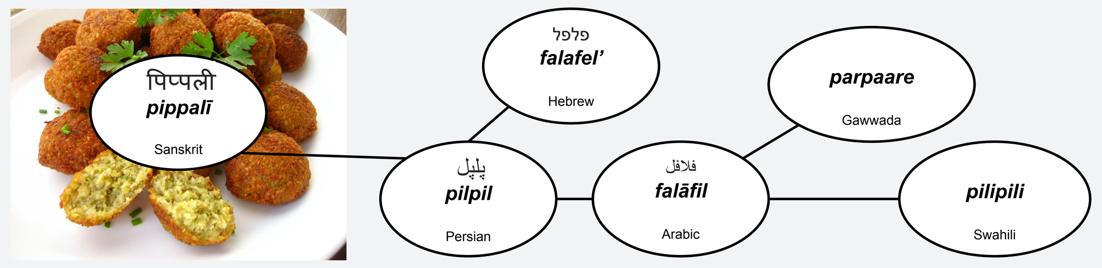
Learning an Optimality Theoretic Model of Lexical Borrowing

Yulia Tsvetkov Waleed Ammar Chris Dyer
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA

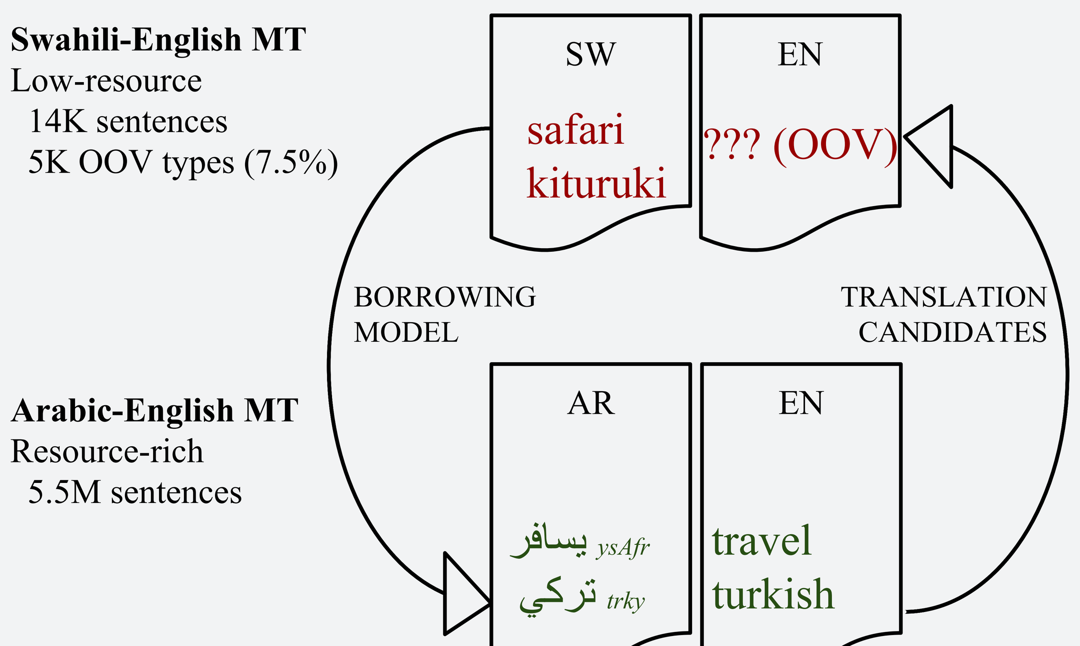
Lexical Borrowing

Lexical borrowing – adoption and nativization of words from another language; it happens when more than one language meet at the same place and over a period of time. Borrowing is pervasive in a majority of the world's languages and is a fundamental research topic in linguistics. In computational linguistics, however, no prior work has addressed modeling language contact–induced linguistic borrowing.

