline translation with translations obtained with the system (S2), in which the placeholders are simply annotated with case. The selected examples illustrate how the generation component can overcome structural differences in source and target language (cf. examples (8) and (9)), or generate a correct preposition in comparison to a literally translated preposition in the baseline (cf. examples (10) and (11)).

(8) EN ... malmon 's team will have to improve **on** recent performances .

BL ... malmon das Team wird **über** die jüngsten Leistungen zu verbessern .
... malmon the team will over the recent performances to improve .

S2 ... malmon das Team hat ∅ die jüngsten Leistungen zu verbessern .
... malmon the team has ∅ the recent performance to improve .

REF ... muss sich das Malmon-Team im Vergleich zu den vergangenen Auftritten
... must -refl- the malmon-team in-the contrast to the past performances
auf jeden Fall steigern .
in any case improve .

In example (8), the verb *improve* is translated by *verbessern*, which subcategorizes a direct object instead of a prepositional phrase with *on*, as is the case for the English verb. Thus, translating the preposition *on* into an overt preposition (such as *über* in the baseline) is bound to lead to an incorrect sentence. In comparison, the system (S2) correctly generates no preposition, thus realizing this argument as a direct object. Note that the reference translation differs a lot from the translation output, both structurally and lexically – thus, this instance cannot be counted in the previously presented evaluation, and is not given credit by BLEU.

(9) EN outer space offers many possibilities for studying substances under extreme conditions ...

BL in den Weltraum bietet viele Möglichkeiten für das Studium ∅ Stoffe unter
in the space provides many possibilities for the $\textit{study}_{noun}$ *∅ substances under*
extremen Bedingungen ...
extreme conditions ...

S2 der Raum bietet viele Möglichkeiten zum Studium **von** Stoffen unter extremen
the space offers many possibilities for $\textit{study}_{noun}$ *of substances under extreme*
Bedingungen ...
conditions ...

REF Das Weltall bietet viele Möglichkeiten , Materie unter extremen Bedingungen
the universe offers many possibilities , substances under extreme conditions
zu studieren ...
to study ...

The sentence in (9) presents the opposite constellation – the direct object in the English sentence cannot be translated as a noun phrase in the sentence structure adopted in the translations. In contrast to the verb+NP structure in the English sentence, both translation systems (BL) and (S2) opt for a complex NP structure by translating *studying* into the noun *Studium*, presumably because a translation as verb would require extensive reordering, as the verb would need to go to the end of the sentence, as can be seen in the reference translation. While *Studium* can also govern a genitive phrase (i.e. without a preposition), the baseline translation is incorrect as this would require a definite article between the two nouns (*Studium der Stoffe*). A better alternative is the translation of the second phrase as the PP *von Stoffen* ('of substances').

(10) EN nowadays there are specialists **in** renovation to suit the needs of the elderly.

BL heutzutage gibt es Spezialisten **in** der Renovierung der Bedürfnisse der älteren
nowadays there are specialists in the renovation of-the needs of the
Menschen.
elderly .

S2 heutzutage gibt es Spezialisten **für** Renovierung , die die Bedürfnisse der älteren
nowadays there are specialists for renovation , that the needs of the
Menschen.
elderly .

REF heute gibt es auch **für** den altersgerechten Umbau Spezialisten .
today there are also for the age-appropriate renovation specialists .
'today, there are also specialists for the age-appropriate renovation.'

In (10), the preposition *in* is literally translated in the baseline output, whereas the sentence produced with system (S2) has the correct preposition *für* in this position. While both translations are not particularly good at a global level (for example, there is no translation whatsoever of *to suit*), the baseline sentence is essentially a very long noun phrase, from *Spezialisten* to *Menschen* at the end of the sentence, whereas the translation obtained with (S2) at least inserts some sort of clause boundary after *Renovierung* which ever so slightly increases the readability of the sentence.

(11) EN ... what role the giant planet has played **in** the development of the solar system

BL ... welche Rolle der riesige Planet gespielt hat , **in** der Entwicklung des
... which role the giant planet played has , in the development of-the
Sonnensystems
solar-system

S2 ... welche Rolle der riesige Planet gespielt hat **bei** der Entwicklung des
... which role the giant planet played has in the development of-the
Sonnensystems
solar-system

REF ... welche Rolle der Riesenplanet bei der Entwicklung des Sonnensystems gespielt
... which role the giant-planet in the development of-the solar-system played
hat
has

Similarly, the preposition *in* is translated literally in the baseline output in example (11). The prepositions *bei* is a better choice, even though the baseline sentence is understandable. As the elements relevant for the triple evaluation match between the reference and the MT outputs, this instance is counted in the evaluation for preposition accuracy.

6.6 Related work

Research on prepositions in machine translation has been mostly reported for rule-based systems, see also section 5.6 in the previous chapter.

Gustavii (2005) employ a conceptually similar approach to the one presented in this chapter; they build a classifier to correct the output of a rule-based translation system for the language pair Swedish–English based on bilingual features. We follow their lead in regarding target-side features as mostly relevant for selectional constraints of functional prepositions, while access to the original source-side preposition guarantees to reproduce the meaning of a lexical preposition. Their classification task consists in recognizing contexts where a “default preposition”, such as the most frequent translation of the source preposition, should be overridden with a more suitable other preposition. The correction rules are restricted to modifying the choice of prepositions, but cannot remove or add prepositions. The classifier is built via transformation-based learning, and is applied to simulated MT output that consists of the target-side part of parallel data, in which the modeled prepositions are replaced by the “default translation” of the source preposition. Applying the classifier leads to an increased amount of correct prepositions; the system has however not been tested on actual MT output.

Agirre et al. (2009) model the translation of prepositions in a rule-based Spanish–Basque translation system. Spanish and Basque are different with respect to the realization of verbal complements in that Basque is an agglutinative language that uses post-positions or case suffixes attached to the nouns to express the function of a complement, where Spanish typically uses prepositions and does not have explicit surface case marking. Their approach includes the translation of “zero prepositions” to model such structural differences, and is integrated into the open source *Matxin* system, a rule-based transfer system relying on deep syntactic analysis. Its preposition translation module is fed with output from the modules for syntactic analysis and lexical transfer. To model verbal complements, they rely on a rich set of linguistic information, including a subcategorization lexicon and attested *verb/post-position/noun* dependency triples. Three strategies for the translation of prepositions are compared: manually encoded selection rules, rules informed with information about the subcategorization lexicon, and a selection based on observed verb/post-position/noun triples. As the individual strategies are prone to coverage problems, a combination of approaches leads to the best results. As baselines, the authors use a dictionary (with the first entry being chosen as the translation), as well as the most frequent translation derived from a pre-processed and aligned corpus. The variants are evaluated on a test set of 300 sentences where sentences with errors on other levels in the translation system that are affecting the translation of prepositions, are filtered out, resulting in a set of 54 sentences. For this set, an improvement with regard to the translation quality of prepositions is reported. Shilon et al. (2012) extend this approach with a statistical component to rank translations; they report a gain in BLEU for a test set consisting of 28 sentences.

Even though the approach presented in this chapter does not improve over the output of a baseline system, the presented experiments are carried out and evaluated within a full-scale statistical machine system and compared to a strong baseline, both in terms of BLEU and a an estimation of translation accuracy of prepositions. In contrast, most previous work measures accuracy gains for simple baselines (such as the most frequent translation) on small or carefully selected test sets.

A related task to generating prepositions on the target side is the generation of determiners, which can be problematic when translating from languages without definiteness morphemes, such as Czech or Russian. Similar to prepositions, determiners can be considered as function words that are determined by the target-side context, while not being completely independent from the source context. Tsvetkov et al. (2013) create synthetic phrase-table entries with added/removed determiners to augment a standard phrase-table. A classifier trained on local contextual features is used to predict whether to add or to remove a determiner on the target-side of the modeled translation rules. They report an improvement in translation quality for a Czech–English and a Russian–English translation system. Another related problem is the translation of pronouns when translating from a pro-drop language; this requires to additionally identify the subject/object to which the pronoun refers, for examplePeral et al. (2003). The proper use of prepositions is also a typical problem when learning a new language – another related task is thus the error correction of second language learners, for example Rozovskaya et al. (2013), which also includes the correction of prepositions.

An important aspect of modeling linguistic phenomena in machine translation is the type of evaluation. There is general agreement that BLEU (Papineni et al., 2002), which measures the general translation quality on document-level, is not well suited to assess the modeling of linguistic phenomena. Most studies presented in the above-cited previous work evaluate their systems in terms of preposition accuracy on carefully designed test sets, or by enforcing constraints in rule-based MT systems to allow for a direct comparison between the baseline and a system enriched with a preposition module. In an SMT setting, a direct comparison of phrases is difficult, if not impossible, as there are often considerable differences between the outputs of two systems.

In the tradition of previous papers and as addition to the standard evaluation metric BLEU, the experiments in this chapter are also evaluated with respect to the accuracy of generated translations. We propose to do this automatically by considering only the immediately relevant components *preposition – governed noun – governing verb*, which are then compared with the respective triples in the reference translation. Conceptually, this is related to semantically focused metrics such as *MEANT* (Lo et al., 2011), as this evaluation goes beyond a "flat" n-gram matching, but aims at evaluating meaningful entities. MEANT relies on semantic role labeling applied to both the reference translation and the machine translation output to measure the translation accuracy of the semantic role fillers of each semantic frame. To avoid the problem of applying semantic role labeling to ill-formed MT output, and being mainly interested in the realization of prepositions at this point, the comparison of the relevant triples seems a sound compromise.

6.7 Summary

This chapter outlined a method to handle prepositions as a target-side generation problem: based on abstract placeholders, prepositions (including a special "empty" preposition to form NPs) are generated in a post-processing step in a morphology-aware English–German translation system.

By deferring the decision *how* a phrase should be realized (NP vs. PP, in combination with selecting the preposition) until after the translation step, structural differences between the source and target side can be effectively handled: first, transforming all NPs on the source and the target side into "pseudo-prepositional phrases" provides a better basis for word alignment as it allows for alignments between empty and overt prepositions, where there is not always an equivalent in the baseline setup. Furthermore, deciding on the realization of complements after the translation step allows to take into account the relevant context from the full source and target sentences, whereas standard translation systems only have access to a limited window during translation. For the prediction task, source- and target-side features are combined in a flexible way that does not require an explicit distinction between functional prepositions (where target-side context is likely more relevant) and content-bearing prepositions (where the source-side content is relevant in combination with potential target-side restrictions).

While some of the systems reach the level of the baseline, none of the systems is able to add some further improvement. However, the evaluation of *preposition-noun-verb* triples indicates that the prediction accuracy of prepositions is slightly better in the generation systems than in the baseline system, despite slightly lower BLEU scores.

The evaluation shows that the translation quality of prepositions is a problem that needs more attention; it also points out typical problems, such as structural mismatches between source and target side, or a tendency to translate prepositions literally if they have a predominant literal sense (such as *in*) in examples (10) and (11). As BLEU is not well suited to measure the quality of specific phenomena in SMT output, the translation quality of prepositions was estimated by considering only the relevant elements (*preposition-noun-verb* triples) in a comparison with the reference sentence. While this only gives a partial insight into the quality of the produced prepositions, it is a first step towards an automatic evaluation of prepositions – with prepositions being a persistent problem in machine translation, a (simple) automatic metric that takes into account morpho-syntactic as well as semantic aspects of the realization of prepositions would be a useful instrument.

An important finding of the experiments in this chapter is that a meaningful representation of the placeholder prepositions is crucial for the general translation quality. In particular, the annotation of case resulted in the best score among the placeholder-only systems – this information can be considered as a "light semantic" annotation that has the additional benefit of being easily obtainable. A more semantically motivated annotation to represent the semantic class of a preposition by a more meaningful representation, such as temporal or local, constitutes an interesting and challenging idea for future work, but will not be further studied in the remainder of this thesis.

The analyses in this chapter suggest that the proposed placeholder representation provides a sound basis to handle prepositions in SMT, in particular with regard to tackling the problem of variations between NPs and PPs. However, the fact that the translation performance is very sensitive to the representation of prepositions indicates that the complete removal of prepositions at translation time might result in a too severe loss of information that the system cannot recover from, regardless of the quality of the prediction model.

A solution might thus be to benefit from the positive aspects of the placeholder representation, but to generate actual prepositions at an earlier stage in the translation process, such that there is no lack of information at decoding time. The next chapter outlines a strategy that, instead of generating prepositions in a post-processing step, creates contextually conditioned phrase-table entries with complement realizations (in the form of prepositions or "empty" case-markers) to provide the translation system with a set of optimal translation rules for different contexts. As the synthetic phrase-table entries represent translation options for a particular context, they can steer the translation model away from predominant, but inappropriate translations options in the respective contexts, thus addressing another problem observed in the evaluation.

Chapter 7

Using Synthetic Phrases to Model Complement Types

While generating prepositions as a post-processing step is not optimal due to the loss of discriminatory power at translation time resulting from the removal of prepositions, the experiments in chapter 6 indicated that the concept of generating prepositions as required by the sentence context is promising, as it addresses two main challenges: structural differences between source and target-side are modeled by temporarily transforming every phrase into a PP, and the preposition can be selected according to token-level requirements, independently from a preposition's type-level translation probabilities. Using the placeholder representation from chapter 6, the prediction step is now shifted from post-processing to an earlier stage to provide the translation system with the fully specified translation options. This is achieved by integrating synthetic phrase-translation pairs containing generated prepositions conditioned mainly on rich source-side context into the translation model. These synthetic phrase-table entries are generated to provide an optimal selection of phrase-translation pairs for the given context.

This chapter presents two methods to model complement types: First, a translation model built on a "preposition-informed" representation (with pseudo-prepositions inserted at the beginning of noun phrases, but all other prepositions intact) already leads to an improvement. The second strategy is the modeling of complement types by means of synthetic phrases, which can additionally improve the translation results in some of the explored settings. The experiments described in this chapter are published in Weller-Di Marco et al. (2017).

7.1 Motivation

The output of machine translation is often incomprehensible because in the translation process, syntactic functions often get confused and complements are realized incorrectly or arranged in a meaningless way. This can mean the generation of an incorrect grammatical case of a phrase, the generation of a prepositional phrase instead of a noun phrase (or vice-versa), and finally the selection of a wrong preposition. However, the realization of complements represents important information at the syntax-semantic interface: the grammatical case of a noun phrase expresses its syntactic function as well as its semantic role; and the choice of preposition in a prepositional phrase sets its semantic role.

While the lexical context of a target-language phrase is defined by the respective phrase in the source sentence, the exact choice of (functional) preposition and case mainly depends on the target-side context, and is in particular determined by the subcategorization frame of the target-side verb. As already illustrated in the previous chapters, source- and target-side subcategorization frames cannot be expected to be isomorphic, but differ with regard to grammatical case and on the level of realization as noun phrase or as prepositional phrase.

Selecting the wrong complement type or an incorrect preposition has a major effect on the *fluency* of SMT output, but can also impact the perception of semantic roles. This is illustrated by the example in (1): when the subcategorized preposition *for* is translated literally into *für*, the meaning of the translated sentence (1-b) shifts, such that the book is no longer the object that John is searching, but rather the recipient of the search.

(1) a. John looks for his book.

b. *John sucht für sein Buch.
'John searches on behalf of his book.'

c. John sucht [∅ sein Buch]$_{\textbf{NP-Acc}}$.
John sucht [nach seinem Buch]$_{\textbf{PP-nach}}$.

(1-c) shows two possible translation options: the phrase *for his book* can be either translated into a direct object phrase with no preposition, or into a prepositional phrase headed by the preposition *nach*. Many prepositions tend to have a predominant translation, such as *for* → *für*, which is correct in many contexts, but unsuitable in settings as illustrated in the example above. Such translation options can be difficult to override, even when there are clues that a literal translation is wrong.

A related problem is that of coverage: even though prepositions are highly frequent words, the *meaning* of a preposition in a particular context, or its use in a subcategorization frame might not be sufficiently well represented in the translation model to be selected at decoding time. For example, it may be the case that the appropriate translation occurs in the phrase-table, but is practically inaccessible due to having a too low translation probability. The example in (2), showing an English input sentence and the translation obtained with the standard inflection prediction system as presented in 3.4.3, illustrates this problem.

