doi:10.1017/S1351324910000124

Directional distributional similarity for lexical inference

LILI KOTLERMAN[1],
IDO DAGAN[2], IDAN SZPEKTOR[3]
and MAAYAN ZHITOMIRSKY-GEFFET[4]
[1]*Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel*
e-mail: lili.dav@gmail.com
[2]*Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel*
e-mail: dagan@cs.biu.ac.il
[3]*Yahoo! Research, Building 30 Matam Park, Haifa 31905, Israel*
e-mail: idan@yahoo-inc.com
[4]*Department of Information Science, Bar-Ilan University, Ramat Gan, Israel*
e-mail: maayan.geffet@gmail.com

(*Received 8 February 2009; revised 2 March 2010; accepted 14 May 2010*)

Abstract

Distributional word similarity is most commonly perceived as a symmetric relation. Yet, directional relations are abundant in lexical semantics and in many Natural Language Processing (NLP) settings that require lexical inference, making symmetric similarity measures less suitable for their identification. This paper investigates the nature of directional (asymmetric) similarity measures that aim to quantify distributional feature inclusion. We identify desired properties of such measures for lexical inference, specify a particular measure based on Average Precision that addresses these properties, and demonstrate the empirical benefit of directional measures for two different NLP datasets.

1 Introduction

Many works on automatic identification of semantic term similarity exploit distributional similarity, assuming that terms that appear in similar contexts are semantically similar. This has been now an active research area for a couple of decades in which many measures were proposed for detecting lexical similarity (Hindle 1990; Jiang and Conrath 1997; Lin 1998a; Turney 2001; Weeds and Weir 2003).

While distributional similarity is most prominently modeled by symmetric measures, quite a few semantic similarity relations are directional (asymmetric). For example, most of the ontological (WordNet-style) word-interrelationships, like hyponym/hypernym ({skyscraper, building}), meronym/holonym ({window, building}), and value/attribute ({slow, velocity}) are directional relations.

In addition to ontological relations between words, in many Natural Language Processing (NLP) applications, such as Question Answering (QA), Information Retrieval (IR), Information Extraction (IE), and Text Categorization (TC), it is crucial to recognize whether a specific target meaning can be inferred from a given text. For example, a QA system has to deduce that '*John bought a boat*' can be inferred from '*John bought a skiff*' to answer '*Did John buy a boat?*'. As another example, a TC system usually expands a target category with related words. Identifying directional inference relations between terms is often required as part of the inference process in these applications, e.g. that *skiff* entails *boat* in the above example. Typically, symmetric distributional similarity measures have been utilized in these applications to identify inferential relations (Jing and Croft 1994; Xu and Croft 1996; Mandala, Tokunaga and Tanaka 1999; Zazo *et al.* 2005). However, the required lexical relation is directional in nature. For example, an IR user looking for 'sport events' will be satisfied with documents about 'baseball events', since baseball is a subtype of sport, but not vice versa. Therefore, symmetric similarity measures might be less suitable for identifying inferential relationships.

Despite the evident need of directional similarity measures for lexical inference, their investigation counts, to the best of our knowledge, relatively few works (Dagan, Lee and Pereira 1999; Lee 1999; Weeds and Weir 2003; Geffet and Dagan 2005; Bhagat, Pantel and Hovy 2007; Szpektor and Dagan 2008; Clarke 2009). So far, existing research in this direction did not penetrate broadly to the common practice in NLP applications, in particular in applied inference, where symmetric association measures, namely LIN (Lin 1998a) and cosine, are prominently used. Furthermore, most of these measures were not compared within the same experiment.

This paper aims to investigate the nature of directional similarity measures for applications that require lexical inference. We first conduct a thorough analysis of the behavior of current directional measures. Based on our analysis we identify several properties that directional measures should desirably follow. Since the state-of-the-art measures do not meet all of these properties, we design a novel directional measure that aims to satisfy these properties.

We demonstrate our measure's advantage over state-of-the-art measures in identifying directional similarity, as well as its empirical advantage under two distinct evaluation settings for lexical inference, Automatic Content Extraction (ACE) and Automatic Text Categorization (ATC). Our carefully designed measure significantly outperformed all other tested measures. In a broader prospect, we suggest that asymmetric measures might be more suitable than symmetric ones for many other settings as well.[1]

2 Background

This section presents the background material necessary to understand the contributions of this paper in the area of directional distributional similarity. First, the Textual Entailment (TE) framework is presented, introducing lexical entailment, a prominent type of lexical inference utilized by many NLP applications.

[1] Our directional term-similarity resource will be available at http://www.cs.biu.ac.il/~nlp/downloads/

Positive example:

T: The drugs that slow down or halt Alzheimer's disease work best the earlier you administer them.

H: Alzheimer's disease is treated using drugs.

Negative example:

T: Arabic, for example, is used densely across North Africa and from the Eastern Mediterranean to the Philippines, as the key language of the Arab world and the primary vehicle of Islam.

H: Arabic is the primary language of the Philippines.

Fig. 1. An example of a positive (entailing) text–hypothesis pair and a negative (nonentailing) pair, taken from the RTE-2 development dataset (Bar-Haim *et al.* 2006).

Following, we describe related work on distributional similarity between terms. The distributional term similarity scheme follows two steps. First, a feature vector is constructed for each term by collecting context words as features. Each feature is assigned a weight indicating its 'relevance' to the given term. Then, term vectors are compared by some vector similarity measure. The main difference between symmetric and directional similarity measures lies in the manner in which the two vectors are compared. This step is the focus of our paper. We present an overview of both symmetric and directional state-of-the-art similarity measures along with the motivations behind their design.

Finally, we briefly describe an alternative approach, which suggests extracting similarities from the existing lexical resources created by humans.

2.1 Textual Entailment

The applied inference required by many NLP applications, such as QA, IE and IR, may be addressed by the generic **Textual Entailment** paradigm (Dagan, Glickman and Magnini 2006). The essence of the TE framework is a directional relation between two texts, termed **text** and **hypothesis**. The TE relation between a text *t* and a hypothesis *h*, denoted t → h, holds if a human reading the text would infer that the hypothesis is most likely true. An example for an entailing and a nonentailing text–hypothesis pairs is presented in Figure 1. Some recent papers showed that utilizing TE engines improves the performance of NLP systems (Harabagiu and Hickl 2006; Harabagiu, Hickl and Lacatusu 2007; Lloret *et al.* 2008; Mirkin, Dagan and Shnarch 2009).

Textual Entailment engines require knowledge resources for identifying entailment relationships between texts. One prominent type of knowledge representation needed for such inference is lexical entailment rules. A **lexical entailment rule** is a directional relation, '*entailing-term* → *entailed-term*', between an entailing term (a.k.a left-hand side or LHS) and an entailed term (a.k.a right-hand side or RHS). A rule is considered correct if the meaning of the RHS term is implied from the meaning of the LHS term. Some rule examples are '*chess* → *game*', '*divorce* → *marriage*' and '*government* → *state*'. We note that synonyms and other symmetric inference relations may be viewed as bidirectional entailment rules, e.g. '*car* ↔ *automobile*' and '*buy* ↔ *purchase*'.

2.2 Symmetric distributional similarity measures

To date, most distributional similarity research concentrated on symmetric measures. Most of these measures assess the relative amount of features common between the two vectors compared to the whole set of features within these vectors. Typical examples are the cosine measure (Salton and McGill 1983; Ruge 1992; Caraballo 1999; Gauch, Wang and Rachakonda 1999; Pantel and Ravichandran 2004) and *Jaccard's coefficient* (Gasperin *et al.* 2001). Another such common symmetric measure is the widely cited and competitive (as shown in Weeds and Weir 2003) LIN measure (Lin 1998a), defined as

$$LIN(u,v) = \frac{\sum_{f \in F_u \cap F_v} [w_u(f) + w_v(f)]}{\sum_{f \in F_u} w_u(f) + \sum_{f \in F_v} w_v(f)}$$

where F_x is the feature vector of a term x, $w_x(f)$ is the weight of the feature f in that term's vector, set to their pointwise mutual information (pmi) (Church and Hanks 1990) and $F_u \cap F_v$ is a set of features that are common for the two vectors F_u and F_v .

In a different approach, Dagan *et al.* (1999) proposed to adopt the *Jensen–Shannon divergence (JS)* measure for feature vector similarity. Viewing each feature vector F as defining a probability distribution over the feature space, JS is the average of the *Kullback–Leibler divergence (KL-divergence)* of each of the two distributions to their average distribution:

$$JS(u,v) = \frac{1}{2}\left[D\left(F_u || \frac{F_u + F_v}{2}\right) + D\left(F_v || \frac{F_u + F_v}{2}\right)\right]$$

2.3 Directional distributional similarity measures

Few works investigated a directional similarity approach for relations that capture some notion of lexical inference (e.g. lexical entailment). To present these measures, we adopt a 'semantic expansion' terminology, in analogy to a typical lexical expansion setting: given a pair of similar terms, we say that a term with a narrower common subcontext *expands* a related term with a broader common subcontext. For example, 'baseball' expands 'sports', but not necessarily vice versa, since typical baseball contexts are also sports contexts. In this paper, we refer to the term with a narrower subcontext as a *narrower term* and to the term with a broader subcontext as a *broader term.* We also denote the relation between a narrower term u and a broader term v by '$u \lesssim v$'.