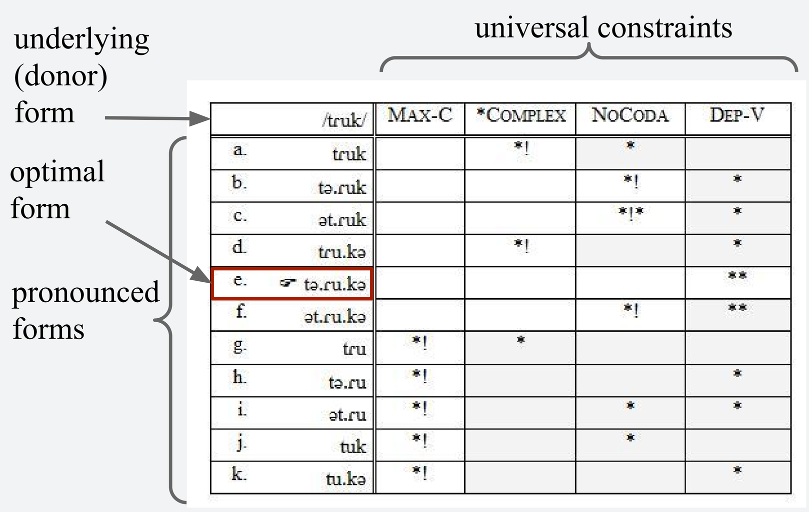


This Work

1. We present a semi-supervised generative model of lexical borrowing based on the Optimality Theory
2. The borrowing model helps improve low-resource statistical machine translation

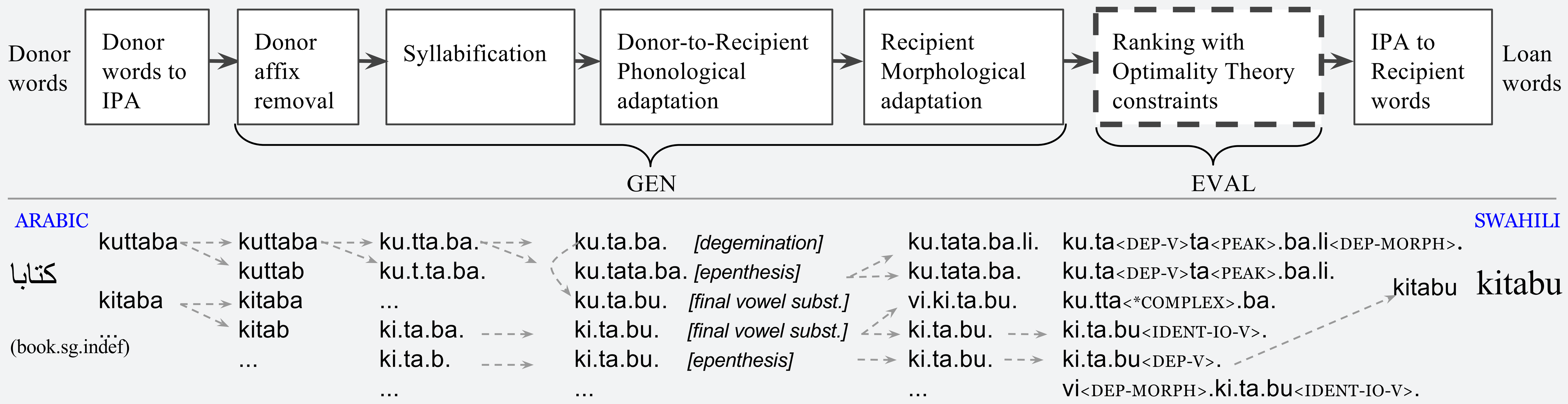


Optimality Theory (OT)



OT is a theory of phonology which accounts for sound patterns via *constraints*. OT analyzes the surface words of a language as emerging from *underlying forms* (abstract phoneme sequences) according to a two-stage process: (1) all candidates are generated (the GEN phase); (2) the candidates are evaluated, and the the most optimal realization of the underlying form wins (this surface form most closely conforms to the phonological preferences of the language).

Lexical Borrowing Model



After generating multiple plausible syllabifications of an Arabic word, each syllabified phonetic sequence undergoes phonological and morphological adaptation to comply with Swahili syllable structure, phonology, and morphology. This adaptation leads to a potentially infinite set of generated 'underlying representations' of a Swahili loanword from which an optimal form must be chosen (*a.k.a.* GEN in OT). The underlying representations are then evaluated using a set of strictly ordered (violable) constraints (CON), and an automatically learned constraint ranking that is aimed to assign higher score to underlying forms with fewer violations of higher-ranked constraints (EVAL). Finally, the winning underlying forms are converted to their surface realizations. We employ the Nelder-Mead simplex method to iteratively optimize constraint weights.