(2) EN many critics of veganism warn in particular **of** the lack of vitamin b12

MT viele Kritiker des Veganismus warnen insbesondere der Mangel an Vitamin B12
many critics of veganism warn in-particular the lack of vitamin b12

REF Insbesondere **vor** dem Mangel an Vitamin B12 warnen viele Vegansimus-Kritiker
in-particular of the lack of vitamin b12 warn many veganism-critics

The translation of this sentence is essentially straightforward, the only difficulty is the translation of the functional preposition *of* into *vor*, as required by the subcategorization frame of *warnen*. The showed translation is mostly correct, save for the missing preposition *vor*[1]. A closer look at the phrase-table reveals that the predominant translation of *of* is the definite article (i.e. resulting in a German genitive phrase), followed by *von*. The translation option *of* → *vor* does exist, but is very unlikely to be selected during the translation process due to its low translation probability of 0.00067 (in contrast to a probability of 0.22 for the top-ranked translation option, the definite article). Similarly, longer phrases such as *of the*, *particular of* or *particular of the* have fitting translation options containing *vor*, which however all have very low translation probabilities and are thus unlikely to be actually used. Generally, translating a verb and its complement(s) by one phrase makes the selection of prepositions considerably easier: the top-ranked translation options for the phrase *warn of* are the expected *warnen vor* or variants thereof. However, the verb *warn* and the complement *of the lack* do not occur next to each other, but are separated by two words, making it more difficult to translate the verb and the preposition within one phrase. As there is no entry for the phrase *warn in particular of*,

[1] The structure in the translation, corresponding to that in the English sentence, is neutral, whereas in the reference translation, the constituent containing *the lack of vitamin b12* is emphasized through its position at the beginning of the sentence.

and with the translation options for *vor* being effectively inaccessible, the preposition *of* is bound to be translated incorrectly in this context.

Outline This chapter presents two approaches to improve the modeling of complement types. Based on the ideas outlined in the previous chapter, the first approach makes use of an abstract representation of "placeholder" prepositions that are added at the beginning of noun phrases on the source and target sides. The inserted placeholder prepositions lead to a more symmetric structure, and consequently to a better coverage of prepositions, as all NPs are effectively transformed into PPs. This allows for the alignment of prepositions in one language with a placeholder in the other language that previously had to remain unaligned or were aligned to adjacent phrases. Furthermore, the placeholder prepositions are annotated with grammatical case, and besides functioning as explicit phrase boundaries, they can provide flat structural information. In contrast to the experiments presented in the previous paper, all overt prepositions are kept during the translation process. Thus, no prediction of prepositions is necessary; and empty prepositions are just removed prior to the prediction of inflectional features. This variant of the placeholder representation (the "preposition-informed system") leads to a significant improvement in translation quality over the standard inflection prediction system.

The second approach aims at generating context-dependent translation options relying on rich context information. Using an abstract representation of prepositions, synthetic phrase-table entries containing prepositions predicted to fit the respective context are generated and integrated into the phrase-table. This method aims (i) at improving the choice of prepositions by conditioning on features extracted from the respective source context (i.e. at token-level); and (ii) at modifying the translation probabilities in the generated entries such that they favour context-appropriate translation options.

Generating phrase-table entries allows to create prepositions in contexts not observed in the parallel training data – in particular, it is possible to generate "missing" or "inaccessible" prepositions, as outlined in example (2). The resulting phrase-table entries are unique for each context and should provide an optimal selection of translation options for the problem of complement realization. Translation systems extended with synthetic phrases are significantly better than a baseline system without a component to handle prepositions, and can also slightly outperform the system with inserted placeholders, i.e. the preposition-informed system.

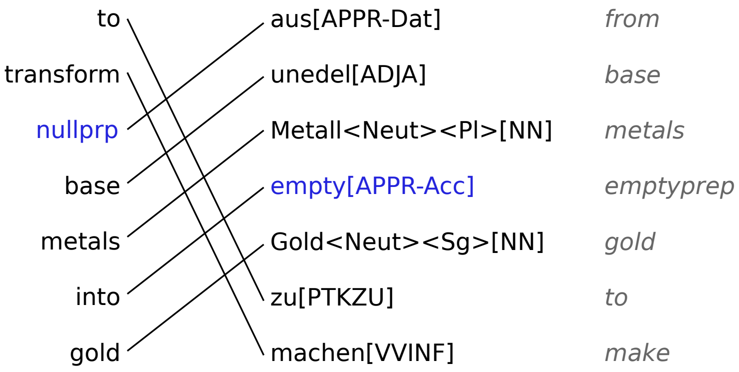

FIGURE 7.1: Example for preposition-informed representation with empty placeholders heading NPs on source and target-side.

7.2 Preposition-informed representation

As a basis for the modeling of complement types, the standard abstract stemmed representation (cf. chapter 3) is extended with pseudo-prepositions markers to indicate the beginning of noun phrases. This form of representation is very similar to the one introduced in the previous chapter, with the only difference that pseudo-prepositions are added at the beginning of noun phrases, while overt prepositions remain in the text, in contrast to being also replaced by placeholder prepositions. Based on the outcome of the previous experiments, these pseudo prepositions are annotated with grammatical case. The addition of pseudo-prepositions serves two purposes: first, structural differences on source and target side at the level of complement types can be adjusted as every phrase is transformed into a (pseudo-)PP. Second, the pseudo-prepositions can be considered as explicit phrase-boundaries providing information about syntactic functions in form of the annotated case.

Relying on parse structures, placeholders for empty prepositions are inserted at the beginning of noun phrases on both source and target side. The example in figure 7.1 shows structural mismatches in the training data, where the PP *into gold* on the source side corresponds to the NP `Gold<Neut><Sg>` on the target side. Similarly, the NP *base metals* on the source side is translated by the PP `aus unedel Metall<Neut><Pl>`. Without the placeholders transforming noun phrases into pseudo-prepositional phrases, the word alignments for such phrases contain unaligned overt prepositions, or imprecise one-to-many alignments such as "*into gold* → `Gold<Neut><Sg>`", resulting in phrase-table entries which are not universally applicable during translation.

The addition of placeholder prepositions leads to a cleaner word alignment as the empty preposition on one side can be aligned to an overt preposition on the other side, such as *nullprp* → `aus`. While this correspondence is obviously also not universally true, enabling the alignment between functional words such as pseudo-prepositions and overt prepositions aims at obtaining a better individual coverage of prepositions, while reducing the amount of prepositions being "lumped-together" with an adjacent content word, such as `aus unedel`. Additionally, placeholders between two immediately adjacent phrases such as $\texttt{[aus Metall]}_{PP}$ $\texttt{[Gold]}_{Dir.OBJ}$ provide an explicit phrase boundary with information about the syntactic function of the phrase, such as a subject (`EMPTY-Nom`) or direct object (`EMPTY-Acc`).

7.3 Creating synthetic phrase-table entries

The creation of synthetic phrases builds on the preposition-informed representation: based on intermediate placeholders representing both empty and overt prepositions, synthetic phrases are generated. Combining source-side and target-side features, unique phrase-table entries optimized for the given context can be synthesized.

7.3.1 Motivation and example

The preposition-informed representation presents a straightforward solution to handle structural differences between source and target side. While it already improves the translation quality, the issue of coverage and availability of translations options remains: the probability distribution of the existing translation options might not favour the required translation in a certain context; and in the worst case, the required translation might not even exist, either on word or on phrase-level. For example, the problem of translating *warn ... of* in example (2) is still not adequately solved in the preposition-informed system, and the translation produced is identical to that in example (2). As a solution, we explore the idea of synthesizing phrase-table entries, in order to adjust the translation options to token-level requirements in a way that allows to take into account relevant information from the entire source sentence, and the target-side context accessible in the relevant phrase.

The prediction of synthetic phrase-table entries is based on intermediate placeholders `PREP`, representing both empty and overt prepositions. This intermediate representation essentially corresponds to that employed for translation in the previous chapter, but it is only used as a basis for the prediction of actual prepositions. In the prediction step,

sentence 1: nullprp beginners look **for weapons** in different ways .

sentence 2: nullprp screenshot of the site that accepts nullprp orders **for weapons** .

	1	2	3	4	5	6	7	8	9	10	11		
	NP/PP	**tag**	**word**	**func**	**head**	**head**	**parent**	**parV**	**parV**	**parN**	**parN**	**best-5**	
	src	**src**	**src**	**src**	**src**	**trg**	**src**	**src**	**trg**	**src**	**trg**	**predicted**	
sentence 1	PP	IN	for	prep	weapon	Waffe	V	look	–	–	–	**nach-Dat**	0.349
												empty-Acc	0.224
												empty-Nom	0.206
												von-Dat	0.067
												für-Acc	0.064
sentence 2	PP	IN	for	prep	weapon	Waffe	N	–	–	order	–	**für-Acc**	0.559
												empty-Nom	0.184
												von-Dat	0.087
												nach-Dat	0.078
												empty-Acc	0.053

TABLE 7.1: Source and target side features for the prediction of placeholders in the phrase *for weapons* → `PREP Waffe<Pl>[NN]` in two sentences, using the top-5 five predictions; appropriate prepositions are bold. The prediction model corresponds to model (2) in table 7.
(Omitted: context of 3 words to each side of the placeholder.)

the generic placeholders `PREP` are realized as either overt or empty prepositions. Every phrase can thus be inflected as either NP or PP, containing those prepositions that the prediction system deemed most appropriate for the given context. For the sentence in (2), this means that the for the phrase-translation pair *of* → `PREP`, the generic placeholder can be predicted as the required `vor` as the top-entry, where previously the definite article and the preposition *von* where by far the most probable translation options. Thus, the translation system obtains the correct translation option for this sentence, while also keeping the default translation *of* → *von* for the many cases that require this translation.

Table 7.1 illustrates the concept of the token-based prediction step for the two sentences containing the phrase *for weapons*. The predominant literal translation of *for* is *für*, which is only correct in the context of sentence 2. When used in combination with the verb *look*, a correct translation is possible with either the preposition *nach* or with an empty preposition to form a direct object in accusative case. Thus, for the translation pair *for weapons* → `PREP Waffe<Masc><Pl>[NN]`, different and disjunct translation options need to be available for the two sentences. The prediction model should generate `nach-Dat Waffe<Masc><Pl>[NN]` and `EMPTY-Acc Waffe<Masc><Pl>[NN]`, conditioned on the context of the governing verb *look*, and `für-Acc Waffe<Masc><Pl>[NN]` in sentence 1 for the complex nominal phrase *orders for weapons*. Table 7.1 lists the respective top-5 predictions ranked by their prediction score, with appropriate predictions being marked in bold.

In particular, it is possible to generate phrases that have not been observed in the training data in combination with the source phrase. This is the case for the phrase `EMPTY-Acc Waffe<Masc><Pl>[NN]`: while `nach-Dat Waffe<Masc><Pl>[NN]` is already the 2nd-ranked translation option for *for weapons* (although with a considerably smaller translation probability than `nach-Dat Waffe<Masc><Pl>[NN]`), `EMPTY-Acc Waffe<Masc><Pl>[NN]` does not occur as a translation option in the preposition-informed system.

The format of the generated phrases corresponds to that of the preposition-informed system; with the only difference that for each source phrase, a unique set of target-phrases, possibly with new word sequences, is generated to provide an optimal set of translation options. Different sets of phrase-table entries for different contexts are distinguished by indices added to the words on the source side: for example, the particular preposition $\textit{for}_{123}$ can only be translated with the matching set of phrase-table entries, whereas $\textit{for}_{124}$ in another context has its own, different set of phrase-table entries.

7.3.2 Preposition prediction model

This section outlines how the model used for the prediction of the complement realizations is trained and applied to obtain the basis on which synthetic phrase-table entries can be created: relying on token-level source-side features, a set of n-best complement realizations is predicted, which are used to create the set of optimal phrase-table entries for the observed context.

Prediction features

Table 7.1 shows the set of features used to train a maximum-entropy classifier for the task of predicting prepositions in phrase-table entries. On the target-side, only the phrase is available, as further context would need to come from the sentence that is being translated, which is of course not available at the moment of training the prediction model. In contrast, the source-side features can be extracted from the entire sentence, and thus go beyond phrase boundaries. With the target-side context being restricted by the (often rather short) phrase length, the model relies heavily on source-side features that are centered around the preposition aligned to the German placeholder preposition. Based on dependency parses, the relevant information can be gathered from the sentence, with the aligned preposition as a starting point.

The features used to train the prediction model are listed below; the numbers refer to the columns in table 7.1:

1. type of the aligned phrase (NP or PP)
2. part-of-speech tag
3. lexical content of the preposition (overt or empty preposition)
4. syntactic function of the phrase in the source sentence (subject, object or prepositional phrase)
5. governed noun in the source-side phrase
6. projection of the governed noun (if available in the target phrase)
7. type of governor (verb or noun)
8. governing verb
9. projection of the governing verb (if available in the target phrase)
10. governing noun
11. projection of the governing noun (if available in the target phrase)
12. omitted in table 7.1: three words to the left/right side of the placeholder (depending on the phrase length) in the target phrase, to provide basic target-side context in addition to the projected source words.

Among these features, information about the source-side syntactic role and the governor, in particular the governing verb, are probably most important, as they set the most relevant context to determine the grammatical case and/or the correct preposition for the phrase. These features are not only used individually, but can also be combined as feature n-grams such as noun-verb tuples, or preposition-noun-verb tuples, which contain relevant information about subcategorization preferences when combined. For example, the content of an NP itself is in many cases not primarily relevant for the prediction of a placeholder[2], but the combination with other features (such as verb-noun tuples) provides useful information.

[2] In fact, experiments indicated that the use of features (5) and (6) individually tends to be harmful, whereas they provide useful information when used as part of feature tuples.

Being restricted to the often short phrases on the target-side is not without problems, in particular as the exact way of complement realization depends to a large part on the target-side verb. However, even if the target-side verb is not contained in the phrase-table entry, it is assumed that the verb on the source side is known: given the source-side verb and the syntactic function, the classifier can estimate the best complement realizations. Looking at the example in (2), the classifier should be able to predict `vor-Dat`, given only source-context *warn+of* for the phrase-pair *of* → `PREP`. As the phrase-table entries are generated for the top-5 predictions, they are not restricted to one particular verb in the translation, but represent the best complement realizations for the given (source) context, and are applied in combination with other relevant statistics provided by the SMT system, such as the target-side language model.

Training and prediction

For the training of the prediction model and to obtain the base-entries from which the synthetic phrases are generated, two variants of the target-side training data are required: the training data for the prediction model is derived from the *preposition-informed representation* (overt/empty prepositions + case). The model is then applied to *placeholder-representation* phrase-table entries, i.e. a phrase-table computed on data where all target-side prepositions have been replaced with a generic placeholder `PREP`.

Training The prediction model is trained as maximum-entropy model, with all extracted source/target/alignment triples from the preposition-informed system containing relevant prepositions as training data. In the training data, the preposition to be predicted is replaced with a generic placeholder `PREP`, and the tuple of *preposition+case* is used as label, i.e. the value to predict.

To obtain token-level source-side context, every word in the source-side data is marked with the unique identifier *SentenceNumber_WordPosition*. The standard Moses phrase-extraction routine is then applied to obtain all source/target/alignment triples. By means of the unique identifiers, the exact position of each phrase can be found in the source corpus, and the features from outside of the phrase can be extracted.

In this model, each preposition is treated as one instance, i.e. each preposition is predicted independently and decisions at previous instances do no not impact later predictions. If several prepositions occur in one phrase, all permutations of the respective n-best predictions are used as the basis for the generation of phrase-table entries.

Prediction The model is applied to phrase-table entries containing the generic placeholder PREP ("placeholder representation"): for each n-gram in the source sentence that contains a placeholder preposition, the relevant phrase-table entries are identified and the respective features are extracted from the sentence. Then, the prediction model is applied, and based on the top-5 predictions along with the prediction scores, phrase-table entries are created. Since the complement realization largely depends on target-side context, i.e. context in the translated sentence (such as the verb), that may not yet be available during the generation of the phrase-table, there are often different valid possibilities for the realization of a complement, and it is impossible to decide for one particular realization at the prediction step. The synthetic phrase-table entries are thus generated on the basis of the top-5 predictions[3] to represent the *selection of the most probable complement realizations* given the available context.