Lee (1999) investigated distributional similarity measures for improving likelihood estimation of unseen co-occurrences and noticed that substitutability of one term for another is not symmetric. To address this asymmetry, Lee (1999) proposed the *α-skew divergence* as a directional similarity measure based on KL-divergence. This measure was adopted from information theory, similarly to the *JS* measure. However, *α-skew* is directional, calculating the distance to the feature distribution of a narrower term u from the distribution of the broader term v:

$$s_\alpha(u \lesssim v) = D(F_v || \alpha F_u + (1-\alpha)F_v)$$

α lies in the range of [0,1] and serves as a smoothing parameter. The best reported results were achieved for $\alpha = 0.99$. We note that this measure is an approximation of the *KL-divergence* measure, which is itself directional, but is undefined for features present only in the first of the two vectors, but not in the second, and thus is not applicable without smoothing. The *α-skew* measure was shown to outperform all other symmetric measures that were examined in that study, as well as the *KL-divergence* measure itself applied to smoothed feature vectors as proposed by Chen and Goodman (1996). As stated in Roberts (2008), the *α-skew* measure naturally embodies the asymmetry of similarity argued by Tversky (1977) in his work on psychological distance.

As a directional measure analogous to the symmetric *LIN* measure, the *Precision* measure, denoted here *WeedsPrec*, was proposed for identifying the hyponymy relation and other generalization/specification relations (Weeds and Weir 2003; Weeds, Weir and McCarthy 2004). *WeedsPrec* quantifies the weighted coverage (or *inclusion*) of the features of the candidate narrower term u by the features of the broader term v:

$$WeedsPrec(u \lesssim v) = \frac{\sum_{f \in F_u \cap F_v} w_u(f)}{\sum_{f \in F_u} w_u(f)}$$

The assumption behind *WeedsPrec* is that if one term is indeed a generalization of the other then the features of the more specific term are likely to be included in those of the more general one (but not necessarily vice versa).

Geffet and Dagan (2005) extended the rationale of Weeds *et al.* (2004) to the lexical entailment setting. They defined a lexical entailment relation in which the meaning of the broader term should be directly implied or entailed from the meaning of the narrower one. In this framework it is expected that if the meaning of a term u entails that of v, then all its prominent context features (under a certain notion of 'prominence') would be included in the feature vector of v as well. This approach is denoted here as *feature inclusion*. Their web-based experiments revealed a strong empirical correlation between such complete inclusion of prominent features and lexical entailment. Yet, such complete inclusion cannot be feasibly assessed using an off-line corpus, due to the huge amount of data needed to overcome the sparseness of word-feature co-occurrences.

Recently, Szpektor and Dagan (2008) tried to identify the entailment relation between predicative lexical–syntactic templates using the *WeedsPrec* measure, but observed that it tends to promote unreliable relations involving infrequent entailing (narrower) templates. To remedy this, they proposed to balance the directional

WeedsPrec measure by geometrically averaging it with the symmetric *LIN* measure, a measure denoted here as *balPrec*:

$$balPrec(u \lesssim v) = \sqrt{LIN(u,v) \cdot WeedsPrec(u \lesssim v)}$$

Effectively, this measure penalizes narrower templates having short feature vectors (vectors with relatively few features), as those usually yield low symmetric similarity with the longer vectors of more common templates.

Another work on lexical–syntactic templates (Bhagat *et al.* 2007) suggested the learning directionality of inference rules (LEDIR) algorithm for defining the direction of valid similarities created by a symmetric measure. Given two semantically similar templates u and v, LEDIR measures the ratio between the number of features of u and v and applies a threshold to define the similarity direction. The underlying assumption is that if v occurs in significantly more contexts than u, then v most likely has a broader meaning. This assumption seems to conform with the idea of feature inclusion, yet Szpektor and Dagan (2008) show that filtering with the directional LEDIR did not improve the performance of the original symmetric discovery of inference rules from text (DIRT) resource (Lin and Pantel 2001) in their evaluation. In addition, LEDIR learns thresholds using supervised techniques and thus cannot be considered completely unsupervised.

In a recent work of Clarke (2009) the *degree of entailment* measure, denoted here as *ClarkeDE*, was proposed:

$$ClarkeDE(u \lesssim v) = \frac{\sum_{f \in F_u \cap F_v} min(w_u(f), w_v(f))}{\sum_{f \in F_u} w_u(f)}$$

It quantifies weighted coverage of the features of the candidate narrower term u by the features of the broader term v and thus resembles the *WeedsPrec* measure.

Finally, we note that lexical similarity is not the only task in need of asymmetric approaches. Human word associations are asymmetric as well (Tversky 1977; Michelbacher, Evert and Schutze 2007). For example, when hearing the word ‘mango’, ‘fruit’ is one of the first associations that come to mind. But when hearing ‘fruit’, we are more likely to come up with common fruits like ‘apple’ or ‘orange’ rather than the less frequent ‘mango’. We also note that research on directionality is not limited to developing directional similarity measures. Significant contributions were proposed in encoding directionality into the feature vectors in the past few years, including the holographic model (Jones and Mewhort 2007) and permutation-based approaches (Sahlgren, Holst and Kanerva 2008). These approaches are beyond the scope of this paper and typically employ the symmetric cosine similarity between the vectors.

2.4 Extracting similarities from existing lexical resources

A different approach for lexical similarity proposes to utilize the existing knowledge resources, e.g. thesauri, semantic networks, taxonomies, or encyclopedias, instead of relying on distributional properties of words. The resource used is viewed as a network or directed graph and then terms are considered similar or dissimilar based on properties of paths in this graph.

The most prominently used measures of this kind utilize WordNet (Fellbaum 1998), a broad coverage lexical network of English words, organized into synonym sets (synsets). Synsets are connected with each other by variety of lexical relations, such as *synonymy*, *antonymy*, and *hyponymy*. Since in this paper we focus on the distributional approach, we do not go beyond listing several popular WordNet-based measures, implemented within the WordNet::Similarity package (Pedersen, Patwardhan and Michelizzi 2004). This open source package is becoming very popular in different NLP fields and implements amongst others the following competitive measures: *Lch* by Leacock and Chodorow (1998), *Lesk* by Banerjee and Pedersen (2002), *Lin* by Lin (1998c), *Path* by Pedersen *et al.* (2004), *Res* by Resnik (1995), *Vector* by Patwardhan (2003), *Wup* by Wu and Palmer (1994), and *Jcn* by Jiang and Conrath (1997).

An overview of the existing measures can be found in Budanitsky and Hirst (2006). We note that, as argued in Budanitsky and Hirst (2006), all measures of this kind are symmetric.

3 Empirical observations and motivations

With lexical inference in mind, our research goal is to develop a directional similarity measure suitable for identifying asymmetric relations between narrower and broader terms. In the previous section, we surveyed several directional measures that were proposed for a variety of NLP tasks, such as language modeling, learning hyponymy, and entailment recognition. A common behavior is evident for most of these measures: they are based on feature inclusion. In some cases, the feature vectors of the related terms can mutually include each other to some extent, but usually the inclusion is stronger in one direction.

Aiming to quantify most effectively the above notion of feature inclusion, we performed a preliminary analysis of available annotated positive and negative examples for narrower-broader similarity relationships. The purpose of this examination was to identify statistical traits that can be used to distinguish the valid word pairs from the invalid ones. In this section, we summarize the findings of the analysis, which underlie the design of our proposed similarity measure described next in Section 4.

For a candidate pair '$u \lesssim v$' we will refer to the set of features of the narrower term u, which are those tested for inclusion, as *tested features*. Amongst these features, those found in the feature vector of the broader term v are denoted *included features*.

As can be seen from the formulae presented in Section 2, existing statistical inclusion measures aim to capture certain aspects of feature inclusion. The *WeedsPrec* measure attempts to: (i) reflect the proportion of included features amongst the tested ones (the core inclusion idea) and (ii) assign greater importance to included features which have higher weights within the vector of the narrower term. In addition, the *balPrec* measure penalizes unreliable pairs containing infrequent narrower terms. The *ClarkeDE* measure is similar to *WeedsPrec*, but it reduces the weight of included features if they have lower weight within the vector of the broader term.

Table 1. *Examples of valid and invalid entailment pairs analyzed for feature inclusion*

Valid pairs	Invalid pairs
air force → warplane	abuse ↛ bribe
argument → reason	broker ↛ journalist
broker → trader	ceasefire ↛ federation
care → treatment	central bank ↛ army
chairman → chief executive	chairman ↛ founder
debt → loan	murder ↛ war
government → state	performance ↛ success
prison term → sentence	research ↛ management
town → city	town ↛ airport
war → aggression	vessel ↛ warplane

In our preliminary analysis we aimed to

- verify the necessity of the aforesaid aspects in quantifying feature inclusion;
- detect additional properties that an inclusion measure should posess.

To that end we used an available manually annotated collection of valid and invalid lexical entailments (Zhitomirsky-Geffet and Dagan 2009), which contains 1067 valid and 2705 invalid directional pairs, produced by the *LIN* similarity measure for two different feature weighting schemes. Examples of valid and invalid pairs from this dataset are presented in Table 1. The entailed (right-hand side) term is the broader term having broader contexts than the narrower entailing (left-hand side) term. We note that lexical entailment subsumes most of the WordNet-style relationships between narrower and broader terms, as well as other directional relationships. We thus assumed that examining positive and negative examples for lexical entailment pairs would yield indicative observations relevant for most types of directional similarity for lexical inference.