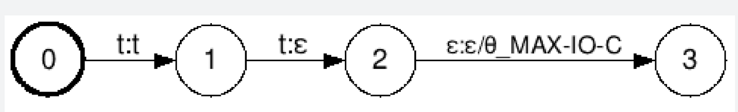
Donor-to-Recipient Phonological Adaptation

Vowel deletion – shortening of Arabic long vowels and vowel clusters
Consonant degemination – shortening of Arabic geminate consonants
Substitution of similar phones – /tˤ/→/t/, /dˤ/→/d/, /sˤ/→/s/, etc.
Vowel epenthesis – eliminating Arabic codas and consonant clusters
Final vowel substitution – /u/, /o/, /i/, /e/

Faithfulness Constraints

MAX-IO-MORPH	no (donor) affix deletion	746
MAX-IO-C	no consonant deletion	310
MAX-IO-V	no vowel deletion	156
DEP-IO-MORPH	no (recipient) affix epenthesis	250
DEP-IO-V	no vowel epenthesis	168
IDENT-IO-P	no pharyngeal consonant substitution	1190
IDENT-IO-C	no consonant substitution	1137
IDENT-IO-G	no glottal consonant substitution	698
IDENT-IO-F	no final vowel substitution	404
IDENT-IO-E	no emphatic consonant substitution	396
IDENT-IO-V	no vowel substitution	0

Faithfulness constraints prefer pronounced realizations completely congruent with their underlying forms.

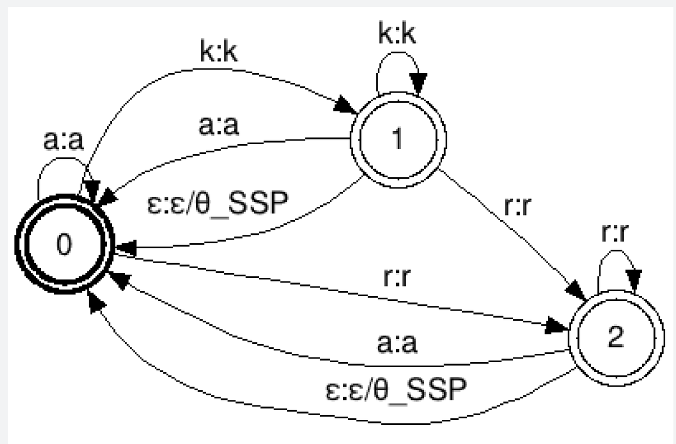


Faithfulness constraints are integrated in phonological transformation transducers as weighted transitions following each transformation. For example, the MAX-IO constraint transition is integrated in the consonant degemination transducer.

Markedness Constraints

NO-CODA	syllables must not have a coda	1722
*COMPLEX-S	no consonant clusters on syllable margins	1047
SSP	complex onsets rise in sonority	589
*COMPLEX-C	no consonant clusters within a syllable	186
PEAK	there is only one syllabic peak	175
*COMPLEX-V	no vowel clusters	173
ONSET	syllables must have onsets	158

Markedness constraints impose language-specific structural well-formedness of surface realizations. Constraints are aligned with their weights, learned by the borrowing model. Higher scores correspond to constraints that are harder to violate, since hypotheses with the highest harmony have shortest paths in the loanwords transducer.



SSP constraint transducer example, for the subset of phonemes /a/, /r/, /k/ and only for complex onsets (codas falling sonority evaluation is not practical in Swahili, as it prohibits codas). According to the sonority scale, /r/ is ranked higher than /k/, therefore when in onset position /kr/ is a non-violating sequence, and /rk/ violates the SSP constraint.

Datasets

1. Arabic and Swahili pronunciation dictionaries (700K and 312K word types)
2. Arabic–English and Swahili–English bitexts (5.4M and 14K sentence pairs)
3. Automatically extracted (bitext alignments plus Levenshtein distance heuristics) 490 Arabic–Swahili borrowing examples: 417 for model parameter optimization, and 73 (15%) for eval.

Intrinsic Evaluation

Model design Reachability is a percentage of donor-recipient pairs that are reachable from Swahili to Arabic. Ambiguity is an average number of outputs that the model generates per one input.

	Dev	Test
Reachability	81.3%	87.7%
Ambiguity	2,033	2,407

Model accuracy The baselines are orthographic (surface) and phonological (based on pronunciation lexicon) Levenshtein distance, heuristic Levenshtein distance with lower penalty on vowel updates (Levenshtein-H), CRF transliteration, and our model with uniform and learned OT constraint weights.

	Acc. (%)
Levenshtein (surface)	8.9
CRF (surface)	16.4
Levenshtein (phon.)	19.8
Levenshtein-H (phon.)	19.7
OT-uniform	35.9
OT	**52.0**

Extrinsic Evaluation

Swahili–English MT performance is improved when we integrate translations of OOV loanwords leveraged from the Arabic–English MT, using generated Arabic donors as pivot.

	BLEU
Baseline	18.0±.2
+ OOV loanwords	**18.5**±.1

Acknowledgements

Sponsored by the U.S. Army Research Lab and the U.S. Army Research Office under grant number W911NF-10-1-0533, and Google Cloud.