7.3.3 Building a system with synthetic phrases

The first step to create a phrase-table with synthetic phrases consists in building a phrase-table on data with generic placeholders on the target-side (i.e. placeholder representation), using the same word alignment as for the preposition-informed system. The entries in this table are then divided into two groups: entries with and without placeholders. Those entries containing no placeholders do not need any processing, and are kept for the final phrase-table including all scores such as translation probabilities and lexical weights. Phrases containing one or several placeholders undergo the prediction step, and phrase-table entries are created for the set of predicted complements (i.e. overt/empty prepositions with case). The format after the prediction/generation step corresponds to that of the preposition-informed system, meaning that the same language model can be used.

As a prediction for all phrase-table entries would be not feasible, the table is decreased in size by keeping only the top-20 entries according to the translation probability $p(e|f)$. This filtering is applied to the phrase-table of the preposition-informed system; the phrase-table entries containing placeholders are then selected accordingly. With this process of phrase selection, the synthetic-phrases system and the preposition-informed system rely on the same phrase inventory.

[3] This assumes one placeholder per phrase; in the case of several placeholders, phrase-table entries representing all permutations of the respective top-5 predictions are created.

7.3.4 Scores in phrase-table and reordering-table

A phrase-table contains translation probabilities and lexical probabilities between the source and the target phrase, each in the translation direction and the reverse direction. For the synthesized entries containing the predicted complement realizations, these four scores have to be computed in order to obtain a complete phrase-table entry.

The lexical weight of a phrase can be calculated from the lexical weights of the words occurring in the phrase (cf. chapter 2); it is thus no problem to compute the lexical weight of a new phrase as the vocabulary of individual words does not change. In contrast, the translation probability of a previously unseen phrase cannot be estimated based on word translation probabilities. Furthermore, as illustrated in example (2), the translation probability of an observed phrase pair might not be appropriate for the given context. When estimating the translation probability, two factors need to be taken into account: (i) how well the target phrase reproduces the content of the source phrase on a lexical level (it is important to keep in mind that a large part of phrase-table entries can be considered "incorrect" in the sense that one side is incomplete, or that they were derived from faulty word alignment); and (ii) how well the generated overt or empty preposition(s) fit into the target phrase given the sentence/phrase context.

These two aspects are modeled by combining the *placeholder translation probability* and the *classifier prediction score*. The underlying assumption is that the placeholder probability (i.e. the translation probability obtained from the placeholder representation) represents the approximate translation probability between a source and a target phrase independent of the actual preposition, thus representing appropriateness on the lexical level. The classifier score (referred to as "ME") indicates how well the generated preposition fits into the target phrase. It is important to keep these two factors balanced – ideally, a "good" translation option has a high placeholder translation probability representing a good lexical fit, as well as high prediction score indicating that the prediction model is confident about the predicted preposition. On the other hand, the effect of "boosting" a phrase pair with a low lexical correspondence by means of a high prediction score or vice-versa should be avoided. In particular the combination of a high prediction score in a generally bad translation option turned out to be harmful in the conducted experiments; and the methods of combining the scores seek to downplay the influence of too high ME scores. In the following, three variants of combining the placeholder probabilities and the prediction scores into features to be optimized by MERT training in the tuning step are presented:

Score variant 1 The placeholder translation probabilities and the ME scores (values between 0 and 1) are used as separate features. Additionally, an indicator feature keeps track of the number of predicted prepositions: non-synthesized phrases get a pseudo ME-score of 1, with the indicator feature set to $exp(0)$. For phrases with $n \geq 1$ prepositions, the ME scores are multiplied and the indicator feature is set to $exp(n)$. The indicator feature serves the purpose of balancing non-synthesized phrases with an ME score of 1 as opposed to synthesized phrases containing several predicted prepositions with a comparatively low ME score.

Score variant 2 Variant 1 is extended with the product of the placeholder translation probabilities and the ME score, to account for cases where lexically bad translations received a high ME score and are thus boosted erroneously. A typical example for such a case would be a phrase pair where the context provides clear indicators and thus allows for a confident prediction of an (appropriate) preposition, but the phrase pair itself is flawed, for example by missing a content word on the target side that was lost due to faulty word alignment.

Score variant 3 This variant aims at providing a new probability distribution: The placeholder translation probability is considered as a translation probability of the phrase containing *some* preposition, and is used as the basis to estimate a score for the phrase to contain the *predicted* preposition, through combination with the ME score. It is important to keep in mind that the prediction score does not provide the probability of the target phrase representing a translation of the source phrase, but only indicates how well the predicted preposition fits into the target phrase in the given context – this means that a bad translation option still can have high ME scores. To avoid that bad translation options are erroneously boosted, the prediction scores are "dampened" with the lexical probability as an indicator for the general translation quality of the phrase pair, resulting in the formula below

$$P_{Prep}(e|f) = P_{PlaceHolder}(e|f) * (ME + lex_{Prep}(e|f))$$

where ME is the prediction score, $P_{PlaceHolder}$ is the translation probability calculated on the placeholder representation, and lex_{Prep} is the lexical probability computed for the phrase containing the generated preposition.

		Target phrase	$p(e\|f)$
Prep-Informed		`für[Acc]Waffe<Fem><Pl>[NN]`	0.333
		`nach[Dat] Waffe<Fem><Pl>[NN]`	0.148
		`für[Acc] nuklear<Pos>[ADJA]`	0.037
		`für[Acc] militärisch<Pos>[ADJA]`	0.037
		`für[Acc] die<+ART>[ART]`	0.037
Synthetic Phrases	**sentence 1**	`nach[Dat] Waffe<Fem><Pl>[NN]`	0.192 ✓
		`empty[Acc] Waffe<Fem><Pl>[NN]`	0.131 ✓
		`empty[Nom] Waffe<Fem><Pl>[NN]`	0.121
		`für[Acc] Waffe<Fem><Pl>[NN]`	0.094
		`von[Dat] Waffe<Fem><Pl>[NN]`	0.038
	sentence 2	`für[Acc] Waffe<Fem><Pl>[NN]`	0.336 ✓
		`empty[Nom] Waffe<Fem><Pl>[NN]`	0.101
		`von[Dat] Waffe<Fem><Pl>[NN]`	0.045
		`nach[Dat] Waffe<Fem><Pl>[NN]`	0.041
		`die<+ART>[ART] Waffe<Fem><Pl>[NN]`	0.037

TABLE 7.2: The top-5 synthetic phrases according to $p(e|f)$ for the phrase *for weapons* based on the predictions from table 7.1. Phrases marked with ✓are correct in the respective context.

In an additional variant (3b), the resulting translation options are normalized such that they sum to 1 with the non-synthesized entries whose probability mass remains unchanged and is equivalent to that in the preposition-informed system. The normalization aims at obtaining a "true" probability distribution with context-dependent scores for phrases containing predicted prepositions that is as close as possible to the preposition-informed system. In particular, it is important to keep the balance between original and synthesized phrases: by normalizing the probability estimations of the synthesized phrases, their overall probability mass corresponds to that in the preposition-informed system, but is distributed differently to represent preferences on token-level.

Example and use in phrase-table There is one main difference between score variants 1 and 2, and score variant 3: while in the first two variants, the ME-based scores are used as additional features to the lexical weights and the placeholder translation probabilities, variant 3 estimates a new phrase translation probability distribution based on the placeholder probabilities and the prediction scores to replace the placeholder probabilities.

The examples in table 7.2 contrast the translation probabilities of the translation options in the preposition-informed system (top part) and the generated phrases and their translation probabilities $p(e|f)$ estimated according to score variant (3b) for the predictions from table 7.1. Suitable translations for the respective context are marked with ✓. For sentence 1 (context *"look for weapons"*), translations containing `nach-Dat` and `EMPTY-Acc` are top-ranked. In comparison to the original distribution from the preposition-informed system, the previously second-ranked option (`nach-Dat`) moved up to the first position, and the option `EMPTY-Acc` constitutes a newly created phrase. While the previously top-ranked option with `für-Acc` is still among the top-5, it now has a comparatively low translation probability. For sentence 2 (context *order for weapons*), the previously top-ranked phrase containing `für-Acc` is still at the top position, and has an even higher translation probability than in the preposition-informed system.

The generated entries in table 7.2 also illustrate another, general effect on the phrase-table containing synthetic phrases: the two top-tanked entries from the preposition-informed system are now expanded and its "descendants" take up the top-5 positions for sentence 1, and the top-4 positions for sentence 2. As a result, the lexically less suiting options from the original positions 3-5 in the preposition-informed distribution are disfavoured by being ranked to considerably lower positions. The advantages and disadvantages of this effect are discussed in section 7.5.2.

Reordering-table To create a reordering-table, the statistics from the placeholder representation are used. Predicting and generating prepositions, even when modifying the complement type, is unlikely to influence the order of a phrase; thus, the statistics calculated on the placeholder representation should be adequate for the synthetic phrases.

7.4 Experiments and results

In this section, the results of the preposition-informed system are compared to systems enriched with synthetic phrases relying on the three presented score variants.

7.4.1 Experimental setup

All systems are built with the Moses phrase-based framework. The parallel English–German training data consists of 4.592.139 parallel sentences aligned with GIZA++. A

5-gram target-side language model is built on 45 million German sentences (News'14 corpus + parallel data). The system is tuned and tested on news data, using the WMT'13 set (3000 sentences) as tuning data, and WMT'14 (3003 sentences) as test data. The four morphological features number, gender, case and strong/weak to generate inflected forms are predicted with four CRFs trained on the target-side part of the parallel data. This setting corresponds to the experimental setup from section 3.4.3.

In contrast to the morphological features, which are modeled by sequence models predicting the features as a sequence of labels (i.e. case/number/etc. of consecutive words in an NP/PP), the prediction model for prepositions is trained as a maximum entropy model on the parallel data. In this setting, each preposition is predicted independently, without knowledge of prior predictions. All prediction models are trained using the Wapiti toolkit (Lavergne et al., 2010).

To obtain the stemmed representation, the German data is parsed with BitPar (Schmid, 2004) and analyzed with SMOR (Schmid et al., 2004), which is also used to generate inflected forms. For the extraction of the English features, the source-side is dependency-parsed with (Choi et al., 2012).

7.4.2 Baselines

For the following discussion, two baselines are defined:

Baseline-1 is a standard phrase-based translation system operating on surface forms without any form of morpological modeling.

Baseline-2 is the inflection prediction system as introduced in chapter 3. There is no modeling of prepositions in this system, except for the splitting of portmanteau prepositions into preposition and article prior to translation, and re-merging in the post-processing step.

7.4.3 Results

Table 7.3 shows the results for the preposition-informed system in comparison to the baselines. As described in section 7.2, the preposition-informed system contains overt "regular" prepositions and additional pseudo-prepositions `EMPTY-Case` at the beginning of a noun phrase, thus transforming it temporarily into a prepositional phrase. After translation, empty prepositions are simply deleted before generating inflected forms.

System		BLEU
baseline-1	Surface forms	19.17
baseline-2	Stemmed	19.35
prep-informed system (P-1)	Stemmed + ∅-CASE	19.76
prep-informed system (P-2)	Stemmed + ∅-CASE-top-20	19.73

TABLE 7.3: Scores for baselines and preposition-informed system.

System	Features used for MERT tuning	BLEU
SP-1	SCORE-VARIANT-1	19.76
SP-2	SCORE-VARIANT-2	19.83
SP-3a	SCORE-VARIANT-3	19.80
SP-3b	SCORE-VARIANT-3, norm. $P_{prep}(e\|f)$	19.86*

TABLE 7.4: Variants of the synthetic-phrases system. * marks significant improvement over system P-2 (with pair-wise bootstrap resampling with sample size 1,000 and a p-value of 0.05)

The introduction of empty prepositions on both sides of the training data leads to statistically significant improvements in BLEU over both the surface system (Baseline-1) and the standard inflection prediction system (Baseline-2).

As the generation of synthetic phrases on the full phrase-table is not manageable, the phrase-table is reduced in size by keeping only using the top-20 translation entries according to $p(e|f)$. The preposition-informed system operating on this reduced phrase-table (P-2) yields essentially the same result, confirming that the restriction to the most probable translation options does not hurt performance. In the following, systems with synthetic phrases are compared to the reduced system P-2, as this relies on the same phrase inventory.

Table 7.4 shows the results for system variants with synthetic phrases, which all outperform Baseline-2, the standard inflection prediction system. All systems using synthetic phrases score in a similar range, and lie slightly over the results of the preposition-informed systems. Even though the difference is small, the best system, SP-3b, is significantly better than system P2, the preposition-informed system reduced to the top-20 phrase-table entries. It is, however, not significantly better than system P-1, which has access to all phrase-table entries. This is reasonable considering that SP-3b relies on the same phrase-inventory as system P-2.

The system with the lowest score (SP-1) uses lexical weights and placeholder translation probabilities in combination with the ME prediction score and an indicator feature to account for the number of prepositions per phrase. System P-2, extended with the product of the phrase translation probability and the ME score, yields a slightly better result. Finally, system SP-3 obtains the best result: here, new phrase translation probabilities replace the placeholder probabilities. Normalizing the scores between all phrases containing generated prepositions such that all translation options sum to 1 leads to the overall best result.

Coming back to example (2), the method of generating synthetic phrases leads to correct translations in the synthetic-phrases systems by promoting the correct translation of *of* in the context of *warn ... of*, which previously was inaccessible. For the phrase pair *of the* → `PREP die<+ART><Def>[ARTdef]`, the top prediction for the placeholder is *vor-Dat*, which leads to a considerably higher probability for using this phrase pair in comparison to the preposition-informed system, where it was available in theory, but very unlikely to be used. For example in system SP-3b, the used phrase-pair *of the* → `vor[APPR-vor-Dat] die<+ART><Def>[ARTdef]` is ranked second (after `EMPTY-Gen die<+ART><Def>[ARTdef]`, due to its very high original base translation probability).

(3) EN many critics of veganism warn in particular **of** the lack of vitamin b12

P-1 viele Kritiker des Veganismus warnen insbesondere der Mangel an Vitamin B12
many critics of veganism warn in-particular the lack of vitamin b12

SP viele Kritiker des Veganismus warnen insbesondere **vor** dem Mangel an Vitamin
many critics of veganism warn in-particular of the lack of vitamin
B12 .
b12

Example (3) contrasts the output of the preposition-informed system with the output obtained from the synthetic phrases system, containing the correct preposition *vor*.

7.5 Discussion

This section focuses on obtaining more insight into the performance of the synthetic phrases systems, particularly with regard to the use of newly generated phrases. Furthermore, two potential side-effects on the phrase-table caused by the addition of synthetic phrases are analyzed: the first concerns the accidental boosting of lexically incorrect translation options by too high prediction scores in certain scenarios; the second concerns the resulting distribution in the phrase-table after expanding one original

	SP-1	SP-2	SP-3a	SP-3b
new	1489	1507	1391	1398
regular	38132	34541	35101	33571

TABLE 7.5: Number of newly generated and regular phrase-table entries used to translate the test set (3003 sentences).

phrase-table entry into 5 descendants representing the top-5 predictions. The section concludes with a small-scale manual evaluation.

7.5.1 Use of newly generated phrases

Motivated by the issue of insufficient coverage of prepositions in non-default translation contexts, as illustrated by example (2), an important aspect in the presented method is its ability to generate new phrases. Table 7.5 shows the amount of newly created and original phrases used to translate the test-set. For 3003 sentence, close to 1500 new phrases have been applied; this corresponds to about one new phrase per two sentences on average. Considering that closed-class function words such as prepositions usually are thought to be well-covered in NLP training data, this number is not insubstantial.

The translations in (4) illustrate how the use of a newly generated translation pair closes a gap in coverage and thus can improve the translation. The translations obtained with the preposition-informed system P2 and the synthetic phrases-system SP-2 are identical with the exception of the wrong preposition in the output of the preposition-informed system. While a combination of *hoffen + zu* ('hope + to') is not incorrect per-se, it is used to adjoin a verbal complement such as *hofft zu verbessern* ('hopes to improve'). Preceding a nominal phrase, *zu* is incorrect in this context.

(4) EN nullprp the deutsche bahn hopes to improve nullprp the kinzigtal railway line in the coming year.

P2 die deutsche Bahn hofft **zur** Verbesserung der kinzigtal Eisenbahnlinie im
the german bahn hopes to improvement of-the kinzigtal railway-line in-the
kommenden Jahr .
coming year .

SP-2 die deutsche Bahn hofft **auf** eine Verbesserung der kinzigtal Eisenbahnlinie im
the german bahn hopes for an improvement of-the kinzigtal railway-line in-the
kommenden Jahr .
coming year .