We analyzed the feature vectors of the terms participating in a sample of this dataset. These vectors were created by parsing the Reuters RCV1 corpus using the Minipar dependency parser (Lin 1998b) and taking as features the words related to each term through a dependency relation. Each feature was weighted by its pmi with the term.

A thorough examination of this data led us to hypothesize the following desired properties for a distributional inclusion measure, further referred as the *desired properties*. We thus hypothesize that a directional similarity measure should reflect:

(1) the relevance of included features to the narrower term
(2) the relevance of included features to the broader term
(3) that inclusion detection is less reliable if the number of features of either the narrower or the broader term is small

We note that all prior measures addressed the first property, while the second property was in some way addressed only by the *ClarkeDE* measure and the third

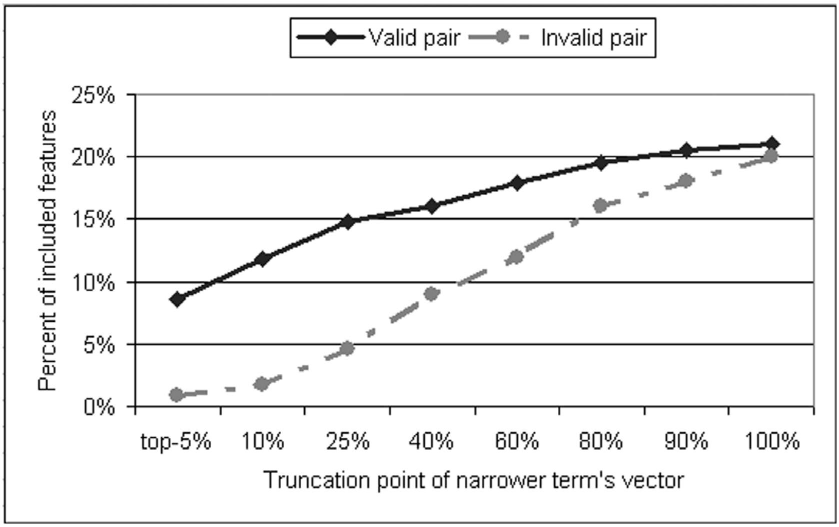

Fig. 2. A typical example of the influence of the tested features according to their rank in the vector of the narrower term. The graph presents the valid pair 'election → vote' compared to the invalid pair 'election ↛ reform'. The x-axis shows different top-n per cents of narrower term's features tested for inclusion; the y-axis shows the corresponding percentage of included features.

property was addressed by the *balPrec* measure only regarding the narrower term. We next explain the rationale behind each of these properties and illustrate them using the analyzed data.

3.1 The relevance of included features to the narrower term

The features within each vector constitute a ranked list according to their weights. It is presumed that features at higher ranks are more important from the perspective of inclusion reliability. For example, inclusion of ten features from the highest ranking tested features (*top features*) is supposed to be of more importance compared to inclusion of ten features scattered along the tail of the feature list.

Indeed, the inclusion of top features of the narrower term was found in our analysis to be more meaningful than inclusion of lower-ranked features. As expected from using pmi as weight, many features at lower ranks are common features in the language that co-occur with many terms and are less indicative specifically for the narrower term. When taking these features into account, the feature inclusion ratios for valid and invalid pairs are becoming similar, making it harder to identify valid pairs. An example of this behavior, which was common to all sampled pairs, is illustrated in Figure 2. As can be seen, when only top features (up to top 30 per cent) of the narrower term are tested for inclusion, the ability to distinguish between valid and invalid pairs is much better, as shown by the larger gap between the inclusion ratios for a valid and an invalid pair.

An additional observation derived from our analysis is that it is preferable to estimate the importance of each included feature based on its rank in the narrower term's vector instead of its pmi weight, as was done in previous works. We observed that pmi-based feature weights demonstrate rather inconsistent behaviors. For example, two highly relevant features, with consecutive ranks, might have very different pmi weights; on the other hand, indicative and nonindicative features that have quite different ranks might have rather similar pmi scores. Moreover, features placed at the same ranks in different vectors typically have considerably different weights, making similarity scores for different pairs incomparable.

Thus, relying on feature weights (as in *ClarkeDE*, *WeedsPrec* and *balPrec*) might yield the following undesired effects:

- Accumulative contribution to the similarity score of many low-ranked, and hence nonindicative, included features might exceed the contribution of fewer but more indicative high-ranked features, thus affecting the ability of the measure to distinguish between valid and invalid pairs.
- Two pairs with exactly the same number of included features placed at exactly the same ranks could receive rather different similarity scores.

In our analysis we did not have a way to measure some quantitative difference between the two approaches, but a qualitative impression indicated that relying on ranking instead of absolute weight score is more stable. We thus expect that considering feature ranks of the narrower term, rather than absolute feature weights, may improve the performance of an inclusion measure. Furthermore, working with ranks is methodologically preferable as it makes the resulting measure applicable for feature vectors created by different feature weighting schemes.

3.2 The relevance of included features to the broader term

We next look at the features of the broader term. In analogy to the narrower term behavior discussed above, we expect to find a correlation between the relevance of the included features to the broader term and the validity of the directional similarity for the tested pair.

This expectation was confirmed by the analyzed data. We observed that ranks of the included features inside the vector of the broader term were usually consistently lower for invalid pairs than for valid ones. A typical example of such behavior is presented in Table 2. The table presents top features of the word 'election' that were included in the vectors of either the word 'vote' (a valid case), or the word 'reform' (an invalid case), or both. From the table we see that

- the features included only for the word 'vote' are ranked higher than the features included only for the word 'reform';
- the features that are included for both words are positioned higher in the vector of 'vote' as well.

This tendency remains the same for all the included features of the narrower term. We note that this distinctive behavior was observed for feature ranks but could not

Table 2. *Comparison between the feature ranks inside the broader term's vector for the valid pair 'election → vote' and the invalid pair 'election ↛ reform'. The upper part shows the top five features of 'election' that were included only in the vector of 'vote'; the middle part shows the top 5 features of 'election' that were included only in the vector of 'reform'; the lower part shows the top 10 features of 'election' that were included in both word vectors. The symbols '>…>' and '<…<' indicate the direction of the dependency relation between the term and a feature*

Top features of 'election'			
Rank	Feature	Rank in 'vote'	Rank in 'reform'
14	>nn>multi-candidate:n	2,927	
49	>nn>midterm:n	2,914	
52	<obj<rerun:v	1,038	
159	>mod>inconclusive:a	2,362	
189	<nn<annulment:n	507	
318	>nn>much-delayed:n		4,050
375	>nn>pluralist:n		4,093
403	>vrel>supervise:v		4,642
458	<nn<promise:n		1,189
599	>mod>thwarting:a		3,662
40	>appo>election:n	44	48
45	>conj>election:n	263	427
286	>nn>multiparty:n	2,928	4,054
533	>mod>forthcoming:a	2,328	3,333
567	<obj<campaign for:v	743	1,404
624	>mod>two-stage:a	2,577	3,675
631	<obj<supervise:v	1,107	1,867
642	>pnmod>scheduled:a	3,226	4,349
723	>mod>legislative:a	2,386	3,417
759	>mod>upcoming:a	2,599	3,692
Average rank of included features		*1,734*	*2,339*

be observed for the inconsistent pmi feature weights, thus confirming the conclusion made in Section 3.1 of the advantage in considering feature ranks rather than pmi weights.

We thus conclude that taking the position of included features in the broader term vector into account may improve the ability of an inclusion measure to identify valid similarity pairs.

3.3 Lower reliability of inclusion detection for short feature vectors

Short vectors, that is, vectors containing relatively few features, are typical for infrequent words in a corpus. With so few occurrences it is hard to collect meaningful statistics about their contexts. Short vectors constitute the majority of vectors in

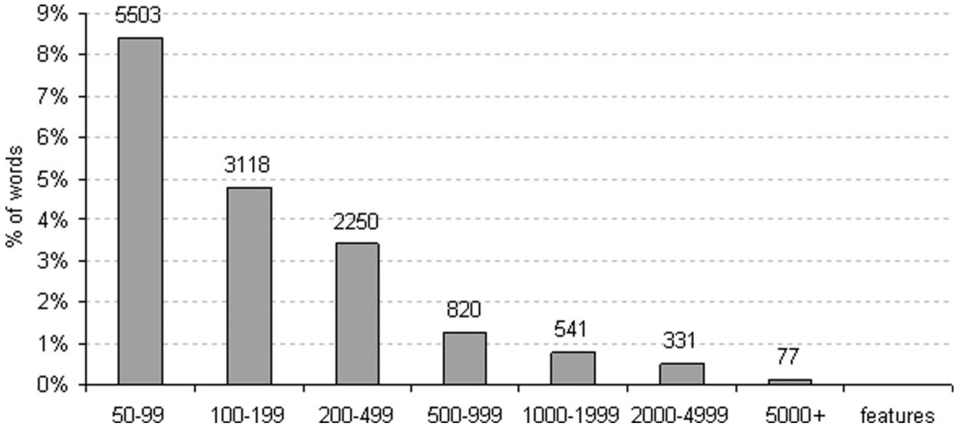

Fig. 3. Distribution of vector lengths for the terms in the analyzed corpus. The diagram shows the number of feature vectors of different lengths out of 12,640 vectors that have more than 50 features each, which makes up 19.3 per cent of all the vectors in our database. Vectors shorter than 50 features constitute the other 80.7 per cent (52,895 vectors).

our database (see Figure 3), as expected from the Zipfian distribution of words. We examined about a hundred of such vectors and saw that most of their features are rather common words, which are not good indicators for a specific term meaning. For example, the vector of the term ‘birthweight’ counts only twenty-eight features, where the top features relate to quite general words like ‘face:n’ and ‘type:n’.