Between the English sentence and the two proposed translations, there is a structural shift from the verbal phrase *to improve* on the source side, and the realization as a noun (*Verbesserung*) on the target side.[4] Furthermore, due to the segmentation in the decoding process, the English verb *hope* is translated in another phrase, which excludes a translation within one phrase such as *hope to* → *hoffen auf*. The used phrase pair in the synthetic-phrases system is based on the phrase *to improve nullprp the* → `PREP`$_1$ `eine Verbesserung PREP`$_2$ `der`, where `PREP`$_1$ is realized as `auf-Dat`; and the realization of `PREP`$_2$ as `EMPTY-Gen` is trivial. To translate the sentence in (4), these new translation options have been used:

the deutsche bahn	→	die ∅-Nom deutsche Bahn
to improve ∅ the	→	**auf-Acc** eine Verbesserung ∅-Gen der
railway line in	→	Eisenbahnlinie in-Dat

Looking at the phrase-pair *to improve ∅ the → auf-Acc eine Verbesserung ∅-Gen der* out of context, it does not seem a particularly good or generally applicable phrase-table entry – it is, however, very useful for the sentence in question. The example illustrates how an underspecified preposition in a phrase can be adapted as needed by the context. For a different subcategorization frame, such as, for example *plans [to improve nullprp the] ...*, the German phrase can be accordingly realized as *plant [∅-Acc eine Verbesserung ∅-Gen der]* ('plans [∅ the improvement of-the]').

7.5.2 Side-effects on the phrase-table

When examining phrase-table entries and produced translations, it becomes evident that a recurring problem are lexically incorrect translations that are unreasonably boosted due to high ME prediction scores in comparison to lexically sounder options. In particular, this happens when a generally infrequent word occurs in a lexically incorrect translation option: in such a scenario, the prediction model is often overly confident, reflected by comparatively high prediction scores derived from insufficient – and seemingly clear – training data. Other, lexically more correct translation options with original translation scores in a similar range as the incorrect option, but with lower prediction scores, are then "overpowered".

[4] It can be assumed that the structurally equivalent translation as *hofft ... zu verbessern* is dis-preferred by the translation system as this would require the positioning of the verb at the end of the clause.

Consider as an example the English phrase *for bags* and two possible translation options with similar translation and lexical probabilities: PREP *Taschen* (general translation: 'bags') and PREP *Müllsäcke* (specific translation: 'garbage bags'). There are only few instances of PREP *Müllsäcke* in the training data for the preposition prediction model, and when applied in the prediction task, the model reproduces the few observed training instances with a score around 0.9 for the top-ranked preposition. In contrast, the predictions for PREP *Taschen* are less spiked, with a score around 0.55 for the top-ranked predictions. As a result, the translation option containing *Müllsäcke*, despite being lexically inappropriate, is boosted by the prediction scores and is consequently more likely to be chosen during translation.

Lexical features, for example verb-noun tuples, are important for training the prediction system as they often carry important information about selectional preferences. However, the example above illustrates how infrequent words can lead to overfitting issues, and thus be more harmful than useful. In the previous experiments, this problem is addressed by weighting down the influence of the prediction score by combining it with the lexical and/or phrase translation probabilities.

An additional experiment aims at reducing the negative influence of rare nouns by replacing them with dummy nouns in the ME training data – this allows to still benefit from lexical information while excluding insufficiently represented nouns. The first line in table 7.6 shows the results obtained with prediction models trained on data where infrequent nouns in the NP/PP ($freq \leq 25$ for features 5 and 6 in table 7.1) are removed from the training data. Comparing to the results in table 7.4, the general outcome is similar in that systems SP-2 and SP-3b are slightly better; the result of system 3b is the overall best result. This suggests that a careful representation of lexical items in the training data, in particular the removal of infrequent nouns, can lead to improved prediction and translation results.

Another potential problem lies in the fact that there is a certain mismatch between the phrase-translation pairs provided to the system (reduced to the top-20 entries per source phrase) and the training data for the prediction model (all source/target/alignment triples extracted from the parallel data). In a further experimental variant, the training examples for the prediction model are restricted to those also occurring in the top-20 filtered phrase table, with the objective to reduce the training data to relevant training instances only. The second line in table 7.6 shows the results for this variant of the training data; however, the results are generally worse than those of the previous system. A possible explanation could be that removing a subset of the training instances leads

	SP-1	SP-2	SP-3a	SP-3b
(1) no infreq nouns	19.59	19.85	19.71	19.94*
(2) reduced data	19.82	19.58	19.73	19.64

TABLE 7.6: Results when filtering out infrequent nouns in the ME training data (1) or reducing the amount of source-target-alignment triples used for ME training (2). * marks significant improvement over system P-2.

to a somewhat unbalanced training set. Furthermore, even though the removed phrases themselves are not immediately relevant due to not occurring in the phrase-table, the training data is still decreased in size as much relevant information comes from outside of the phrase: for example, a training phrase that is lexically irrelevant can still contribute to the prediction model by providing information about subcategorization preferences of a (relevant) source verb observed *outside* of the actual phrase context.

7.5.3 Distribution in the phrase-table

Another, potentially negative effect on the phrase-table stems from integrating translation options containing the n-best predictions per placeholder preposition: if a translation option is already dominant, it can be further enhanced – not only by receiving a high score derived from the translation probability and the prediction score, but also by being represented by several "descendants" of one original phrase. As a result, equally (or even more) valid translation options with only slightly lower translation and/or prediction scores can then be dispreferred not only on the level of scores, but also by the amount of competing phrase-table entries stemming from the original top-ranked translation option.

Consider, for example, the phrase *expand nullprp their*: the best lexically correct translation option, *erweitern* EMPTY-Acc in the preposition-informed system, is ranked third according to $p(e|f)$, with two meaningless translation options (only determiner or only preposition) being ranked first and second. This is already a bad starting point, and when adding synthetic phrases, the meaningless top-ranked translation options are further expanded, and their descendants fill the positions 1-5 in the resulting phrase-table. This makes it even more difficult for the translation system to select a lexically appropriate translation option in such a scenario.

On the other hand, this effect can also promote lexically correct translation options – this is the case when a top-ranked translation is closely followed by a less suited, rivaling translation that has a practical chance to be selected by a baseline system.

	prep-informed		**synth-phrases**	
	missing	wrong	missing	wrong
verbs	32	11	23	10
nouns	2	15	2	17
prepositions	6	6	3	8
gram. case	–	4	–	3

TABLE 7.7: Manual error analysis of 50 randomly selected sentences.

This effect can be seen in table 7.2 where lexically incorrect phrases (such as `für-Acc nuklear<Pos>[ADJA]` ('for nuclear'[5]) as translation for *for weapons*) are moved to a lower position, to the benefit of variants of the lexically sound translation option. While this side-effect can indeed be positive, it may also happen that literal translations are preferred over less common senses in case of word-sense ambiguities. The manual evaluation presented in the next section comes to the result that slightly more verbs are translated with the synthetic-phrases system in comparison to the preposition-informed system. As the omission of verbs is a general problem in English–German translation, the fact that there is a slight improvement with regard to the amount of translated verbs can be seen as an indicator that the side-effect of enhancing lexically sound translations leans to be slightly positive on average.

7.5.4 Manual evaluation

A subset from the output of the overall best system (SP-3b in table 7.6 was manually compared to the output from the preposition-informed system P-2. For a set of 50 randomly chosen sentences (length 10-20 words), 2 native speakers annotated errors concerning missing/incorrect verbs, nouns and prepositions, as well as incorrect grammatical case. This type of error analysis loosely follows the concept employed in Williams et al. (2015); and aims at assessing the quality of prepositions and case, while also covering some relevant lexical aspects that might be influenced from the addition of synthetic phrases.

The result of the evaluation is listed in table 7.7: the number of errors for the categories *preposition* and *grammatical case* are similar for both systems. However, a slight improvement was found with regard to the number of translated verbs, which are known to be generally difficult for English–German translation due to the structural differences between these two languages.

[5] "Nucular – it's pronounced nucular!" (Homer Simpson)

It is important to keep in mind that there are many other factors that are not considered in this evaluation, such as, for example, the overall structure of a sentence. Furthermore, it is generally difficult to evaluate verbs in combination with their subcategorized arguments if there are several valid possibilities to realize and arrange the arguments, and/or when the sentence is only partially translated. Such a problem of unclear error categories is illustrated by the following example:

(5) EN this is mainly due to the higher contribution from the administrative budget

P2 das ist hauptsächlich auf die höheren Beiträge aus dem Verwaltungshaushalt
this is mainly to the higher contributions from the administrative-budget

SP3b das ist vor allem wegen den höheren Beiträgen aus dem Verwaltungshaushalt
this is mainly because-of the higher contributions from the administrative-budget

The translations P2 and SP-3b in (5) are nearly identical, but differ with respect to the preposition used to translate *due to*, which also influences the grammatical case in the German translation. In both translations, the verb is not fully translated, as the German *ist* plays rather the role of an auxiliary. The synthetic-phrases system SP3b preserves the structure in the English sentence by translating *due to* into *wegen-Dat* (*wegen-Gen* is also correct, probably even slightly preferred over dative case). The preposition *auf-Acc* and the accordingly adjusted case in the translation produced by system P-2 can also be part of a valid translation, but only in combination with the verb *zurückführen auf* ('attribute to'). For such cases, it is thus difficult to decide which error category to annotate (*missing verb* or *wrong preposition*).

7.6 Related work

The method of using synthetic phrases to model complement types combines three research areas: integrating rich source-side information, previous approaches to model case and prepositions, and the generation of new phrase-table entries.

7.6.1 Integrating source-side information

The integration of source-side features into SMT is a widely studied problem, but often with an objective for improvement at the lexical level, i.e. with regard to word sense disambiguation and lexical choice, for example Carpuat et al. (2007), Gimpel et al. (2008), Jeong et al. (2010) or Tamchyna et al. (2014), and less with a focus on the modeling of syntactic-semantic aspects or the synthesis of new phrases. Section 4.5 provides more details about the integration of source-side features with regard to modeling case.

7.6.2 Prepositions in SMT

The fact that prepositions are difficult to translate has been shown in many evaluations of machine translation. For example, Williams et al. (2015) present an error analysis of their shared task submission, listing the number of missing and wrong content and function words. In their English–German translation system, the combined number of *missing/wrong/added prepositions* is among the most observed error types. Similarly, Popović et al. (2015), who focus on translating into morphologically rich languages, find that mistranslations of prepositions are a recurring error type. Sections 5.6 and 6.6 already provide an overview of some previous approaches modeling prepositions, often in rule-based systems or in a simulated translation scenario.

In this chapter, the idea of placeholder prepositions that can be transformed into overt prepositions to form a PP or into empty prepositions at the beginning of noun phrases is a key concept. Modeling prepositions, including empty prepositions, based on rich linguistic features has been proposed by Agirre et al. (2009) who successfully apply this strategy to a rule-based system. Their approach is extended by Shilon et al. (2012) by adding a statistical re-ranking component, cf. section 6.6.

A related concept has also been applied to Finnish, which is a highly inflective language with a very complex case and preposition system. Tiedemann et al. (2015) experiment with "pseudo-tokens" in a Finnish–English translation system to account for the fact that Finnish morphological (case) markers often correspond to a separate English word, typically a preposition. Due to the complexity of Finnish, they consider only a subset of the morphological markers. They report mixed results for their approach, and a manual evaluation remains inconclusive about the effectiveness of their method. The main difference to the approach presented in this chapter is the translation direction *into* the morphologically complex language, which requires morphological modeling on the target-side. Furthermore, Tiedemann et al. (2015) only use pseudo-tokens on the source side, whereas the preposition-informed approach in this chapter adapts both source and target side to obtain more isomorphic parallel data.

7.6.3 Synthetic phrases

The use of synthetic phrases has already been mentioned in this thesis, for example in section 3.5 and section 6.6, and is contrasted here to the approach of synthesizing function words presented in this chapter.

Chahuneau et al. (2013) propose the use of synthetic phrases to augment the rule inventory when translating into morphologically rich languages. They use a discriminatory model relying on source-side features (dependency information and word clusters) to predict target-side inflected words that are used to generate new phrase-table entries. They report improved translation quality for several language pairs. While their approach aims at providing better coverage for inflectional variants of (presumably) content words, the strategy in this chapter aims at the generation of function words to obtain the most appropriate complement type realization given the source sentence. Being closed-class words, prepositions are not under-represented in the same way as inflectional variants of low-frequency content words, but their "meaning" or subcategorized use in a particular setting might be not well-attested in the training data, resulting in a coverage problem at phrase-level. The generated synthetic phrases to model complement types include *sequences* not observed in the training data by adding or changing prepositions for a different PP, or by removing prepositions to form an NP.

A related task to synthesizing prepositions is the generation of determiners when translating from a language that has no explicit definiteness morpheme. Tsvetkov et al. (2013) create synthetic phrases with predicted determiners to augment a phrase-table in a Russian–English translation system. They use a classifier trained on local contextual features to predict whether to add or to remove a determiner from the target-side of a translation rule. In contrast to determiners, which can be modeled with local context, prepositions are function words that also have a semantic content and that depend on complex interactions between the verb and other subcategorized elements throughout the sentence.

Huck et al. (2017) generate unseen inflection variants in English–Czech translation using a discriminatory model that combines rich source-side information with target-side information derived from the (partially) translated input sentence.

7.7 Summary

This chapter presented two approaches to model complement types in English–German SMT to balance structural difference between source and target language, and to take into account token-level requirements to find the correct complement realization:

- **Preposition-informed system:** The introduction of pseudo-prepositions to transform every nominal phrase into a prepositional phrase, with the objective to adjust structural differences at the level of complement types between the source and target side, leads to an improvement in translation quality.
- **Synthetic-phrases system:** Extending the preposition-informed system with synthetic phrases conditioned on the source side on token-level tends to perform slightly better than the preposition-informed system

The differences between the preposition-informed systems and the synthetic-phrases systems are rather small and only apply to some system pairs, which makes it difficult to draw a clear conclusion on the effectiveness of generating synthetic phrases to improve the realization of complement types. However, the examples illustrated clearly that there can be, somewhat unexpected for closed-class words, coverage gaps for prepositions. The evaluation showed that newly generated phrases are indeed used by the system. This outcome can be considered as a confirmation that the generation of synthetic phrases to handle subcategorization, be it with the method proposed here or by employing a different technique, is a sound and useful approach.

A somewhat problematic aspect in the presented approach is the fact that the prediction step only predicts how well a preposition fits in a given target phrase based on relevant context in the source phrase. An important part of integrating synthetic phrases thus consists in combining two sets of scores, namely the *lexical/translation probabilities* and the *prediction scores* to take into account both the general quality of a phrase as translation of the source side and the prediction confidence. The different score variants explored in the previous sections aim to find a combination that considers all relevant factors, but the results show that it is a difficult task to account for all possible interactions.

An idea for future work, that is however beyond the scope of this thesis, is the exploration of models that predict the complete *target phrase* given relevant context, instead of predicting only the *preposition* in the target-phrase.

Chapter 8

Addressing Problems across Linguistic Levels in SMT: Modeling Morphology, Syntax and Lexical Choice

In the previous chapters, the main research interest was target-side inflection, with a focus on modeling subcategorization and preposition choice by making use of rich source and target-side features. However, inflection is not the only challenge one faces in a translation task. The typical main error types in machine translation can be attributed to (i) differences at the syntactical level; (ii) problems at lexical choice; and (iii) morphological complexity. The first part of the thesis mainly focused on target-side morphology, in combination with adjusting local syntactic differences, while not much attention was paid to phenomena concerning the lexical level. Generally, there is much research on each of these topics individually, but so far, not much is known about what to expect when combining approaches aiming at different error types, both in terms of being complementary and possible interactions and side-effects.

This chapter explores the combination of established approaches: source-side reordering to overcome global structural differences at the syntactic level (pre-processing), a discriminative model integrated into the translation system to provide a better basis for lexical decisions (at decoding-time), and the inflection prediction approach to handle target-side morphology (combined pre/post-processing). The main findings of this experiment are that the three approaches are indeed complementary, but also that source-side pre-ordering has negative effects on the lexical level, which can be compensated for by the use of the discriminative model. The settings and experiments presented in this chapter are published in Weller-Di Marco et al. (2017).

8.1 Motivation

Many of the errors occurring in machine translation can be attributed to problems on three linguistic levels: structural difference between the source and target side, morphological complexity and lexical choice. These categories are often intertwined, as linguistic phenomena that are expressed syntactically in one language can be expressed on a morphological level in another language. For example, the syntactic function of an argument can be expressed morphologically through grammatical case, (e.g. in German or Slavic languages), or on the syntactic level by means of a fixed word order (such as SVO in English or French).

This chapter focuses on the question of how to combine approaches to address all three linguistic levels. Previous work proposes many strategies to address these problems, but the proposed methods are only studied individually. The experiments presented in this chapter explore system variants that combine structural adaptation between source and target side, a discriminative model to take into account richer context features during translation, and a component to model target-side nominal morphology. The strategies are integrated into an English–German phrase-based system, with a particular focus on the following two research questions:

- Do the individual gains of one strategy add up when being combined with other methods?
- Are there interactions between the linguistic levels and potential side-effects caused by some strategies?

The experiments show that the different strategies aiming at different linguistic levels are complementary. The second question particularly concerns the effect of structural adaptation, which is applied in form of pre-ordering the source-side. While this strategy generally leads to improved translation results, it also entails the loss of relevant predicate-argument structures. The presented experiments indicate that the use of explicit context information, here in form of a discriminative classifier, can re-introduce the lost context information.

8.1.1 Overview of individual strategies

The following section gives an overview of some main strategies to address the different linguistic levels individually, and explains how the methods employed in this chapter relate to them.

Syntactic level

Structural differences between the source and target language are generally problematic for statistical machine translation, in particular when they involve reordering of constituents over large distances. First, the alignment of constituents separated by long distances is difficult, and the alignment process often fails to capture such equivalences, leading to problems already at the training stage of the translation system. Second, the reordering of phrases over long distances is typically costly for phrase-based decoding algorithms, thus also creating a problem during translation. Hierarchical systems (Chiang, 2005) can bridge distances up to a certain length by allowing for "gaps" in the translation rules. Strategies to further model long distance phenomena have been proposed, for example, by Braune et al. (2012).

An alternative method, especially for phrase-based systems, consists in structurally adapting the source language: in a pre-processing step, the source-side data is arranged such that it corresponds to the expected target-side structure. Such a pre-processing step improves the alignment quality and avoids long distance reordering at translation time. Variants of this method have been found to improve translation quality, for example Collins et al. (2005) for German–English and Gojun et al. (2012) for English–German translation.

For the language pair English–German, structural differences mainly concern the position of verbs in the sentence: while English adheres to an SVO-ordering with verbs typically occurring at the beginning of a clause, German has several clause ordering types that require verbs to be positioned at the end of clause in some settings. To handle the differences at the syntactic level, the strategy of pre-ordering verbs on the source side is employed.

Morphological level

The correct inflection of the translation output is one of the main problems when translating into a morphologically rich language. There two aspects to consider: first, morphological complexity is tied to data sparsity issues, as inflected forms need to be observed in the parallel training data in order to be generated by a translation system. Furthermore, inflection is subject to language-specific restrictions, such as local agreement in nominal phrases, but also sentence-level interactions, such as subject-verb agreement, or the assignment of syntactic functions through grammatical case.

There are many strategies to model target-side morphology in SMT, for example by means of computing inflectional features in combination with the generation of inflected forms, as in Toutanova et al. (2008), Fraser et al. (2012) and Burlot et al. (2016). An alternative strategy supplements the translation system with synthetic phrases to provide the full set of word inflections, which was first presented by Chahuneau et al. (2013). Huck et al. (2017) present an extension of this variant that additionally takes into account context information from source and target side. Finally, consistent inflection can be enforced by using agreement restrictions in a hierarchical system, as proposed by Williams et al. (2011). Section 3.5 provides a more detailed discussion on modeling morphology in SMT.

In this chapter, the inflection prediction system that first translates into a stemmed representation with a component for the generation of inflected forms is used, following the idea in Fraser et al. (2012) as described in chapter 3.

Lexical level

Translation problems on the lexical level are diverse and include aspects such as word-sense disambiguation, selectional preferences and the translation of various types of multi-word structures. Many methods studied in previous work rely on rich source-side features to provide more context during decoding, for example Carpuat et al. (2007); Jeong et al. (2010); Tamchyna et al. (2014) and Tamchyna et al. (2016).

Another strand of work focuses on improving the translation of multi-word items which often cannot be translated literally. One popular approach consists in merging multi-word expressions into “super-tokens” such that they can be translated as a single unit and thus preserve their meaning, for example Carpuat et al. (2010) for English–Arabic and Cholakov et al. (2014) for English–Bulgarian translation. Since the approach of merging is problematic for non-contiguous multi-word expressions, Cap et al. (2015b) propose to mark the semantically void verbs in support verb constructions to distinguish between literal translations and translations in an idiomatic context. Further approaches to handle multi-word expressions consist in handling multi-word expressions at the phrase-table level, for example with boolean variables to mark support verb status, for example Carpuat et al. (2010).

In this chapter, a discriminative model using rich source-side features to score translation rules aims at providing more context to allow for better informed lexical decisions. In a variant of the experiment, features to explicitly model support verb constructions and verbal inflection are added to the discriminative model.

8.1.2 Combining approaches

To address the three different linguistic levels in one system, individually established and well-researched methods are employed, resulting in a an English–German translation system that combines the strategies of *source-side reordering* (pre-processing), a *discriminative classifier* (at decoding time) and *target-side inflection generation* (combined pre/post-processing). Thus, the translation system incorporating all three strategies is built on reordered English data and an underspecified German representation as basis for the post-processing inflection prediction, with a discriminative classifier added as feature function into the decoder.

While the individual methods are established and usually improve translation results when applied on their own, it is not clear (i) whether individual gains add up when being combined with other approaches and (ii) how targeting one linguistic level impacts other levels. While the first question is straightforward, the second question is more complex and mainly concerns the effects of source-side reordering, namely whether introducing German clause ordering in the English data causes negative side-effects: while verbs and their arguments typically occur close to each other in "regular" English, they can be separated by large gaps.

Source-side reordering improves the translation quality, but separating the verb from its arguments also entails negative consequences. First, Ramm et al. (2016) point out that the number agreement between verb and subject is impaired due to introducing a gap between verb and subject. Second, there can be a negative effect on the lexical level, for example when translating multi-word expressions. Consider, for example, the phrase *cut interest rates* which can be considered non-compositional, as the verb *cut* does not contribute its literal meaning, but rather a figurative sense. If the parts occur close to each other, there is enough context to translate *cut* according to its metaphoric sense into *senken* ('to decrease'). However, with too large a gap between *cut* and *interest rates*, it becomes difficult to disambiguate the use of *cut* into literal and metaphoric.

8.2 Morpho-syntactic modeling

The components handling the morpho-syntactic level are implemented as pre- and post-processing steps in the system setup. The pre-processing consists in applying source-side reordering to the English data and the preparation of the stemmed representation of the German data for modeling target-side inflection in a post-processing step (cf.

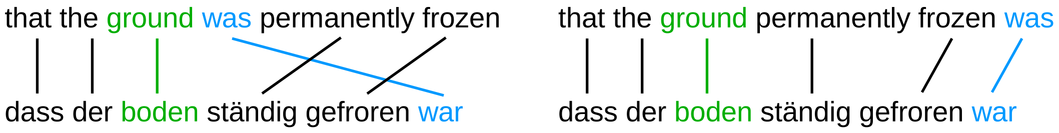
that the ground was permanently frozen
dass der boden ständig gefroren war
that the ground permanently frozen was
dass der boden ständig gefroren war

FIGURE 8.1: Verbal reordering in the training data: moving the verb to the *verb-final position*, i.e. towards the end of the clause/sentence.

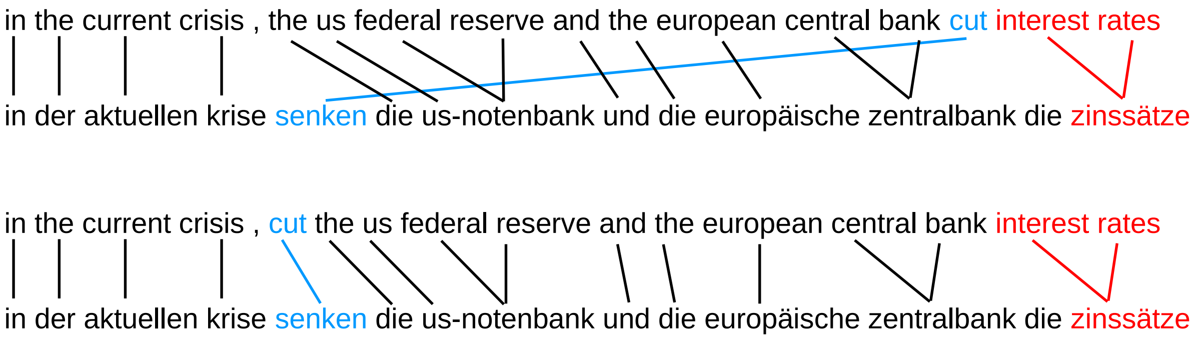
in the current crisis , the us federal reserve and the european central bank cut interest rates
in der aktuellen krise senken die us-notenbank und die europäische zentralbank die zinssätze
in the current crisis , cut the us federal reserve and the european central bank interest rates
in der aktuellen krise senken die us-notenbank und die europäische zentralbank die zinssätze

FIGURE 8.2: Verbal reordering in the training data: moving the verb to the *verb-second position*, i.e. the second constituent in the clause/sentence.

chapter 3 for more details on the stemmed representation). This system is then further enriched with the discriminative classifier described in section 8.3.

8.2.1 Source-side reordering

To adapt the syntactic structures between the source and target side, English verbs are moved to the expected German position, following the rules formulated in Gojun et al. (2012). The resulting structure is fundamentally different from "regular" English, as illustrated in figures 8.1 and 8.2. The sentence in figure 8.1 shows the movement of an English verb to the *verb-final position* in a subordinated clause, inserting a gap between subject and verb. This might well have a negative effect on subject-verb agreement: even though number is typically expressed on English nouns, English verbs are only marked for number (by the suffix *-s* in the 3rd person) in present tense – thus, verbs in past tense, as well as modal verbs, require context to determine number. The example illustrates how the context between the verb and its subject can be lost, in particular if the introduced gap spans over several words. The sentence in 8.2 depicts a movement to the *verb-second position*, where the finite verb is moved to the second constituent in the sentence. In the case of a compound tense, the finite verb (i.e. the auxiliary) is positioned in the verb-second position, whereas all involved further verbs (such as participles or modal verbs) are moved to the end of the sentence, thus separating the verbs of a complex tense structure. In contrast, English clause structure typically has the verbs of a compound tense occur close to one another.

Long-distance reorderings as illustrated in figures 8.1 and 8.2 are not uncommon, and the benefit of adapting the source-side structure is intuitively clear. However, the verbal pre-ordering comes at the price of separating verbs and its arguments. This is particularly problematic when the verb and its object form a non-compositional multi-word expression. Then, the entire expression (or at least parts thereof) cannot be translated literally, but the object must be known to correctly translate the verb and vice-versa. While the reordered representation provides a better basis for word alignment, in particular for capturing verbal translations, there is also less relevant context to distinguish between translation senses.

Furthermore, the separation of finite/non-finite verbs in compound tenses is problematic as German past tense auxiliaries depend on the verb: the past tense of most verbs is built with the auxiliary *haben* ('to have'), but the set of *motion verbs* (*gehen, kommen, ...*: 'to go, to come, ...') requires the auxiliary *sein* ('to be').

8.2.2 Morphological modeling

Nominal morphology is handled by a morphology-aware translation system that first translates into an underspecified representation, with a component to generate inflected forms in a post-processing step, cf. chapter 3. The stemmed representation is enriched with translation-relevant features such as number on nouns, to ensure that the number of a phrase as expressed on the source-side is preserved during translation. To re-inflect the underspecified SMT output, inflectional features are predicted in a sequence model, based on the values specified in the stem-markup as input. Then, inflected forms can be generated from the stem-feature pairs using the morphological resource SMOR.

8.3 Context features for lexical modeling

Rich context features provide valuable information on the lexical level, but can also contribute on the morpho-syntactic level by presenting information for number agreement or auxiliary choice. The context information is integrated by the use of a *discriminative classifier* (VowpalWabbit[1]) that is integrated into the Moses framework and scores translation rules using rich context information (Tamchyna et al., 2014).

8.3.1 Integration of the context features

The information used within the classifier mainly stems from the source side, where the entire sentence is available and can be parsed, whereas the target-side sentence is (i) not yet complete during translation, and (ii) likely not well-formed, making parsing or similar analyses difficult and inaccurate. The context from the target side is thus limited to the respective phrase.

8.3.2 Classifier variants

A "basic" classifier to be combined with the morpho-syntactic strategies is built on standard features such as POS-tags, lemmas and source-side dependencies, targeting mainly the lexical level. With regard to the potential problems introduced by the verbal pre-ordering on the source-side, this first classifier is extended with features aiming at providing information for the translation of *support verb constructions* and for better modeling of *number and tense* for verbal translation.

Standard features

On the source-side, the standard features comprise part-of-speech tags and lemmas within the phrase and a context window set to 5 words on each side for tags, and 3 words for words and lemmas. Information across larger distances is captured by dependency relations such as verb-object pairs or verb-subject pairs, as depicted in the 4th and 5th column in table 8.1. On the target-side, lemmas and part-of-speech tags for the current phrase are given.

Support verb constructions

Support verb constructions, also known as light-verb constructions, are multi-word expressions that are formed by a verb and a predicative noun, for example *make a contribution*. Typically, the verb does not contribute its full meaning, but is semantically

[1] https://github.com/JohnLangford/vowpal_wabbit/wiki

word	pos	lemma	associated verb/noun	relation	svc
cut	vvd	cut	rate	dobj	250
the	dt	the	–	–	–
us	np	us	reserve	nn-mod	–
federal	np	federal	reserve	nn-mod	–
reserve	np	reserve	cut	nsubj	–
...					
interest	nn	interest	rate	nn-mod	–
rates	nns	rate	cut	dobj	–

TABLE 8.1: Subject/object relations (columns 4 and 5) and support verb status annotated as *degree of association* (column 6) on the reordered sentence from figure 8.2.
.

light. Support verb constructions are problematic in machine translation, because the system often cannot distinguish between literal and idiomatic uses of a verb, where the latter often requires a non-literal translation.

As the examples in section 8.2.1 demonstrated, non-continuous multi-word expressions such as support verb constructions might be particularly impaired by the reordering. A variant of the classifier is thus extended with explicit information about support verb constructions. Cap et al. (2015b) propose a method to improve German–English phrase-based translation by annotating the *support verb status* on source-side verbs, thus disambiguating the verbs into "non-literal use" in the context of a support verb construction, and "literal use" otherwise. The set of support verb constructions to be annotated consists of highly associated verb-noun tuples, where the degree of association is measured by their log-likelihood. Cap et al. (2015b) opt for a hard annotation by adding markup to verbs occurring in a verb-noun paired scoring above a given threshold[2].

In the experiments in this chapter, an equivalent to a hard annotation as in Cap et al. (2015b) is compared to a second, more flexible annotation variant:

- The use of a *binary support verb feature* (yes/no) to mark the support verb status of a verb for a fixed set of verb-noun tuples, using a log-likelihood threshold of 1000[3]. There is no dependency information used in this variant of the classifier,

[2] Cap et al. (2015b) report results for different thresholds, but find that varying the thresholds has little to no influence on the results.

[3] This threshold lead to one of the best results in Cap et al. (2015b).

only the basic features lemma and part-of-speech tags, making the setup similar to the one employed in Cap et al. (2015b).

- The annotation of the *degree of relatedness* between verb and noun in form of the log-likelihood score, see rightmost column in table 8.1. In contrast to the first variant, this allows to annotate the spectrum of association (i.e. low association to high association) instead of arbitrarily deciding on a threshold.

Number and tense information

The complexity of verbal inflection is generally difficult to capture, in particular when complex interactions between several verbs in a compound tense are involved. As reordering leads to problems at the level of verb-translation (such as verb-subject agreement or choice of auxiliary), the basic discriminative classifier is extended with regard to the modeling of verbal features. It also has to be noted that the inflection prediction process (cf. 8.2.2) only covers nominal morphology, but no verbal morphology. While an attempt has been made to extend the inflection prediction setup with a component to generate inflected verb forms on the target-side based on an underspecified intermediate representation (Ramm et al., 2016), the authors come to the conclusion that this strategy is not optimal due to the comparatively complex nature of verbal inflection. Instead, Loáiciga et al. (2014) investigate rich source-side features, leading to an improvement for the translation of tense in an English–French factored MT system.