Another reason of lower reliability of infrequent narrower terms for inclusion testing is that inclusion-based measures calculate the proportion between the number of included features and the number of tested ones. Thus, if the number of features in the vector of a narrower term is relatively small, it becomes relatively easy to cover most of them and get a high inclusion rate. Consequently, some arbitrary terms with short feature vectors are likely to have higher inclusion rates in general than similar terms having longer vectors.

In a similar manner, the number of included features is limited by the number of the features of the candidate broader term. When a candidate broader term has a short vector, candidate narrower terms with longer vectors cannot yield high inclusion ratios, while some (often arbitrary) terms with shorter vectors might get promoted. We thus expect inclusion detection to be less reliable for broader terms with short vectors. Table 3 presents an analysis of this expected behavior on our entailment pairs collection. The table shows that the longer the vector of a broader term is, the lower is the number of invalid pairs extracted for this term, and vice versa.

Overall, our analysis of valid and invalid pairs reaffirmed our hypothesis that short vectors, both for narrower and broader terms, seem to be unreliable for detecting valid similarity pairs based on feature inclusion. It confirms the well-known fact that data sparseness is a problem when dealing with statistical measures. We note, however, that this issue is not addressed by standard distributional similarity

Table 3. *Distribution of 1,067 valid and 2,705 invalid pairs in the entailment-annotated set according to the number of features in vectors of the candidate broader terms*

Vector length (features)	Number of pairs	Invalid pairs (%)	Valid pairs (%)
Less than 630	758	24	11
630–1,900	761	21	19
1,900–2,900	776	21	19
2,900–4,100	756	19	23
4,100–30,000	721	16	28
total	3,772	100	100

measures, such as *LIN*, *cosine*, etc., widely used in the common practice. We also note that ignoring this problem is, in our opinion, the main reason of the LEDIR algorithm's failure, since more frequent terms have more features (contexts), but they do not necessarily have a broader meaning.

4 A proposed inclusion measure

As identified in Section 3, a directional similarity measure based on feature inclusion should desirably satisfy several properties when measuring the degree of feature inclusion:

(1) Promoting the similarity scores if included features are highly relevant for the narrower term; the estimation of feature relevance may be better based on feature ranks rather than on feature weights.
(2) Promoting the similarity scores when included features are placed higher in the vector of the broader term as well.
(3) Demoting similarities for short feature vectors.

As a starting point for designing our new directional measure for lexical inference, we suggest utilizing a common IR evaluation method, namely *Average Precision* (AP), adapted to our problem. The rest of the section is organized as follows: first, we discuss the AP evaluation method and its applicability to our problem, then we introduce several adaptations for this method, resulting in our new proposed similarity measure.

In IR evaluation the task is to compare different retrieval systems: given a query, each system returns a list of documents, ranked according to their relevance to the query. A good system should thus:

- Retrieve many relevant documents (high recall).
- Retrieve as few irrelevant documents as possible (high precision).
- Place relevant documents at the top of the ranked list (high relevance).

The common IR measure that captures these properties is the *Average Precision (*AP*)* metric (Voorhees and Harman 1999):

$$AP = \frac{\sum_{r=1}^{N}[P(r) \cdot rel(r)]}{total\ number\ of\ relevant\ documents}$$

where r is the rank of a retrieved document amongst the N retrieved, $rel(r)$ is an indicator function for the relevance of that document, and $P(r)$ is precision at the given cut-off rank r. Thus, AP combines precision, relevance ranking and overall recall.

We adapted AP for measuring lexical similarity. In our case the features of the broader term are analogous to the set of all relevant documents, while the tested features correspond to retrieved documents. Included features thus correspond to relevant retrieved documents, yielding the following measure under our terminology:

$$AP(u \lesssim v) = \frac{\sum_{r=1}^{|F_u|}[P(r) \cdot rel(f_r)]}{|F_v|}$$

$$rel(f) = \begin{cases} 1, & \text{if } f \in F_v \\ 0, & \text{if } f \notin F_v \end{cases}$$

$$P(r) = \frac{|included\ features\ in\ ranks\ 1\ to\ r|}{r}$$

where F_x is a feature vector of term x and f_r is the feature at rank r in F_u.

This new intermediate feature inclusion measure partly addresses our desired properties. Its score increases with a larger number of included features (the core inclusion principle), while giving higher weight to highly ranked features of the narrower term (1st desired property).

To further meet the desired properties, we introduce two modifications to the above measure. First, we use the number of tested features $|F_u|$ for normalization instead of $|F_v|$. This captures better the notion of feature inclusion, which targets the proportion of included features relative to the tested ones.

Second, in the classical AP formula all relevant documents are of the same relevance to the input query. However, as suggested by our 2nd desired property, higher-ranking features of the broader term should have bigger influence on the final measure score. We thus reformulate $rel(f)$ to reflect the feature rank in a simple linear manner:

$$rel'(f) = \begin{cases} 1 - \frac{rank(f,F_v)}{|F_v|+1}, & \text{if } f \in F_v \\ 0, & \text{if } f \notin F_v \end{cases}$$

where $rank(f, F_v)$ is the rank of f in F_v. $rel'(f)$ provides a real number in the range (0,1) for relevance estimation, with higher values for higher ranking features. This modification also leads to some demotion of broader terms with short feature vectors, thus addressing also our third desired property: the ratio $rank(f, F_v)/(|F_v| + 1)$ will be higher for smaller denominators, thereby decreasing the value of $rel'(f)$ for all included features, even when they are ranked highly in the vector.

Incorporating our two modifications yields the *APinc* measure:

$$APinc(u \lesssim v) = \frac{\sum_{r=1}^{|F_u|}[P(r) \cdot rel'(f_r)]}{|F_u|}$$

Finally, we adopt the balancing approach in (Szpektor and Dagan 2008) to yield our proposed directional measure *balAPinc*:

$$balAPinc(u \lesssim v) = \sqrt{LIN(u,v) \cdot APinc(u \lesssim v)}$$

As explained in Section 2, balancing penalizes similarities containing infrequent narrower terms. Hence, *balAPinc* addresses also the third inclusion measure property. We note that balancing actually penalizes infrequent broader terms as well, since the *LIN* measure is symmetric and yields lower scores if one of the two compared terms has a short vector. We note, however, that this approach does not demote similarities when both the narrower term and the broader term vectors are short.

5 Evaluation and results

In this section we present experiments for evaluating our novel directional measure, comparing it to state-of-the-art measures. We first evaluate the ability of the various measures to detect the direction of similarity relation between candidate word pairs. We then perform application-based evaluations for the tested measures.

5.1 Learning corpus

In all of our distributional similarity measure implementations we use feature vectors created by parsing the Reuters RCV1 corpus with Minipar and taking the words related to each term through a dependency relation as its features (coupled with the relation name and the corresponding feature role as head or modifier, as in Lin 1998a).[2] We considered only terms that occur at least ten times in the corpus, and as features only words that occur at least twice.

In Table 4, anticipating the subsequent quantitative evaluation, we present top similarities extracted by the *LIN* and *balAPinc* measures for several common words, taken as the broader terms for *balAPinc*. The table shows that the directional *balAPinc* generates more accurate similarity lists than the symmetric *LIN* measure.

5.2 Detecting similarity direction

In our first evaluation, we would like to directly compare between the ability of various similarity measures to detect the direction of the similarity relation between candidate word pairs. To this end we used the same manually annotated collection

[2] Our implementations compute the top 1,000 terms similar to each given term.

Table 4. *Top similarities learned by* LIN *and* balAPinc *for exemplary common words*

word	*LIN*	*balAPinc*
food	meat, beverage, goods, medicine, drink, clothing, foodstuff, textile, fruit, feed, water, coffee, meal, tobacco, fuel, sugar, material, chemical, equipment, rice	food stuff, food product, food company, noodle, canned food, feed, salad dressing, bread, food aid, drink, ration, drinking water, wheat flour, grocery, beverage, snack, dairy product, hamburger, chocolate, sea food
vehicle	car, truck, bus, model, equipment, aircraft, engine, plane, boat, tank, helicopter, sedan, ship, train, weapon, machine, automobile, jet, goods, motorcycle	truck, jeep, pickup truck, sedan, minivan, car, motor vehicle, personnel carrier, passenger car, bus, motorcycle, lorry, wagon, motorbike, automobile, scooter, tractor, limousine, trailer
university	college, school, campus, hospital, church, municipality, institution, embassy, student, political party	highschool, college, seminary, campus, educational institution, faculty, high school, harvard, higher education, institute
airport	port, hospital, hub, hotel, city, suburb, terminal, station, centre, root	gatwick, airfield, heathrow, london heathrow, airbase, hub, expressway, terminal, port , air base

of valid and invalid lexical entailments (Zhitomirsky-Geffet and Dagan 2009), which was used in our preliminary analysis in Section 3. The collection was generated by manually judging a sample of 1,886 term pairs, which were produced by the symmetric distributional similarity measure of *LIN*. Each pair was assessed in both directions for lexical entailment, resulting in 1,067 valid and 2,705 invalid directional pairs.

5.2.1 Evaluation setting

To compare between different similarity measures, we provided the above pair collection as input to each measure. For each pair $\{u, v\}$, two candidate directional pairs were provided as input: '$u \rightarrow v$' and '$v \rightarrow u$'. Each measure provided a score for each candidate directional pair, quantifying its belief that the pair is valid. For each measure we then sorted the directional pairs by their scores and assessed the quality of the ordered list by calculating its Average Precision score, based on the gold-standard annotation.

As explained in Section 4, Average Precision (AP) combines precision, relevance ranking and overall recall. Thus, the AP score will be higher for measures that identify many valid rules by giving them a relatively high similarity score.

Table 5. *Results of the tested distributional measures on the directionality detection experiment*

Measure	AP	Precision	Recall	Pairs retrieved
LIN	0.41	0.32	0.88	2,922
JS	0.15	0.36	0.38	1,132
0.99-skew	0.21	0.36	0.39	1,184
WeedsPrec	0.43	0.32	0.91	3,060
balPrec	0.45	0.32	0.92	3,095
ClarkeDE	0.47	0.32	0.92	3,080
balAPinc	0.47	0.32	0.92	3,030

5.2.2 Results for distributional similarity measures

We evaluated our proposed *balAPinc* similarity measure as well as the following distributional similarity measures: *LIN*, *Skew divergence* with $\alpha = 0.99$ (*0.99-skew*), *Jensen-Shannon divergence* (*JS*), *WeedsPrecision* (*WeedsPrec*), *balPrecision* (*balPrec*) and *degree of entailment* (*ClarkeDE*) (see Section 2).

To obtain a clearer picture, in addition to the AP score we report the number of directional pairs retrieved by each of the measures as at least slightly similar (being included in the top 1,000 similarities), and the corresponding precision and recall values. Table 5 presents the results for the tested measures.

From the table we see that, overall, directional inclusion-based measures show better performance than symmetric ones (*LIN* and *JS*). In addition, our novel *balAPinc* measure outperforms all of the symmetric and directional baselines, excluding the *ClarkeDE* measure, which shows identical performance. The improvement over all other measures is statistically significant according to the two-sided Wilcoxon signed-rank test at the 0.01 level (Wilcoxon 1945).[3]

We note that the *ClarkeDE* and *balAPinc* measures are the only measures that comply with the second desired property presented in Section 3. Moreover, complying with this property is the only difference between the *ClarkeDE* measure and the *WeedsPrec* measure, which performs significantly worse.

We also note that in this experiment similarity measures were used to grade pre-given term pairs that were already found sufficiently similar by the symmetric measure of *LIN*. As explained earlier in Sections 2 and 4, this measure promotes pairs with rather common terms and demotes those with infrequent terms. Thus, in this evaluation the influence of reflecting the third desired property, which suggests to demote similarities of short feature vectors, cannot be adequately shown. We note that the third property is addressed by the *balAPinc* measure and is not addressed by the *ClarkeDE* measure. To illustrate the importance of following this property,

[3] We use the two-sided Wilcoxon signed-rank test at the 0.01 level for measuring statistical significance in all our evaluations in this section.

Table 6. *Top similarities learned by* balAPinc *and* ClarkeDE *for exemplary common words*

Word	*balAPinc*	*ClarkeDE*
jail	prison, prison term, custody, probation, imprisonment	bangladesh jute association, mountjoy, tanjung priok, bureau of prisons, ryszard wesolowski
money	monies, amount of money, sum of money, sum, wealth	eiichiro, up to *xxx* billion escudos, cretafund, close-season, kiyoshi
attack	bombing, raid, ambush, bombardment, assault	m. yoshikawa, rocketing, public interest, massing, ambush

Table 7. *Results of the* balAPinc *measure and the tested WordNet measures on the directionality detection experiment*

Measure	AP	Precision	Recall	Pairs retrieved
balAPinc	0.47	0.32	0.92	3,030
Jcn	0.40	0.31	0.76	2,597
Lch	0.44	0.29	0.89	3,313
Lesk	0.42	0.29	0.89	3,318
Lin	0.38	0.32	0.71	2,348
Path	0.44	0.29	0.89	3,313
Res	0.38	0.29	0.84	3,090
Vector	0.40	0.29	0.90	3,328
Wup	0.43	0.29	0.90	3,310

Table 6 presents the top five similarities produced by the two measures for exemplary common words. From the table we see that, unlike *balAPinc*, the *ClarkeDE* measure gives high scores to many incorrect similarities with infrequent terms. As will be shown later in our application-based evaluations, the *ClarkeDE* measure performs quite poorly when applied independently (rather than on the output of the *LIN* measure).

We conclude thus, that in this evaluation we showed the validity of the inclusion-based approach and our identified desired properties for recognizing directionality of the distributional similarity relation, as well as the advantage of our novel *balAPinc* measure, designed according to all of these properties.

5.2.3 Results for wordNet-based measures

We also compared our *balAPinc* measure to state-of-the-art WordNet-based measures, using their implementations from the WordNet::Similarity package (see Section 2.4). In Table 7 we report the results of the WordNet-based measures. The results for *balAPinc* are shown for comparison.

Table 8. *Results of combining the* balAPinc *measure and the* LIN *measure with each of the tested WordNet measures on the directionality detection experiment*

	Combined with *LIN*			Combined with *balAPinc*		
Measure	AP	Precision	Recall	AP	Precision	Recall
Jcn	0.50	0.29	0.98	0.52	0.29	0.99
Lch	0.50	0.29	0.98	0.53	0.29	0.99
Lesk	0.50	0.29	0.98	0.52	0.29	0.99
Lin	0.49	0.31	0.93	0.52	0.31	0.96
Path	0.50	0.29	0.98	0.53	0.29	0.99
Res	0.47	0.30	0.96	0.50	0.30	0.97
Vector	0.48	0.29	0.98	0.50	0.29	0.99
Wup	0.50	0.29	0.98	0.52	0.29	0.99

The table shows that the results provided by the WordNet-based measures are lower than those of *balAPinc* under all evaluation metrics. In terms of the AP score, the advantage of *balAPinc* over the *Jcn*, *Lin*, *Res*, and *Vector* measures is statistically significant. Thus, we may conclude that our proposed distributional measure performs not worse and sometimes even better than state-of-the-art WordNet-based measures, not only in terms of relevance ranking (AP), but also when directly measuring recall and precision. This is an interesting result, since distributional measures are typically known to yield high recall while having low precision, while WordNet-based similarities are considered precise, but suffering from insufficient recall (Mirkin *et al.* 2009).

In recent research, Agirre et al. (2009) showed that combination with symmetric distributional similarity improves the performance of WordNet-based methods. In a further evaluation we checked whether this conclusion can be confirmed for our proposed directional measure as well. For this purpose we combined the output of either *balAPinc* or the symmetric *LIN* measure with each of the WordNet-based methods.

We combined the output of any two measures by first ranking the pairs based on each measure alone. Then, each pair was re-ranked according to its average rank in the two lists. If a pair was not present in one of the lists, it was treated as if placed at the end of that list. We chose to combine pairs based on ranks rather than scores because many WordNet-based measures assign unnormalized scores, which are not comparable to distributional similarity scores. Table 8 summarizes the results of this evaluation.

As can be seen from the table, combining distributional similarity with WordNet-based measures yielded considerably better results than any of the two measures alone. These results are statistically significant compared to each of the combined measures alone. The considerable improvement in recall is achieved without any significant drop in precision. This shows that distributional and WordNet-based measures are complementary. We also note that combining WordNet measures with the directional *balAPinc* consistently outperforms combining them with *LIN* in terms of Average Precision.

5.3 Manual vs. application-based evaluation

The results presented in the above evaluation show that *balAPinc*, which is designed to follow all the desired properties presented in Sections 3 and 4, achieves best performance in detecting directional similarity. In addition, the performance improvement achieved when combining *balAPinc* with WordNet-based measures is higher than when combining a symmetric measure with WordNet-based measures.

Still, such evaluation does not fully reflect the quality of the tested similarity measures. The actual utility of a lexical semantic resource can be measured best in an instance-based evaluation, examining the correctness of applying similarity pairs for lexical inference instead of directly assessing their correctness (Szpektor, Shnarch and Dagan 2007; Mirkin *et al.* 2009). Hence, we perform additional application-based evaluations of distributional measures, choosing our evaluation framework to be lexical-expansion. Lexical expansion is widely employed to overcome lexical variability in applications like IR, IE, and QA (e.g. Xu and Croft 1996; Mandala *et al.* 1999) and it is one of the prominent utilizations of distributional similarity for lexical inference. The expansion task is to augment a given textual input (e.g. a query) with *expansion terms*, terms of meanings similar to the input terms, which are produced in our case by the distributional similarity measures.

We tested lexical expansion within two different application settings – ACE and Unsupervised keyword-based TC. In the following subsections we describe each of these two evaluations. We underline that in both evaluations the applications are used as a comparative tool to evaluate the quality of the measures; the results are not claimed to be in any way optimal for the given task, since we were not employing an optimal comprehensive system for each application.