To compensate for problems introduced by the source-side reordering, the basic classifier is extended with features providing number and tense information; this has the additional benefit of supplementing the morphological level with a component aiming at verbal inflection, in addition to nominal inflection. Two types of verbal features are added to the classifier:

- *Number*, as derived from the subject, is used as an additional feature for finite verbs: while the number of a verb in present tense is often obvious (e.g. *goes* vs. *go*), the number of modal verbs (e.g. *can*), verbs in past tense (e.g. *went*) or progressive form (e.g. *going*) can only be derived through the context, namely the subject. While German does not have as rich a verbal morphology as for example many Romance languages, singular and plural usually require different forms for all tenses and verb classes.

- The status of *past* vs. *non-past* in combination with the related other verb(s) is annotated as extra information onto verbs. This serves the purpose of "connecting" verbs that are separated in the reordering process in order to help decide for the correct tense (*has ... gone, went*) and to select the correct auxiliary (*sein*: 'to be' vs. *haben*: 'to have') for German compound past tenses.

8.4 Experiments and results

This section presents the results of combining the strategies for the three linguistic levels and the extended variants of the discriminative classifier.

8.4.1 Data and resources

To build the SMT systems, a phrase-based translation model is learned from 4.592.139 parallel English–German word-aligned sentences. A 5-gram language model is built on 45 million German sentences. As development set, NewsTest'13 (3000 sentences) and as test set, NewsTest'14 (3003 sentences) are used. All training/test data was released as part of the WMT-2015 Translation Shared Task[4].

The linguistic pre-processing for the inflection prediction includes parsing the German data (Schmid, 2004) and morphological analysis to obtain the underspecified representation and to generate inflected forms (Schmid et al., 2004). To predict the features for nominal inflection, four CRF sequence models (one for each inflectional feature) are trained on the German part of the parallel data. The reordering rules from Gojun et al. (2012) are applied to the English data parsed with Charniak et al. (2005).

All systems are built using a version of Moses with the integrated discriminative classifier VowpalWabbit (Tamchyna et al., 2014)[5]. The training examples are extracted from the parallel data based on phrase-table entries. In order to keep the amount of training examples manageable, the phrase-table is reduced with *sigtest-filtering* with the recommended default setting *-l a+e -n 30*[6]. The results of all experiments are reported for sigtest-filtered phrase-tables, including the baseline systems and system without using the discriminative classifier. To train the classifier, 50 training iterations are run before applying early-stopping on the development set to identify the optimal model.

[4] `http://www.statmt.org/wmt15/`

[5] `https://github.com/moses-smt/mosesdecoder/tree/master/vw`

[6] `https://github.com/moses-smt/salm`

system	**basic**	**VW-1 pos/lem**	**VW-2 pos/lem/dep**
Surface	19.45	19.81*	19.90*
Surface V-Reordered	19.71*	20.24*	20.27*
MorphSys	19.81*	19.80*	19.93*
MorphSys V-Reordered	20.08*	20.51*	20.50*

TABLE 8.2: Results for morpho-syntactic and lexical strategies in case-insensitive BLEU. *: significantly better than Surface-basic (19.45)

8.4.2 Morpho-syntactic strategies with basic lexical model

To assess the influence of the different strategies, and in particular the impact of the lexical model combined with a reordered versus non-reordered system, we look at two dimensions: first, the combinations of the morphological and syntactic strategies are analyzed; then, the addition of the discriminative model at the lexical level is discussed.

The column "basic" in table 8.2 shows the results for combining strategies at the morpho-syntactic level: "Surface" refers to a baseline system trained on simple surface forms, "MorphSys" denotes the inflection prediction system, and "V-Reordered" refers to systems with reordered source-side data. Both inflection prediction and verbal reordering lead to an improvement on their own. Combining the two strategies adds up to a further improvement. In total, the system employing both morphological modeling and source-side reordering (20.08) gains 0.63 BLEU points in comparison to the surface baseline system (19.45).

The columns "VW-1" and "VW-2" show the effect of adding two variants of the basic discriminative classifier to the respective systems. Classifier VW1 uses only *word/lemma/pos* information, whereas VW-2 additionally used source-side dependency information to capture verb-argument relations. The difference between the two classifiers is small in every system setting. While the baseline system benefits from the added classifier, there is not much improvement to be observed for the "MorphSys" system. A reason might be that the classifier in the baseline surface system also contributes to a certain extent on the morphological level, such as triggering consistent inflection, which is already an integral part in the morphology-aware system. On the other hand, both systems built on reordered source-side data seem to benefit more from information at the lexical level. This outcome supports the initial hypothesis that verbal reordering, while generally improving the translation quality, comes at the price of introducing new

system	VW-2	VW-1 +threshold	VW-2 +degree
MorphSys	19.93	20.07	19.98
MorphSys V-Reordered	20.50	20.40	20.46

TABLE 8.3: Annotating support verb status.

problems. Combining all three strategies leads to the overall best results, a gain of 1.05 BLEU points over the surface baseline system.

With these results, the first question asked at the beginning of the chapter – do the individual gains add up – can be answered positively. The answer to the second question – are there side-effects or interactions between the different strategies – can be answered rather in an indirect way: the results indicate that the use of additional (source) context features by integrating a discriminative classifier can compensate for a certain negative effect introduced by the reordering approach.

8.4.3 Extended variants of the discriminative classifier

The results obtained with the basic classifier trained on *word/pos/lemma/dependency* information lead to further improvements in combination with the morpho-syntactic strategies. With regard to the two problems suspected to be introduced by the reordering, loss of context for multi-word expressions and verbal translations, two extensions to the basic discriminative model are investigated. For these experiments, the reordered and non-reordered variants of the inflection prediction model are contrasted.

Support verb constructions

To integrate explicit information about support verb constructions in addition to the dependency information, two annotation variants are compared: the addition of support verb status as a binary feature (yes/no conditioned on a log-likelihood threshold) to the features of classifier VW-1, and the degree of association (log-likelihood score) to the features of VW-2. The results are displayed in table 8.3. While both variants are better than the surface baseline, they do not result in further improvements over the respective basic classifiers VW-1 and VW-2. The only variant that seems to benefit at least a bit is the non-reordered system with the simpler classifier without dependency features. A gain in this setting is plausible, as the addition of support verb status is a new type of information; however, it is not clear why the corresponding classifier in the reordered setting does not benefit from the support verb information to a similar extent. No benefit is obtained with classifier VW-2: here, information about support verb constructions is however already indirectly contained in the dependency information. The explicit annotation of the degree of association does not seem to provide extra knowledge.

An interesting extension to this type of annotation, though beyond the scope of this work, is the addition of noun information in relation to the expected translation. On one hand, this can concern the similarity of nouns in a similar construction, for example *die Gelegenheit/Möglichkeit/Chance ergreifen* ('to take/seize the opportunity/possibility'), where *Gelegenheit*, *Möglichkeit* and *Chance* are semantically close words with similar translation options for *ergreifen* (literally: 'to grasp', 'to take/seize' in the context of the support verb construction). On the other hand, it seems that many support verb constructions can be translated with a light verb of the same general meaning, regardless of the noun (such as variants of 'take/seize' for constructions with *egreifen*, e.g. *Macht ergreifen*: 'to seize power', *Besitz ergreifen*: 'to take possession' or *Partei ergreifen*: 'to take sides'), while there are some constructions that do not fit into the scheme and require considerably different translations, for example *Wort ergreifen*: 'to rise to speak', *Flucht ergreifen*: 'to escape' or *Beruf ergreifen*: 'to choose a profession/become'. Thus, indicating translation preferences based on noun generalization and translation features might be useful as treatment for support verb constructions.

Verb features

Table 8.4 shows the results obtained when adding number and tense features to the classifiers. In all variants, small improvements can be observed[7], again with a tendency for the reordered system to benefit more from the verbal features.

As BLEU is not an optimal metric to assess small differences in machine translation output, the output of the system with the basic classifier (VW-2-reordered) was compared with the output of the enriched system (reordered VW-2 +Num+Tense). As test set, sentences containing at least one difference in verb translations were extracted, with additionally restricting the source sentence to 8–20 words. After removing sentences with only lexically different verbs, which are not of interest in this evaluation, 155 sentences remained. Three native speakers of German manually rated each pair of differently translated verbs (ignoring all other words) with respect to the following five categories:

[7] Even though small, the difference between 20.50 and 20.62 is statistically significant with pair-wise bootstrap resampling with sample size 1,000 and a p-value of 0.05.

system	**VW-2**	**VW-1 +num**	**VW-2 +num**	**VW-2 +num +tense**
MorphSys	19.93	20.00	20.00	20.02
MorphSys V-Reordered	20.50	20.60	20.57	20.62

TABLE 8.4: Results for systems with a classifier using additional number and tense information.

	better	**worse**	**equal**
number agreement	20	2	4
auxiliary (past/passive)	11	5	2
tense	4	4	2
missing/extra verb	61	20	14
none of the above	0	0	17

TABLE 8.5: Manual evaluation of 155 sentences with regard to the translation of verbs.

- **Number agreement:** subject and verb agree in the feature *number*. The value "equal" can apply if the subject is translated differently, e.g. *research shows* vs. *studies show*.

- **Auxiliary:** presence, absence and the choice of auxiliary are rated. As the translations can have different structures (e.g. compound tense vs. simple tense), the sentences are rated with regard to their respective "technical correctness".

- **Tense**: the translation reproduces the tense in the source sentence, while being technically well-formed at the same time, e.g. *has done* vs. **has did* vs. * $\emptyset$ *done*

- **Missing/extra verb**: refers to the number of full verbs in the sentence. In this category, it is mostly the case that verbs are missing, but it is also possible that a superfluous verb appears in the translation.

- **None of the above:** this category mostly applies to translations of such poor quality that the realization of verbs could not be analyzed properly.

The results in table 8.5 show that the enriched system is better with regard to number agreement between verb and subject, the choice of auxiliary and the number of missing/superfluous verbs. While the differences are not enormous, there is a clear tendency that goes in line with the BLEU improvement. The following examples are selected to illustrate the functioning and the effect of the tense and number features.

(1) EN i really feel that he should follow in the footsteps of the other guys .
really feel that he in the footsteps of the other guys follow should .

VW2 ich bin wirklich der Meinung , dass er in die Fußstapfen der anderen Jungs
i am really of-the opinion , that he in the footsteps of-the other guys
folgen **sollten**$_{\mathbf{PL}}$.
follow should .

VW2$_{NT}$ ich bin wirklich der Meinung , dass er in die Fußstapfen der anderen Jungs
i am really of-the opinion , that he in the footsteps of-the other guys
folgen **sollte**$_{SG}$.
follow should .

The annotation of number is very straightforward, as it is a single piece of information that can be easily obtained by identifying the subject associated to a verb. The effect of annotating number information to a verb is demonstrated by the example in (1), where EN contrasts the original and the reordered input sentence. The output of the system using the basic classifier VW-2 does not have access to the subject's number and as a consequence, produces a verb in plural form for a singular object. Enriched with number/tense information, the output of VW2$_{NumberTense}$ contains the correctly inflected verb form.

(2) EN television footage revealed how numerous ambulances and police cars arrived at a terminal .
television footage revealed how numerous ambulances and police cars at a terminal arrived .

VW2 das Fernsehen zeigte Bilder , wie zahlreiche Rettungswagen und Polizei Autos
the television showed images , how numerous ambulances and police cars
an einem Terminal .
at a terminal .

VW2$_{NT}$ das Fernsehen zeigte Bilder , wie zahlreiche Rettungswagen und Polizei Autos
the television showed images , how numerous ambulances and police cars
an einem Terminal **angekommen** .
at a terminal arrived .

The modeling of tense features is more complex as it involves several verbs, and their effect cannot be explained as easily as in the previous example. A plausible explanation may be that the generally richer feature set leads to a slightly more precise estimation of translation probabilities, resulting in a slight overall improvement concerning verbal translations. For example, the output of the system enriched with verbal features in (2) contains a verb that is missing in the output of the system with the basic classifier VW-2.

(3) EN it would thus be suitable to assist illegal immigration into the usa .
it would thus suitable be illegal immigration into the usa to assist .

VW2 es **wäre** daher geeignet **sein** , die illegale Einwanderung in die USA zu
it would-be thus suitable be , the illegal immigration into the usa to
unterstützen .
assist .

VW2$_{NT}$ es **wäre** daher ideal , illegale Einwanderung in die USA zu unterstützen .
it would-be thus ideal , illegal immigration into the usa to assist .

In contrast, the VW-2 system in (3) produces the superfluous verb *sein* ('to be'), at the position corresponding to the source-side *be*. However, the verb *wäre* is already a finite verb with the meaning of *would be*, thus making the second *sein* redundant. In the number/tense-enriched classifier VW-2$_{NT}$, *be* is connected with its related verb *would*, and thus might trigger a preference for a translation without a verb in this context, given that *would* $\rightarrow$ *wäre* is already a perfect translation.

8.5 Summary

This chapter presented and combined established approaches to address the linguistic levels *Morphology*, *Syntax* and *Lexical Choice* in an English–German phrase-based system. Combining the strategies showed that the presented approaches are complementary to one another; the overall best result was achieved when combining all three strategies.

Furthermore, the results support the initial assumption that verbal reordering on the source side can have a negative effect on the morphological level (such as subject-verb agreement), and on the lexical level (such as the translation of multi-word expressions) due to separating the verb from its arguments. The addition of a discriminative classifier trained on rich source-side context can compensate for the loss of context. Enriching the standard features with information covering verbal inflection leads to a further improvement, that has also been confirmed in a manual evaluation.

Chapter 9

Conclusion and Ideas for Future Work

This chapter summarizes the main findings and results of the thesis, and outlines some ideas for future work to address the limitations of the thesis, and to also go into the direction of the newly evolving neural machine translation.

9.1 Conclusion

This thesis investigated strategies to integrate morpho-syntactic and semantic information into an English–German statistical machine translation system. The basis of the research, a morphology-aware translation system that first translates into an underspecified stemmed representation with a component to generate target-side inflected forms, was extended with rich source- and target-side features to better handle subcategorization and to support the selection of prepositions.

The *two-step underspecification-generation system* setup allowed to efficiently arrange *source-side* and *target-side processes*: for the basic inflection prediction (chapter 3), the translation step on the lexical level was handled separately from target-side inflection; to improve the modeling of grammatical case, the feature prediction model was enriched with selected additional features (chapter 4). In a first attempt to improve the translation of prepositions, chapter 5 made use of noun class information to model selectional preferences. To model subcategorization across complement types, NPs and PPs, the realization of prepositions and case markers was then moved to a stage that allowed better access to relevant features (chapters 6 and 7). The generation of phrase-table entries conditioned on the translation context finally led to improved translation results. In chapter 8, the focus shifted to a higher level by investigating combinations of strategies to model problems across linguistic levels, rather than focusing on one level. The methods to address morpho-syntactic issues as well as lexical choice turned out to be complementary; in particular, the use of rich source-context features was found to compensate for the loss of context information introduced by source-side reordering.

An important step in explicitly modeling source-side and target-side processes consisted in *balancing differences between the source and the target side* to make relevant features accessible to the system – either during translation or at post-processing time – while deleting features from stages where they are not needed. The effect of adapting source- and target-side (structural) differences was investigated at different stages: (i) on the *morphological level*: adapting inflectional features on the source and target side (chapters 3 and 4); (ii) for *local syntactic structure*: representation of complements (NPs vs. PPs) in chapters 6 and 7; and (iii) for *global syntactic structure* where source-side reordering of verbs was applied as one (already established) strategy in chapter 8.