5.4 Evaluation within ACE

5.4.1 Evaluation setting

As a typical lexical expansion task we used the ACE 2005 events dataset.[4] This standard IE dataset contains thirty-three event types, such as *Attack*, *Divorce* and *Law Suit*, with all event mentions annotated in the corpus. It was previously successfully utilized for evaluating distributional similarity methods for recognizing lexical–syntactic entailment relations (Szpektor and Dagan 2008; Szpektor *et al.* 2008; Szpektor and Dagan 2009). For our lexical expansion evaluation we considered the first IE subtask: finding sentences that contain mentions of a target event.

For each event we manually selected a few typical terms for the event, denoted *seeds*. These terms (four on average) were selected from the textual description of the event definition in the ACE guidelines. The seeds serve as a baseline query for retrieving sentences containing the event occurrences. For example, the words ‘meet:n’, ‘meet:v’, ‘meeting:n’, and ‘meeter:n’ were used as the seeds for the *Meet* event.

[4] http://projects.ldc.upenn.edu/ace/, training part.

Table 9. *MAP scores of the tested measures on the ACE experiment*

	MAP	
Measure	22 events	9 events subset
LIN	0.04	0.08
JS	0.02	0.03
0.99-skew	0.02	0.04
WeedsPrec	0.02	0.04
ClarkeDE	0.02	0.05
balPrec	0.12	0.26
balAPinc	0.16	0.34

To evaluate each similarity measure, the terms found similar to each of the event's seeds ('$u \lesssim seed$') by that measure were taken as expansion terms. All sentences containing the expansion terms were retrieved. To measure the sole contribution of the applied similarity list, we removed from the retrieved list all the sentences that contain at least one seed. For 11 out of thirty-three events, less than ten sentences were retrieved in this manner, providing insufficient statistics for comparison. These events were excluded from our evaluation, leaving twenty-two events. For these twenty-two events there are overall 3,789 positive annotated instances in the corpus (excluding those containing one of the seeds). This amounts to an average of 172 positive instances per event, within more than 15,700 sentences in the corpus.

The expansion quality for each event by a given measure was calculated by scoring each retrieved sentence with the sum of the similarity scores of the expansion terms it contains. For each event, the ranked list of retrieved sentences was generated and the list quality was assessed by the Average Precision (AP) evaluation measure (based on the ACE gold-standard annotation). We report Mean Average Precision (MAP) over all tested events for each tested measure.

5.4.2 Results

We evaluated *balAPinc* on the ACE setup as well as the following distributional similarity measures: *LIN*, *Skew divergence* with $\alpha = 0.99$ (*0.99-skew*), *Jensen-Shannon divergence* (*JS*), *WeedsPrecision* (*WeedsPrec*), *degree of entailment* (*ClarkeDE*) and *balPrecision* (*balPrec*). Table 9 presents the results for the tested measures. We report the MAP score calculated for all twenty-two tested events. In addition, we report MAP value for a subset of nine events for which at least one of the evaluated measures achieved AP value of at least 0.1; these events are those for which expansion by distributional similarity methods is reasonably effective (for the remaining events purely distributional methods may not be suitable for lexical expansion).

The results show that the α-*skew* , the *WeedsPrec* and the *ClarkeDE* directional measures are not competitive for this application setting, showing lower performance than the symmetric measure of *LIN*. On the other hand, *balPrec* yields considerably better results. Our main result is that *balAPinc* is the best-performing measure,

Table 10. *MAP scores of the* balAPinc, balPrec *and* LIN *measures on the ACE experiment without excluding the sentences with seed terms*

	MAP	
Measure	22 events	9 events subset
Seed terms only	0.28	0.33
LIN	0.31	0.40
balPrec	0.38	0.53
balAPinc	0.40	0.57

showing statistically significant improvement over all other measures according to the two-sided Wilcoxon signed-rank test at the 0.01 level. These results support our hypothesis that carefully designed directional approaches should produce improved similarity lists.

We next discuss these results in view of the desired properties of an inclusion measure, as defined in Sections 3 and 4. First, we note the utility of the balancing approach. It was introduced in order to demote similarities of short feature vectors, as suggested by the 3rd desired property. We remind that the *balPrec* measure is actually the balanced version of *WeedsPrec*. Thus, its considerable advantage over the *WeedsPrec* measure is achieved due to satisfying the third property by filtering out infrequent terms. Indeed, similarity lists produced by the *WeedsPrec* measure are full of arbitrary infrequent terms and thus appear as completely useless. For example, the top similarities produced by this measure for the seed term ‘meeting’ are ‘yield-pct’, ‘wheat tender - sri lanka’, ‘washington research group’, ‘walheim’, ‘vienna museum for applied arts’, ‘value-usda’ etc. (the same holds for the *ClarkeDE* measure, as shown earlier in Table 6).

We also see improved results for *balAPinc* over *balPrec*, which shows that meeting our other desired properties improves performance as well: (a) *balPrec* does not reflect different relevance of included features to the broader term and thus does not satisfy our 2nd desired property, while *balAPinc* does; (b) based on *WeedsPrec*, *balPrec* uses pmi for feature weights to reflect feature relevance to the narrower term, while *balAPinc* uses feature ranks, which seems advantageous from the perspective of satisfying the first desired property (as discussed in Section 3.1).

As an additional analysis of the results, Table 10 reports the results of the ACE experiment without excluding sentences that contain seed terms. The configurations included in the table are: using only seed terms without any lexical expansion, the two best-performing directional measures – *balAPinc* and *balPrec*, and the best-performing state-of-the-art symmetric *LIN* measure. From the table we see that directional measures introduce considerably higher improvement as compared to the symmetric approach. Furthermore, our proposed directional *balAPinc* measure outperforms the other baselines in this setting as well, and its improvement is statistically significant compared to all other configurations.

Finally, we performed ablation tests on *balAPinc* by running the ACE experiment over intermediate measures constructed by removing one of the three improvements in *balAPinc* over the baseline *AP* measure (Section 4) at a time. The tests showed that each of the improvements introduces a statistically significant improvement in results.

To conclude, the results in this experiment support our suggestion to satisfy each of the desired inclusion measure properties presented in Sections 3 and 4 when designing distributional similarity measures.

5.4.3 Error analysis

To better understand the quality of the similarity lists generated by directional and symmetric measures, we performed error analysis on the results of the symmetric *LIN* measure and the directional *balAPinc* measure. To that end, we selected from each event the two highest scoring retrieved sentences that did not contain an event mention (*false positives*). We analyzed all matched expansions in these sampled sentences. Alltogether, over 600 expansion matches were evaluated for each measure.

In our analysis we aimed to identify in which cases false positive sentences were retrieved by clearly invalid expansions (e.g. '*war* $\lesssim$ *demonstration*', '*work* $\lesssim$ *marry*') versus being retrieved by potentially valid expansions, which yielded an irrelevant sentence in a specific context.

Such potentially valid lexical expansions were further split into two subtypes. The first one corresponds to valid entailment relations, both substitutable (Geffet and Dagan 2005) and nonsubstitutable (Mirkin *et al.* 2009). These are high quality expansions, where the meaning of the broader term can be directly implied from the meaning of the narrower one, e.g. '*sit-in* $\lesssim$ *demonstration*'.

The other type of valid expansions refers to strong context relatedness, i.e. there is no entailment relation between the two terms but the context of the narrower term is quite prominently connected with the context of the broader one, e.g. '*solidarity* $\lesssim$ *demonstration*'. These are somewhat less accurate expansions. Nevertheless, they could be successfully utilized both to detect mentions of the broader term's meaning and to support decisions involving additional similarity-based expansions.

We illustrate the two valid expansion types via the text '*…Jordanian lawyers staged a* ***sit-in*** *at the main court house after being forcibly blocked by* ***riot police*** *from* ***marching*** *towards the Iraqi* ***embassy*** *to show their* ***solidarity****…*', which was extracted as mentioning the *Demonstration* event. The matching expansions consisted of three entailing terms, 'sit-in', 'riot' and 'marching', and three context-related terms 'police', 'embassy' and 'solidarity' ('riot' and 'police' were matched separately). In this example, context-related terms increased the matching score of the text, reflecting the level of certainty that it indeed contains a mention of the target event.

Table 11 presents the distribution of valid and invalid matched expansions in the false-positive sentences.

Invalid expansions. We see in the table that the majority of false-positive cases in our sample were caused by completely invalid expansions. For example, the sentence

Table 11. *The distribution of valid and invalid expansions, whose matching resulted in false-positive errors in the ACE task. Average scores of the corresponding expansion types are given in parentheses*

	Valid expansions		
Measure	Entailment	Strong context relatedness	Invalid expansions
LIN	4.3% (0.059)	6.5% (0.052)	89.2% (0.052)
balAPinc	12.3% (0.044)	7.3% (0.020)	80.4% (0.013)

'*We're **planning** on **going** July 4th week – what better **time** to be in Boston … **stay** with **friends** living in Boston …*' was retrieved as mentioning the *Phone-Write* event. This means that if it were possible to filter out the invalid expansions, manually or using some automatic techniques, this would yield much better performance. We note that often weight thresholds, either set up manually or tuned using some positive and negative examples, are used to filter out invalid expansions. Since in our evaluation we aimed at comparing different similarity measures rather than obtaining the optimal utility of different lexical resources for specific tasks, we did not apply such thresholds. This also explains the considerable amount of completely invalid expansions displayed in Table 11: sentences with many matches of invalid expansions got higher scores than those with one or two valid matches and thus were retrieved in our experiment.

From the average scores presented in Table 11 we see that scores produced by *balAPinc* allow much better distinction between valid and invalid expansions. The scores assigned by the *balAPinc* measure to valid entailments are noticeably higher than the scores for the context-related expansions, which in turn are much higher than those for completely invalid expansions. The scores assigned by the *LIN* measure do not show such sharp distinctions at all.

Valid expansions: passing reference. We observed that many flase-positive decisions were made using valid similarities. One of the main reasons is passing references, when a valid expansion term occurs in a text fragment, but it is not the focal point of a sentence. For example, a sentence containing the following fragment '*Call Carnival Wedding Dept. at 1 800 933-4968*' was mistakenly extracted as mentioning the *Marry* event by applying the similarity pair '*wedding* $\lesssim$ *marriage*'. We conclude that developing techniques for more accurate matching of expansion terms in order to avoid passing reference would improve performance.

Valid expansions: semantic ambiguity. Another typical issue observed in our analysis is that many valid expansions cause errors due to their semantic ambiguity. In some cases errors were caused by matching a wrong sense of an expansion term in the text, e.g. incorrectly extracting the sentence '*Everest is the highest summit in the world*' as mentioning the *Meet* event, as a result of applying the similarity pair '*summit* $\lesssim$ *meeting*'. In other cases errors occurred due to ambiguous seeds, e.g. when looking for the *End-Position* event mentions using pairs like '*shoot* $\lesssim$ *fire*'. Ambiguity of these two types was found in 35 per cent of matched valid expansions

Table 12. *Performance of the* balPrec *and* balAPinc *measures on the ACE experiment when using for inclusion testing different top-n percents of the context features of candidate narrower terms*

	MAP			
Measure	Top 25%	Top 50%	Top 75%	100%
balPrec	0.17	0.15	0.13	0.12
balAPinc	0.18	0.17	0.17	0.16

generated by the *balAPinc* measure (within our sampled false-positive sentences) and in 23 per cent of the matched valid expansions produced by the measure of *LIN*. We conclude that solving the problem of semantic ambiguity either explicitly via classical word sense disambiguation (WSD) techniques or implicitly via contextual preferences (e.g. Szpektor *et al.* 2008) would considerably improve the performance of similarity-based expansion.

5.4.4 Analysis of feature vector truncation

As discussed in Section 3.1, the distinguishing ability of inclusion measures should be higher when applying inclusion testing only to the top features of the candidate narrower term. Table 12 presents the performance of the *balPrec* and *balAPinc* measures when applied to such truncated vectors of the narrower term, which include only the top-*n* per cent of the ranked feature list, over the ACE experiment for all twenty-two tested events.

From the table we see that *balAPinc* consistently shows better performance. More generally, these results show that these inclusion measures indeed perform better when using truncated vectors of the candidate narrower term. When observing the optimal threshold in retrospect, we see that with top 25 per cent cut-off the results achieved by the two measures become rather close, but *balAPinc* still enjoys the benefit of being much more stable with respect to the cut-off point selection. Since in unsupervised settings we do not expect optimal tuning of this parameter, such robustness is indeed an advantage.

We also examined the highest scoring terms produced for the same seeds using vectors truncated at different points. We noticed that these lists typically look considerably different. An example for the seed verb ‘die’ is presented in Table 13. We thus conclude that developing techniques to identify appropriate feature vector cut-off points will allow inclusion measures to produce more precise similarity lists and is a worthwhile pursuit for future work.

5.5 Evaluation within keyword-based text categorization setting

5.5.1 Evaluation setting

As an additional application setting for lexical expansion, we evaluated the various similarity measures on a TC dataset. To that end we used an available keyword-based

Table 13. *Top similarities learnt by the* balAPinc *measure for a seed verb 'die' for different truncation points of the narrower term vector*

Top 25%	Top 50%	Top 75%	100%
perish	perish	perish	perish
drown	riot	drown	sleep
breathe	drown	riot	drown
mourn	starve	sleep	open fire
come home	open fire	emigrate	emigrate
succumb	sleep	open fire	disappear

TC system (Barak, Dagan and Shnarch 2009). Such methods aim at topical categorization of documents based on sets of terms, without requiring a supervised training set of labeled documents (McCallum and Nigam 1999; Ko and Seo 2004; Liu *et al.* 2004). Generally speaking, such systems operate in two phases: (i) a setup phase, in which a set of characteristic terms for the category is assembled, constituting the category's feature vector; (ii) a classification phase, in which the term-based feature vector of a classified document is compared with the feature vectors of all categories.

In our evaluation, as in typical query expansion, category names were taken as seeds and expanded by distributional similarity to form category vectors. Document and category vectors were then compared with the cosine similarity measure, producing the categorization score for each document with respect to each category. Document features consist of POS-tagged lemmas of single words and bigrams, limited to nouns, verbs, adverbs, and adjectives, with term frequency as the feature value.

We note that for the eventual classification phase different keyword-based TC systems employ different strategies – some assign a document only to the category with the highest score, others select the top-*n* per cent of the documents classified for each category. Often, an additional bootstrapping step was conducted, feeding a standard supervised classifier with the classifications produced by the keyword-based categorization method. We note that in our case the final classification phase is not the focus of our evaluation. Since our goal is to comparatively evaluate the performance of various similarity measures in expanding the category name, we consider the ranked lists of documents assigned to each category. We thus view our categorization setting as similar to query expansion in IR and report the results in terms of Average Precision for the ranked lists of categorized documents, yielding the same evaluation procedure as in Section 5.4.

For our evaluation we used the Reuters-10 corpus, constructed of the 10 most frequent categories of the Reuters-21578 collection.[5] We used the Apte split of the Reuters-21578 collection, which is often used in TC tasks. The complete collection contains 12,902 documents for ninety categories, where the top ten categories include

[5] Available at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

Table 14. *MAP scores of the tested measures on the TC experiment*

Measure	MAP
LIN	0.47
JS	0.39
0.99-skew	0.41
WeedsPrec	0.41
ClarkeDE	0.42
balPrec	0.64
balAPinc	0.66

Table 15. *Performance of the* balPrec *and* balAPinc *measures on the TC experiment when using for inclusion testing different top-n percents of the features of candidate narrower terms*

	MAP			
Measure	Top 25%	Top 50%	Top 75%	100%
balPrec	0.66	0.65	0.64	0.64
balAPinc	0.69	0.68	0.67	0.66

9,296 documents. The documents are divided in advance to train (70 per cent) and test (30 per cent) parts. Since in our settings no training was performed, we only used the test part for our comparative evaluation.

5.5.2 Results

Table 14 presents the MAP values obtained for the tested measures on the TC experiment. The table shows that, as in the ACE setting, *balAPinc* is the best-performing measure.

Table 15 presents the performance of the *balPrec* and *balAPinc* measures when applied to truncated vectors of candidate narrower terms, analogously to Table 12. The results affirm the finding for the ACE evaluation, that inclusion measures benefit from using truncated vectors of candidate narrower terms. We also see that in this evaluation *balAPinc* preserved its advantage over the *balPrec* measure for the optimal cut-off as well.

Overall, summarizing the evaluations, we conclude that satisfying all the desired properties indeed allowed our *balAPinc* measure to show best results in detecting the direction of similarity relation and to achieve the best performance within both application-oriented tasks, as well as to show higher robustness with respect to the truncation level of the vectors of the narrower terms.

6 Conclusions and future work

This paper advocates the use of directional similarity measures for lexical inference and expansion. In particular, we focus on directional measures based on distributional inclusion scores for feature vector pairs. Based on a thorough analysis, we identified desired properties for an inclusion-based directional similarity measure. Showing that state-of-the-art directional measures do not satisfy all these properties, we designed a novel directional measure, *balAPinc*, based on the standard IR Average Precision evaluation measure, which addresses all of the desired properties.

We compared our proposed measure to other state-of-the-art symmetric and directional measures on detecting the direction of entailment relations between terms, under which *balAPinc* showed the best results. In addition, we tested our measure as well as the other baseline state-of-the-art measures on lexical expansion for two application settings, IE and keyword-based TC. In both settings, our carefully designed directional measure performed significantly better than the other tested measures. We also observed that only one of the previous directional measures, *balPrec*, achieved competitive results, but our measure was more robust. In general, our experiments show the advantage of directional measures that satisfy the proposed desired properties over both typical symmetric measures and directional measures that do not follow these properties.

In a future work, we plan to explore other measures that meet our suggested properties. In addition, we found that truncated feature vectors for narrower terms perform better than the complete vectors. We thus plan to investigate automatic techniques for choosing the optimal truncation point. Also, many correctly learned rules were incorrectly applied in invalid contexts. In a future work, we aim at automatically learning each rule's context model as well, to improve rule application (Szpektor *et al.* 2008). Finally, it would be interesting to apply our novel directional measure as a kernel function for supervised learning like in Bloehdorn and Moschitti (2007) or in Basili, Cammisa and Moschitti (2006), as well as to apply it in other related fields, such as simulating human associations and psychological distance and compare it with measures designed for this kind of tasks (e.g. Michelbacher, Evert and Schutze 2007).