While not all experiments presented in the thesis were successful, the concept of generating surface words from an intermediate representation – as post-processing step for the inflection prediction system, or to create synthetic phrases to model complement types – turned out to work well for the studied problems. It is a powerful concept as it allows to address the main problems when translating into a morphologically complex language: data sparsity and generalization issues can be diminished by using an underspecified representation to learn lexical relations between the source side and the target side. This effect is obvious for the reduction of surface forms to stems, but also comes into play for the placeholder prepositions, where the main benefit may rather lie in balancing source- and target-side structures. To a certain extent, however, the "meaning" or the use of a preposition in a particular context can be considered as insufficiently covered or even uncovered, even though the preposition itself does occur in the training data; thus creating a different type of sparsity problem. In the generation step, the surface form (either an inflectional variant of the stem, or a preposition) can be produced according to the needs of the context based on selected token-level features instead of relying on contextually potentially inadequate type-level statistics. Also, with the generation process being independent from the observed forms in the training data, novel items can be generated, either inflectional variants from seen stems, or phrases containing "new" prepositions with their respective translation probabilities.

9.2 Ideas for future work

This section discusses the limitations of the research presented in this thesis and outlines ideas for future work, both for statistical and neural machine translation scenarios.

9.2.1 Addressing the limitations of the thesis

The research as presented in this thesis has two main limitations: the modeling of inflection prediction and the extensions to improve the translation of complements is tailored to the *translation direction* English–German, with a heavy focus on modeling *noun phrases and prepositional phrases*. The following sections outline ideas to overcome these limitations.

Limitations to noun phrases and prepositional phrases The main objective of the thesis, improving the realization of complement types and the choice of prepositions, concentrated on modeling nominal inflection. So far, verbs constituted relevant context information in many of the experiments, but they were only "consulted" as a lexical feature, or as a predicate with particular subcategorization requirements. The modeling of verbal morphology has mostly been neglected in this work, with the exception of using simple verb information as part of the discriminative classifier in chapter 8.

German verbal morphology is quite complex: verbal inflection is often not restricted to a continuous, locally restricted phrase such as an NP or PP, but it can be discontinuous with several auxiliary and/or modal verbs distributed over the sentence in addition to the main verb. Also, the modeling of verbs is tied to the complements it governs; to a certain extent, one could argue that the assignment of syntactic functions is a prerequisite to the modeling of verbal inflection – at least, the subject needs to be identified to ensure subject-verb agreement. On the other hand, tangible verb information, such as number information, is likely to help the prediction task. An obvious approach is thus to combine the modeling of verbal and nominal inflection; this is in line with Ramm et al. (2016) who find that the prediction of verbal features as a further post-processing step in the inflection prediction setting does not lead to much improvement.

When adding verbal morphology to the translation setup, there are two layers to be considered: general morphological features such as number, tense, mood, and the question of how to realize them. For example, past tense can be expressed through one verb form (*I made*), or through the combination of auxiliary verbs and a participle (*I have made*). Of course, this decision is not arbitrary, but it is subject to different factors in the sentence context. Preferences and restrictions to form verb tenses vary between languages: for example, German tends to be rather lenient concerning the selection of past tense, while the use of English past tense is more rigorously defined. Furthermore, English and German do not have a morphologically expressed future tense and thus have to resort to a compound tense (*I will make*), whereas Romance languages, for example, can form a future tense morphologically (*je ferai*) or by means of a compound tense (*je vais faire*). Typical errors concerning the realization of verb tenses are technically incorrect compound tenses (such as missing or wrong auxiliary, several finite verbs or no finite verb at all), or a mismatch with the source-side tense.

The "placeholder prediction" technique employed to handle different complement types in chapters 6 and 7 might also be applicable to modeling *chains of verbs*, i.e. the connected verbs within a clause. For example, consider verb formation to have two interacting slots: one slot for the full verb, which always needs to be occupied, and another slot dedicated to auxiliaries, that can hold any number of verbs. By having the auxiliary slot filled with a placeholder in an intermediate representation, the decision about the realization of tense (i.e. the main verb in combination with the auxiliaries), can be shifted to a convenient stage in the translation process. For example, this can be a post-processing step connected to the modeling of nominal inflection, or even involve the creation of synthetic phrases to be used during translation. Using such a method to model verbs is much more complicated than modeling complement types, as it does not only require the *consultation* of relevant information from the entire sentence, but also the *modification* of potentially several verbs distributed throughout the sentence.

Limitations to translation direction and language pairs While the experiments conducted in the scope of this thesis were only carried out for the language pair English–German, the underlying ideas can generally be applied to other languages. For example, the concept of synthesizing complement realization in context-aware phrase-table entries would be interesting for a language with a complex case system such as Finnish.

One of the central research questions of the thesis was the arrangement of linguistic information at different stages in the translation process with the objective to handle source- and target-side processes more efficiently. This question increases in relevance the more the source and the target languages differ. In particular a translation scenario where both the source and the target language are morphologically complex, but have a (mostly) disjoint feature set, is likely to benefit from an optimized arrangement of the respective relevant features during translation and in post-processing components.

The findings and observations made in this work for the language pair English–German thus need to be transferred to a new translation scenario by identifying features that are target-side relevant only, as opposed to features that need to be integrated into the translation model at some point in the process. On a more abstract level, working towards less language-specific specifications to represent relevant linguistic features, including the (temporal) removal or addition of features at particular stages in the translation process, as well as how to efficiently arrange source- and target-side processes during translation, constitutes an interesting extension to this thesis.

9.2.2 Linguistic modeling in neural machine translation

With Neural Machine Translation (NMT) becoming the new state-of-the-art in MT, a different strand of future work is linguistic modeling in NMT. Neural translation systems depart from the setup in SMT to employ a combination of several interacting components (such as translation model, language model, etc), and instead use a single neural network: a recurrent neural network *encodes* the source sentence for another neural network to *decode* by predicting target-side words (Bahdanau et al., 2015).

Modeling morphology in NMT One main bottleneck in neural machine translation is the vocabulary size – a too large vocabulary will result in memory problems and intractable training times. The most effective method so far to reduce vocabulary size is Byte Pair Encoding (BPE: Sennrich et al. (2016)), that segments words based on non-linguistically motivated optimization heuristics. Applying the concept of generating inflected forms from an intermediate abstract representation is a straightforward strategy to address this problem, as mapping inflected forms to lemmas considerably reduces the vocabulary size in a linguistically sound way. For the generation of inflected forms, there are two ways to proceed: the first possibility consists in applying the inflection prediction setup as-is within an NMT system, i.e. the prediction of inflectional features followed by a generation step. The second method foregoes the prediction step by combining the lemma with the full set of inflectional features during translation, followed by a deterministic generation of inflected forms. This is conceptually close to a factored translation model with a generation component operating on lemmas and morphological tags (Koehn et al., 2007a); however, the independent modeling of lemmas and features in a factored translation model turned out to be problematic for the generation step. In contrast, NMT systems can handle the task of effectively learning lemmas and associated tag-feature sequences, which makes an external feature-prediction step redundant. In particular, this avoids the somewhat unsatisfactory setup of the four individual CRFs in the inflection prediction system that model the four inflectional features independently.

Such a lemma-generation approach provides several advantages: a linguistically sound reduction of the word inventory by mapping inflected forms to lemma-feature sequences coupled with better generalization of lexical relations between the source and the target side, and the ability to generate context-appropriate (and potentially novel) inflections from seen stems based on morphological patterns learned from the tag-feature sequences. In a first experiment, the lemma-generation strategy led to promising results for NMT systems trained on small English–German data sets.

As lemmatization on its own is likely not sufficient to bring down the vocabulary size to an amount that NMT systems can comfortably handle, it needs to be combined with an additional segmentation strategy such as BPE. However, using a morphologically complex analysis instead of a "flat" lemmatization containing information about sub-words and derivational morphology (as is provided by SMOR) might be used as a basis for a linguistically sound segmentation into smaller units. In particular, handling productive word formation processes such as the decomposition of compounds into individual components is promising as new and infrequent compounds are likely to occur even in large corpora.[1] The need to further segment words into smaller units depends on the individual properties of the respective language, but for morphologically complex languages with highly productive word-formation processes such as compounding or agglutination (e.g. German, Finnish or Turkish), linguistically motivated segmentation and generation strategies that are also able to cover non-concatenative phenomena such as *Umlautung* or *vowel harmony* might be a good choice.

Evaluation Departing from new ideas for modeling linguistic phenomena in NMT, another perspective is that of evaluation: as NMT is fundamentally different from SMT, an obvious research question is how well problems across linguistic levels are handled in NMT generally, as well as in contrast to SMT systems. In a first evaluation, Bentivogli et al. (2016) find that NMT outperforms state-of-the-art SMT systems with regard to morphology, lexical choice, and in particular, that it is considerably better for long-distance reorderings. Extending evaluation studies to also capture performance at the syntax-semantic interface, such as the realization of complement types or the correspondence of semantic roles between the source and target language, might provide interesting insights into the performance of NMT on its own and in relation to SMT.

[1] For example, the evaluation in chapter 3 showed that novel word forms in the inflection prediction system are mostly compounds.

Summary and outlook While neural machine translation systems are very different from statistical machine translation, many of the key questions explored in this thesis remain the same – how can linguistic information be represented efficiently in the translation model, and to what extent can or should target-side processes be modeled separately from the translation step? With regard to the properties of NMT, particularly concerning the restriction of the vocabulary size, exploring the adaptation of the discussed strategies to an NMT setting is an interesting task for future research, starting with a lemma-generation model to handle target-side inflection as a basis for better word segmentation and for the integration of further linguistic information.

Declaration of Authorship

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen Quellen und Hilfsmittel als die angegebenen benutzt habe.

Marion Di Marco

Bibliography

Agarwal, Abhaya and Alon Lavie (2008). “METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output”. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio, pp. 115–118.

Agirre, Eneko, Aitziber Atutxa, Gorka Labaka, Mikel Lersundi, Aingeru Mayor, and Kepa Sarasola (2009). “Use of Rich Linguistic Information to Translate Prepositions and Grammatical Cases to Basque”. In: *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT)*. Barcelona, Spain, pp. 58–65.

Avramidis, Eleftherios and Philipp Koehn (2008). “Enriching Morphologically Poor Languages for Statistical Machine Translation”. In: *Proceedings of ACL08-HLT*. Columbus, Ohio, pp. 763–770.

Badr, Ibrahim, Rabih Zbib, and James Glass (2008). “Segmentation for Englis-to-Arabic Statistical Machine Translation”. In: *Proceedings of ACL-08: HLT (Short Papers)*. Columbus, Ohio, pp. 153–156.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bazrafshan, Marzieh and Daniel Gildea (2013). “Semantic Roles for String-to-Tree Machine Translation”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pp. 419–423.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico (2016). “Neural versus Phrase-Based Machine Translation Quality: a Case Study”. In: *arXiv:1608.04631*.

Bohnet, Bernd (2010). “Top Accuracy and Fast Dependency Parsing is not a Contradiction”. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*. Beijing, China, pp. 89–97.

Bojar, Ondřej and Kamil Kos (2010). “2010 Failures in English–Czech Phrase-Based MT”. In: *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden, pp. 60–66.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi (2015). "Findings of the 2015 Workshop on Statistical Machine Translation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pp. 1–46.

Braune, Fabienne, Anita Gojun, and Alexander Fraser (2012). "Long Distance Reordering During Search for Hierarchical Phrase-based SMT". In: *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT)*. Trento, Italy, pp. 177–184.

Burlot, Franck, Elena Knyazeva, Thomas Lavergne, and François Yvon (2016). "Two Step MT: Predicting Target Morphology". In: *Proceedings of 13th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, Washington.

Cap, Fabienne, Alexander Fraser, Marion Weller, and Aoife Cahill (2014). "How to Produce Unseen Teddy Bears: Improved Morphological Processing of Compounds in SMT". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Göteborg, Sweden, pp. 579–587.

Cap, Fabienne, Marion Weller, Anita Ramm, and Alexander Fraser (2015a). "CimS - The CIS and IMS Joint Submission to WMT 2015 - Addressing Morphological and Syntactical Differences in English to German SMT". In: *Proceedings of the 10th Workshop on Statistical Machine Translation at EMNLP (System papers)*. Lisbon, Portugal, pp. 84–91.

Cap, Fabienne, Manju Nirmal, Marion Weller, and Sabine Schulte im Walde (2015b). "How to Account for Idiomatic German Support Verb Constructions in Statistical Machine Translation". In: *Proceedings od the 11th Workshop on Multiword Expressions (MWE)*. Denver, Colorado, pp. 19–28.

Carpuat, Marine and Dekai Wu (2007). "Improving Statistical Machine Translation Using Word Sense Disambiguation". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, pp. 61–72.

Carpuat, Marine and Mona Diab (2010). "Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation". In: *Proceedings of the 2010 Annual Conferencce of the North American Chapter of the Association for Compuational Linguistics (NAACL)*. Los Angeles, California, pp. 242–245.

Chahuneau, Victor, Eva Schlinger, Noah A. Smith, and Chris Dyer (2013). "Translating into Morphologically Rich Languages with Synthetic Phrases". In: *Proceedings of the*

2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). Seattle, Washington, pp. 1677–1687.

Charniak, Eugene and Mark Johnson (2005). "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguisticcs (ACL)*. Ann Arbor, Michigan, pp. 173–180.

Chiang, David (2005). "A Hierarchical Phrase-Based Model for Statistical Machine Translation". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguisticcs (ACL)*. Ann Arbor, Michigan, pp. 263–270.

Choi, Jinho D. and Martha Palmer (2012). "Getting the Most out of Transition-Based Dependency Parsing". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, pp. 687–692.

Cholakov, Kostadin and Valia Kordoni (2014). "Better Statistical Machine Translation through Linguistic Treatment of Phrasal Verbs". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pp. 196–201.

Clifton, Ann and Anoop Sarkar (2011). "Combine Morpheme-based Machine Translation qith Post-processing Morpheme Predition". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, pp. 32–42.

Collins, Michael, Philipp Koehn, and Ivona Kučerová (2005). "Clause Restructuring for Statistical Machine Translation". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguisticcs (ACL)*. Ann Arbour, Michigan, pp. 531–540.

Daiber, Joachim and Khalil Sima'an (2015). "Machine Translation with Source-Predicted Target Morphology". In: *roceedings of MT Summit XV (MT Researchers' Track)*. Miami, Forida , USA, pp. 283–296.

Eckle, Judith (1999). "Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora". PhD thesis. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Erk, Katrin, Sebastian Padó, and Ulrike Padó (2010). "A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences". In: *Computational Linguistics* 36.4, pp. 723–763.

Faaß, Gertrud and Kerstin Eckart (2013). "SdeWaC – A Corpus of Parsable Sentences from the Web". In: *Proceedings of the 25th International Conference of the German Society* *for Computational Linguistics and Language Technology (GSCL)*. Darmstadt, Germany, pp. 61–68.

Faruqui, Manaal and Sebastian Padó (2010). “Training and Evaluating a German Named Entity Recognizer with Semantic Generalization”. In: *Proceedings of KONVENS 2010*. Saarbrücken, Germany.

Fellbaum, Christiane (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Formiga, Lluís, Adolfo Hernández, José B. Mariño, and Enric Monte (2012). “Improving English to Spanish Out-of-Domain Translation by Morphology Generalization and Generation”. In: *Proceedings of the Monolingual Machine Translation-2012 Workshop*. San Diego, California, pp. 6–16.

Fraser, Alexander (2009). “Experiments in Morphosyntactic Processing for Translating to and from German”. In: *Proceedings of the fourth Workshop on Statistical Machine Translation (WMT)*. Athens, Greece, pp. 115–119.

Fraser, Alexander, Marion Weller, Aoife Cahill, and Fabienne Cap (2012). “Modeling Inflection and Word-Formation in SMT”. In: *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*. Avignon, France, pp. 664–674.

Fritzinger, Fabienne and Alexander Fraser (2010). “How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing”. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden, pp. 224–234,

Galley, Michael, Mark Hopkins, Kevin Knight, and Daniel Marcu (2004). “What’s in a Translation Rule?” In: *Proceedings of the Human Language Technology Conference, North American Chapter of the Association for Computational Linguistic (HLT-NAACL)*. East Stroudsburg, PA, pp. 273–280.

Gao, Qin and Stephan Vogel (2011). “Utilizing Target-Side Semantic Role Labels to Assist Hierarchical Phrase-Based Machine Translation”. In: *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical TRanslation*. Portland, Oregon, pp. 107–115.

Gimpel, Kevin and Noah A. Smith (2008). “Rich Source-Side Context for Statistical Machine Translation”. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio, pp. 9–17.

Gojun, Anita and Alexander Fraser (2012). “Determining the Placement of German Verbs in English-to-German SMT.” In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Avignon, France, pp. 726–735.