Acknowledgements

This work was partially supported by the NEGEV project (www.negev-initiative.org), the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886, the Israel Science Foundation grant 1112/08, and the FBK-irst/Bar-Ilan University collaboration under the ITCH Project. We thank Roy Bar-Haim and Libby Barak for providing their systems for the evaluations, and Benjamin Zhitomirsky for his help with the WordNet::Similarity package. Finally, we thank the anonymous reviewers for their useful comments and suggestions.

References

Agirre, E., Enrique A., Keith H., Jana K., Marius P., and Aitor S. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *NAACL HLT '09:*

Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, CO, USA: Association for Computational Linguistics, pp. 19–27.

Banerjee, S., and Ted, P. 2002. An adapted lesk algorithm for word sense disambiguation using WordNet. In *CICLing*, Mexico City, ME, pp. 136–145.

Barak, L., Dagan, I., and Shnarch E. 2009. Text categorization from category name via lexical reference. In *Proceedings of NAACL HLT 2009: Short Papers*, pp. 33–36, Boulder, Colorado, USA.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., and Szpektor, I. 2006. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge*, Venice, Italy, pp. 33–36.

Basili, R., Cammisa, M., and Moschitti, A. 2006. A semantic kernel to classify texts with very few training examples. *Informatica (Slovenia)* **30**(2): 163–172.

Bhagat, R., Pantel, P., and Hovy, E. 2007. LEDIR: an unsupervised algorithm for learning directionality of inference rules. In *Proceedings of EMNLP-CoNLL*, Prague, Czech Republic.

Bloehdorn, S., and Moschitti, A. 2007. Structure and semantics for expressive text kernels. In *CIKM*, Lisbon, Portugal, pp. 861–864.

Budanitsky, A., and Hirst, G. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics* **32**(1): 13–47.

Caraballo, S. A. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Thirty-Seventh Annual Meeting of the ACL*, College Park, MD, USA.

Chen, S. F., and Goodman, J. 1996. An empirical study of smoothing techniques for language modeling. In *ACL*, Santa Cruz, CA, USA, pp. 310–318.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics* **16**(1): 22–29.

Clarke, D. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pp. 112–119. Athens, Greece: Association for Computational Linguistics.

Dagan, I., Glickman, O., and Magnini, M. 2006. The PASCAL recognising textual entailment challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, F. d'Alché-Buc (eds.), Machine learning challegues. *Lecture Notes in Computer Science*, vol. 3944, pp. 177–190. Springer.

Dagan, I., Lee, L., and Pereira, F. 1999. Similarity-based models of cooccurrence probabilities. *Machine Learning* **34**(1–3): 43–69.

Fellbaum, C. 1998. *WordNet – An Electronic Lexical Database*. MIT Press.

Gasperin, C., Gamallo, P., Agustini, A., Lopes, G., and de Lima, V. 2001. Using syntactic contexts for measuring word similarity. In *In the Workshop on Semantic Knowledge Acquisition and Categorisation (ESSLI 2001)*, Helsinki, Finland.

Gauch, S., Wang, J., and Rachakonda, S. M. 1999. A corpus analysis approach for automatic query expansion and its extension to multiple databases. *ACM Transactions on Information Systems (TOIS)* **17**(3): 250–269.

Geffet, M., and Dagan, I. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of ACL*, Michigan, USA.

Harabagiu, S., and Hickl, A. 2006. Methods for using textual entailment in open-domain question answering. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 905–912, Morristown, NJ.

Harabagiu, S. M., Hickl, A., and Lacatusu, V. F. 2007. Satisfying information needs with multi-document summaries. *Information Processing and Management* **43**(6): 1619–1642.

Hindle, D. 1990. Noun classification from predicate-argument structures. In *Proceedings of ACL*, Pittsburgh, Pennsylvania, USA.

Jiang, J. J., and Conrath, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Tapei, Taiwan, pp. 19–33.

Jing, Y., and Croft, W. B. 1994. An association thesaurus for information retrieval. In *Proceedings of RIAO 94*, Rockefeller University, NY, USA, pp. 146–160.

Jones, M. N., and Mewhort, D. J. K. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review* **114**(1): 1–37.

Ko, Y., and Seo, J. 2004. Learning with unlabeled data for text categorization using a bootstrapping and a feature projection technique. In *ACL 2004*, Barcelona, Spain, pp. 255–262.

Leacock, C., and Chodorow, M. 1998. *WordNet: An Electronic Lexical Database – Combining Local Context and WordNet Similarity for Word Sense Identification, in Wordnet: An Electronic Lexical Database*, chap. 11, pp. 265–283. MIT Press.

Lee, L. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, USA, pp. 25–32.

Lin, D. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*, Montreal, Quebec, Canada.

Lin, D. 1998b. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on Evaluation of Parsing Systems at LREC 1998*, Granada, Spain.

Lin, D. 1998c. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, Madison, WI, USA.

Lin, D., and Pantel, P. 2001. DIRT – discovery of inference rules from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*, San Francisco, CA, USA, pp. 323–328.

Liu, B., Li, X., Lee, W. S., and Yu, P. S. 2004. Text classification by labeling words. In *AAAI-2004*, San Jose, CA, USA, pp. 425–430.

Lloret, E., Ferra'ndez, O., Mun oz, R., and Palomar, M. 2008. A text summarization approach under the influence of textual entailment. In B. Sharp and M. Zock (eds.), *NLPCS*, pp. 22–31. INSTICC.

Mandala, R., Tokunaga, T., and Tanaka, T. 1999. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of SIGIR*, Berkeley, CA, USA.

McCallum, A., and Nigam, K. 1999. Text classification by bootstrapping with keywords, em and shrinkage. In *ACL '99 Workshop for Unsupervised Learning in Natural Language Processing*, pp. 52–58, College Park, Maryland, USA.

Michelbacher, L., Evert, S., and Schutze, H. 2007. Asymmetric association measures. In *Proceedings of RANLP*, Borovets, Bulgaria.

Mirkin, S., Dagan, I., and Shnarch, E. 2009a. Evaluating the inferential utility of lexical-semantic resources. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece: Association for Computational Linguistics, pp. 558–566.

Mirkin, S., Specia, L., Cancedda, N., Dagan, I., Dymetman, M., and Szpektor, I. 2009b. Source-language entailment modeling for translating unknown terms. In *Proceedings of ACL-IJCNLP*. Singapore.

Pantel, P., and Ravichandran, D. 2004. Automatically labeling semantic classes. In *Proceedings of Human Language Technology/North American chapter of the Association for Computational Linguistics (HLT/NAACL-04)*, pp. 321–328, Boston, MA, USA.

Patwardhan, S. 2003. *Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness*. Master's thesis, Palo Alto, CA, USA: University of Minnesota.

Pedersen, T., Patwardhan, S., and Michelizzi, J. 2004. Wordnet: Similarity – measuring the relatedness of concepts. In *AAAI*, pp. 1024–1025, San Jose, CA, USA.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, pp. 448–453, San Francisco, CA: Morgan Kaufmann Publishers Inc.

Roberts, M. A. J., and Chater, N. 2008. Using statistical smoothing to estimate the psycholinguistic acceptability of novel phrases. *Behavior Research Methods* **40**(1): 84–93.

Ruge, G. 1992. Experiments on linguistically-based term associations. *Information Processing and Management* **28**(3): 317–332.

Sahlgren, M., Holst, A., and Kanerva, P. 2008. Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci'08)*, Washington, DC, USA, pp. 1300–1305.

Salton, G., and McGill (eds.) 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill.

Szpektor, I., and Dagan, I. 2008. Learning entailment rules for unary templates. In *Proceedings of COLING*, Manchester, UK.

Szpektor, I., and Dagan, I. 2009. Augmenting WordNet-based inference with argument mapping. In *Proceedings of the 2009 Workshop on Applied Textual Inference*, Singapore.

Szpektor, I., Dagan, I., Bar-Haim, R., and Goldberger, J. 2008. Contextual preferences. In *Proceedings of ACL-08: HLT*, Columbus, OH, USA, pp. 683–691.

Szpektor, I., Shnarch, E., and Dagan, I. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of ACL 2007*, Prague, Czech Republic.

Turney, P. D. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, pp. 491–502, London, UK: Springer-Verlag.

Tversky, A. 1977. Features of similarity. *Psychological Review* **84**: 327–352.

Voorhees, E. M., and Harman, D. K., (eds.) 1999. *The Seventh Text REtrieval Conference (TREC-7)*, vol. 7. NIST.

Weeds, J., and Weir, D. 2003. A general framework for distributional similarity. In *Proceedings of EMNLP*, Sapporo, Japan.

Weeds, J., Weir, D., and McCarthy, D. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of COLING*, Geneva, Switzerland.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* **1**: 80–83.

Wu, Z. and Palmer, M. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, Las Cruces, NM, USA, pp. 133–138.

Xu, J., and Croft, W. B. 1996. Query expansion using local and global document analysis. In *Proceedings of SIGIR*, Zurich, Switzerland.

Zazo, Á. F., Figuerola, C. G., Alonso Berrocal, J. L., and Rodríguez, E. 2005. Reformulation of queries using similarity thesauri. *Information Processing and Management* **41**(5): 1163–1173.

Zhitomirsky-Geffet, M., and Dagan, I. 2009. Bootstrapping distributional feature vector quality. *Journal of Computational Linguistics* **35**(3).