Goldwater, Sharon and David McClosky (2005). "Improving Statistical MT through Morphological Analysis". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*. Vancouver, BC Canada, pp. 676–683.

Gustavii, Ebba (2005). "Target Language Preposition Selection – an Experiment with Transformation-Based Learning and Aligned Bilingual Data". In: *Proceedings of the 1oth Annual Conference of the European Association for Machine Translation (EAMT)*. Budapest, Hungary, pp. 112–118.

Hamp, Birgit and Helmut Feldweg (1997). "GermaNet - a Lexical-Semantic Net for German". In: *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, Spain, pp. 9–15.

Huang, Bryant and Kevin Knight (2006). "Relabeling Syntax Trees To Improve Syntax-Based Machine Translation". In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (NAACL-HLT)*. New York, pp. 240–247.

Huck, Matthias, Aleš Tamchyna, Ondřej Bojar, and Alexander Fraser (2017). "Producing Unseen Morphological Variants in Statistical Machine Translation". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Short Paper)*. Valencia, Spain, pp. 369–375.

Husain, Samar, Dipti Misra Sharma, and Manohar Reddy (2007). "Simple Preposition Correspondence: A Problem in English to Indian Language Machine Translation". In: *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*. Prague, Czech Republic, pp. 51–58.

Hálek, Ondrej, Rudolf Rosa, Aleš Tamchyna, and Ondrej Bojar (2011). "Named Entities from Wikipedia for Machine Translation". In: *Proceedings of the Conference on Theory and Practice of Information Technologies*. Košice, Slovakia, pp. 23–30.

Jeong, Minwoo, Kristina Toutanova, Hisami Suzuki, and Chris Quirk (2010). "A Discriminative Lexicon Model for Complex Morphology". In: *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas.* Denver, Colorado.

Joanis, Eric, Suzanne Stevenson, and David James (2008). "A General Features Space for Automatic Verb Classification". In: *Journal of Natural Language Engineering* 14.3, pp. 337–367.

Kholy, Ahmed El and Nizar Habash (2012). "Translate, Predict or Generate: Modeling Rich Morphology in Statistical Machine Translation". In: *Proceedings of the 16th*

Annual Conference of the European Association for Machine Translation. Trento, Italy, pp. 27–34.

Koehn, Philipp (2010). *Statistical Machine Translation*. 1st. New York, NY, USA: Cambridge University Press. ISBN: 0521874157, 9780521874151.

Koehn, Philipp and Hieu Hoang (2007a). "Factored Translation Models". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, pp. 868–876.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst (2007b). "Moses: Open Source Toolkit for Statistical Machine Translation". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Demonstration Papers)*. Prague, Czech Republic, pp. 177–180.

Koehn, Phillip (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". In: *Proceedings of Machine Translation Summit X*. Phuket, Thailand, pp. 79–86.

Koehn, Phillip and Kevin Knight (2003). "Empirical Methods for Compound Splitting". In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Budapest, Hungary: Association for Computational Linguistics, pp. 187–193.

Kunze, Claudia (2000). "Extension and Use of GermaNet, a Lexical-Semantic Database". In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*. Athens, Greece.

Lavergne, Thomas, Olivier Cappé, and François Yvon (2010). "Practical Very Large Scale CRFs". In: *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Uppsala, Sweden, pp. 504–513.

Liu, Ding and Daniel Gildea (2008). "Improved Tree-to-String Transducers for Machine Translation". In: *Proceedings of the Third Workshop on Statistical Machine Translation (WMT)*. Columbus, Ohio, USA, pp. 62–69.

Liu, Ding and Daniel Gildea (2010). "Semantic Role Features for Machine Translation". In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*. Beijing, China, pp. 716–724.

Lo, Chi kiu and Dekai Wu (2011). "MEANT: An Inexpensive, High-Accuracy, Semi-Automatic Metric for Evaluation Translation Utility via Semantic Frames". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, Oregon, USA, pp. 220–229.

Loáiciga, Sharid, Thomas Meyer, and Andrei Popescu-Belis (2014). "English–French Verb Phrase Alignment in Europarl for Tense TranslationModeling". In: *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*. Reykjavik, Iceland, pp. 674–681.

Luong, Minh-Thang, Preslav Nakov, and Min-Yen Kan (2010). "A Hybrid Morpheme-Word Representation for Machine Translation of Morphologically Rich Languages". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Massachusetts, USA, pp. 148–157.

Marton, Yuval and Philip Resnik (2008). "Soft Syntactic Constraints for Hierarchical Phrase-Based Translation". In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*. Columbus, Ohio, pp. 1003–1011.

Minkov, Einat, Kristina Toutanova, and Hisami Suzuki (2007). "Generating Complex Morphology for Machine Translation". In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic, pp. 128–135.

Naskar, Sudip Kumar and Sivaji Bandyopadhyay (2006). "Handling of Prepositions in English to Bengali Machine Translation". In: *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*. Trento, Italy, pp. 89–94.

Navigli, Roberto (2006). "Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance". In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, pp. 105–112.

Nießen, Sonja and Hermann Ney (2001). "Toward Hierarchical Models for Statistical Machine Translation of Inflected Languages". In: *39th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Workshop on Data-Driven Machine Translation*. Toulouse, France, pp. 1–8.

Nădejde, Maria, Alexandra Birch, and Philipp Koehn (2016). "Modeling Selectional Preferences of verbs and Nouns in String-to-Tree Machine Translation". In: *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pp. 32–42.

Och, Franz Josef and Hermann Ney (2003). "A Systematic Comparison of Various Statistical Alignment Models". In: *Computational Linguistics* 29(1), pp. 19–51.

Oflazer, Kemal and İlknur Durgar El-Kahlout (2007). "Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation". In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, pp. 25–32.

Palmer, Martha, Paul Kingsbury, and Daniel Gildea (2005). “The Proposition Bank: An Annotated Corpus of Semantic Roles”. In: *Computational Linguistics* 31.1, pp. 71–105.

Palmer, Martha, Hoa Trang Dang, and Christiane Fellbaum (2007). “Making fine-grained and coarse-grained sense distinctions, both manually and automatically”. In: *Journal of Language Engineering* 13.2, pp. 137–163.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). “BLEU: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, pp. 311–318.

Peral, Jesús and Antonio Ferrández (2003). “Translation of Pronomial Anaphora between English and Spanish: Discrepancies and Evaluation”. In: *Journal of Artificial Intelligence Research* 18, pp. 117–147.

Popović, Maja and Hermann Ney (2004). “Towards the Use of Word Stems and Suffixes for Statistical Machine Translation”. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal, pp. 1585–1588.

Popović, Maja and Mihael Arčan (2015). “Identifying Main Obstacles for Statistical Machine Translation of Morphologically Rich South Slavic Languages”. In: *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*. Antalya, Turkey, pp. 97–104.

Popović, Manja and Hermann Ney (2009). “Syntax-Oriented Evalaution Measure for Machine Translation Output”. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, pp. 29–32.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ramm, Anita and Alexander Fraser (2016). “Modeling Verbal Inflection for English to German SMT”. In: *Proceedings of the First Conference of Machine Translation (WMT16)*. Berlin, Germany, pp. 21–31.

Rozovskaya, Alla and Dan Roth (2013). “Joint Learning and Inference for Grammatical Error Correction”. In: *Proceedings of the 2013 Conference on Empirical Maethods in Natural Language Processing*. Seattle, Washingtion, USA, pp. 791–802.

Sadat, Fatiha and Nizar Habash (2006). “Combination of Arabic Preprocessing Schemes for Statistical Machine Translation”. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics,* Morristown, New Jersey, pp. 1–8.

Scheible, Silke, Sabine Schulte im Walde, Marion Weller, and Max Kisselew (2013). "A Compact but Linguistically Detailed Database for German Verb Subcategorisation Relying on Dependency Parses from a Web Corpus: Tool, Guidelines and Resources". In: *Proceedings of the 8th Web as Corpus Workshop*. Lancaster, UK, pp. 63–72.

Schmid, Helmut (2000). "LoPar: Design and Implementation". In: *Arbeitspapiere des Sonderforschungsbereiches 340*. 149. IMS Stuttgart.

Schmid, Helmut (2004). "Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors". In: *Proceedings of the International Conference on Computational Linguistics*. Geneva, Switzerland, pp. 162–168.

Schmid, Helmut (2006). "Trace Prediction and Recovery With Unlexicalized PCFGs and Slash Features". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Sydney, Australia, pp. 177–184.

Schmid, Helmut, Arne Fitschen, and Ulrich Heid (2004). "SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection". In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal, pp. 1263–1266.

Schulte im Walde, Sabine (2002). "A Subcategorisation Lexicon for German Verbs Induced from a Lexicalized PCFG". In: *Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC)*. Las Palmas de Gran Canaria, Spain, pp. 1351–1357.

Schulte im Walde, Sabine (2006). "Experiments on the Automatic Induction of German Semantic Verb Classes". In: *Computational Linguistics* 32.2, pp. 159–194.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pp. 1715–1725.

Shilon, Reshef, Hanna Fadida, and Shuly Wintner (2012). "Incorporating Linguistic Knowledge in Statistical Machine Translation: Translating Prepositions". In: *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*. Avignon, France, pp. 106–114.

Stymne, Sara and Maria Holmqvist (2008). "Processing of Swedish Compounds for Phrase-Based Statistical Machine Translation". In: *In Proceedings of the 12th annual conference of the European Association for Machine Translation (EAMT)*. Hamburg, Germany, pp. 182–191.

Stymne, Sara, Nicola Candedda, and Lars Ahrenberg (2011a). "Generation of Compound Words in Statistical Machine Translation into Compounding Languages". In: *Computational Linguistics* 39(4), pp. 1067–1108.

Stymne, Sara and Nicola Cancedda (2011b). "Productive Generation of Compound Words in Statistical Machine Translation". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT)*. Edinburgh, Scotland/UK, pp. 250–260.

Subotin, Michael (2011). "An Exponential Translation Model for Target-Language Morphology". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, Oregon, pp. 230–238.

Suzuki, Hisami and Kristina Toutanova (2006). "Learning to Predict Case Markers in Japanese". In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*. Sydney, Australia, pp. 1049–1056.

Tamchyna, Aleš, Fabienne Braune, Alexander Fraser, Marine Carpuat, Hal Daume III, and Chris Quirk (2014). "Integrating a Discriminative Classifier into Phrase-based and Hierarchical Decoding". In: *The Prague Bulletin of Mathematical Linguistics* 101, pp. 29–41.

Tamchyna, Aleš, Alexander Fraser, Ondřej Bojar, and Marcin Junczys-Dowmunt (2016). "Target-Side Context for Discriminative Models in Statistical Machine Translation". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Berlin, Germany, pp. 1704–1714.

Tiedemann, Jörg, Filip Ginter, and Jenna Karvena (2015). "Morphological Segmentation and OPUS for Finnish–English Machine Translation". In: *Proceedings of the fourth Workshop on Statistical Machine Translation (WMT)*. Lisboa, Portugal, pp. 177–183.

Tiedemann, Jörg, Fabienne Cap, Jenna Karvena, Filip Ginter, Sara Stymne, Robert Östling, and Mariob Weller-Di Marco (2016). "Phrase-Based SMT for Finnish with More Data, Better Models and Alternative Alignment and Translation Tools". In: *Proceedings of the First Conference of Machine Translation (WMT16): Shared Task Papers*. Berlin, Germany, pp. 391–398.

Toutanova, Kristina and Hisami Suzuki (2007). "Generating Case Markers in Machine Translation". In: *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*. Rochester, NY, pp. 49–56.

Toutanova, Kristina, Hisami Suzuki, and Achim Ruopp (2008). "Applying Morphology Generation Models to Machine Translation". In: *Proceedings of ACL08-HLT*. Columbus, Ohio, pp. 514–522.

Tsvetkov, Julia, Chris Dyer, Lori Levin, and Archna Bhatia (2013). "Generating English Determiners in Phrase-Based Translation with Synthetic Translation Options". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria, pp. 271–280.

Virpioja, Sami, Jaakko Väyrynen, Mathias Creutz, and Markus Sadeniemi (2007). "Morphology-Aware Statistical Machine Translation Based on Morphology Induced in an Unsupervised Manner". In: *Proceedings of the Machine Translation Summit XI*. Copenhagen, Denkmar, pp. 1–8.

Virpioja, Sami and Stig-Arne Grönroos (2015). "LeBLEU: N-gram-based Translation Evaluation Score for Morphologically Complex Languages". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pp. 411–416.

Weller, Marion (2009). "Separate Morphologiebehandlung als Methode zur Verbesserung statistischer maschineller Übersetzung". Studienarbeit. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Weller, Marion, Max Kisselew, Svetlana Smekalova, Alexander Fraser, Helmut Schmid, Nadir Durrani, Hassan Sajjad, and Richárd Farkas (2013a). "Munich-Edinburgh-Stuttgart Submissions at WMT13: Morphological and Syntactic Processing for SMT." In: *In Proceedings of the Eighth Workshop on Statistical Machine Translation (System papers)*. Sofia/Bulgaria, pp. 232–239.

Weller, Marion, Alexander Fraser, and Sabine Schulte im Walde (2013b). "Using Subcategorization Knowledge to Improve Case Prediction for Translation to German". In: *Proceedings of the Association for Computational Linguistics (ACL)*. Sofia, Bulgaria, pp. 593–603.

Weller, Marion, Alexander Fraser, and Ulrich Heid (2014a). "Combining Bilingual Terminology Mining and Morphological Modeling for Domain Adaptation in SMT". In: *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*. Dubrovnik, Croatia, pp. 11–18.

Weller, Marion, Sabine Schulte im Walde, and Alexander Fraser (2014b). "Using Noun Class Information to Model Selectional Preferences for Translating Prepositions in SMT." In: *Proceedings of the Association for Machine Translation in the Americas*. Vancouver, BC Canada, pp. 275–287.

Weller, Marion, Alexander Fraser, and Sabine Schulte im Walde (2015). "Target-Side Generation of Prepositions for SMT". In: *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*. Antalya, Turkey, pp. 177–184.

Weller-Di Marco, Marion, Alexander Fraser, and Sabine Schulte im Walde (2016). “Modeling Complement Types in Phrase-Based SMT”. In: *In Proceedings of the First Conference of Machine Translation (WMT16)*. Berlin, Germany, pp. 43–53.

Weller-Di Marco, Marion, Alexander Fraser, and Sabine Schulte im Walde (2017). “Addressing Problems across Linguistic Levels in SMT: Combining Approaches to Model Morphology, Syntax and Lexical Choice”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Short Paper)*. Valencia, Spain, pp. 625–630.

Williams, Philip and Philipp Koehn (2011). “Agreement Constraints for Statistical Machine Translation into German”. In: *Proceedings of the 6th Workshop on Statistical Machine Translation (WMT)*. Edinburgh, UK, pp. 217–226.

Williams, Philip and Philipp Koehn (2012). “GHKM-Rule Extraction and Scope-3 Parsing in Moses”. In: *Proceedings of the 7th Workshop on Statistical Machine Translation*. Montréal, Canada, pp. 388–394.

Williams, Philip, Rico Sennrich, Maria Nădejde, Matthias Huck, and Philpp Koehn (2015). “Edinburgh’s Syntax-Based Systems at WMT 2015”. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pp. 199–209.

Wu, Dekai and Pascale Fung (2009a). “Can Semantic Role Labeling Improve SMT?” In: *Proceedings of the 13th Annual Conferencce of the European Association for Machine Translation (EAMT)*. Barcelona, Spain, pp. 218–225.

Wu, Dekai and Pascale Fung (2009b). “Semantic Roles for SMT: A Hybrid Two-Pass Model”. In: *Proceedings of NAACL HLT 2009 (Short Papers)*. Boulder, Colorado, pp. 13–16.

Wu, Hua, Haifeng Wang, and Chengquing Zong (2008). “Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual Corpora”. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling)*. Manchester, UK, pp. 993–1000.

Yang, Mei and Katrin Kirchhoff (2006). “Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages”. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Trento, Italy, pp. 41–48.

Yeniterzi, Reyyan and Kemal Oflazer (2010). “Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish”. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Uppsala, Sweden, pp. 454–464.

Zollmann, Andreas and Stephan Vogel (2011). “A Word-Class Approach to Labeling PSCFG Rules for Machine Translation”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, Oregon, pp. 1–11.