Experiments on the difference between semantic similarity and relatedness

Peter Kolb
Universität Potsdam
Potsdam, Germany
kolb@ling.uni-potsdam.de

Abstract

Recent work has pointed out the difference between the concepts of semantic similarity and semantic relatedness. Importantly, some NLP applications depend on measures of semantic similarity, while others work better with measures of semantic relatedness. It has also been observed that methods of computing similarity measures from text corpora produce word spaces that are biased towards either semantic similarity or relatedness. Despite these findings, there has been little work that evaluates the effect of various techniques and parameter settings in the word space construction from corpora. The present paper experimentally investigates how the choice of context, corpus preprocessing and size, and dimension reduction techniques like singular value decomposition and frequency cutoffs influence the semantic properties of the resulting word spaces.

1 Introduction

A growing number of applications in natural language processing rely on knowledge about the semantic similarity between words. These similarities are used for example in ontology learning (Cimiano et al., 2005), information retrieval (Müller et al., 2007), and word sense disambiguation (Patwardhan et al., 2007).

One has to differentiate between semantic "similarity" and semantic "relatedness" (Budanitsky and Hirst, 2006). The first is a narrower concept that holds between lexical items having a similar meaning, like *palm* and *tree*. It is usually defined via the lexical relations of synonymy and hyponymy. (Geffet and Dagan, 2005) require that semantically similar words can be substituted for each other in context, which must not be true for semantically related words.

The broader concept semantic relatedness holds between lexical items that are connected by any kind of lexical or functional association. Dissimilar words can be semantically related, e.g. via relations like meronymy (*palm – leaf*), or when they belong to the same semantic field (*palm – coconut*). (Turney, 2008) seems to equate "related" with "associated" and defines: "Two words are associated when they tend to co-occur (*doctor* and *hospital*)".

Unfortunately, measures of semantic similarity and relatedness rely on hand-crafted lexical resources like WordNet, which are not available for many languages and have limited coverage, particularly in specialized domains. Therefore, (Kilgarriff, 2003) and others have argued for using "distributional similarity" as a proxy for semantic similarity. Distributional semantics is based on the assumption that words with similar meaning occur in similar contexts (Harris, 1968). Several successful methods to compute the distributional similarity of words from text corpora have been proposed, including (Landauer and Dumais, 1997), (Grefenstette, 1994), and (Sahlgren, 2001).

(Budanitsky and Hirst, 2006) emphasize the difference between semantic and distributional similarity. Methods that measure the similarity of the distributional behaviour of words do not take into account the different senses a word has, and therefore mix up the similar words for all the word senses. While semantic similarity is a relation between concepts, distributional similarity is a relation between words.

Finally, (Mohammad and Hirst, 2005) differentiate between distributional relatedness and distributional similarity. Two words are distributionally similar if they have many common co-occurring words in the same syntactic relations. By contrast, distributional measures that use a bag-of-words

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 81–88

context capture distributional relatedness. (Kilgarriff and Yallop, 2000) call these two variants "tight" and "loose" word similarities. (Sahlgren, 2006) comes to the conclusion that word spaces based on direct co-occurrences capture relatedness, while spaces that are based on indirect or second-order co-occurrences capture similarity.

The difference between semantic similarity and relatedness is not only of theoretical interest. In fact some NLP applications require measures of semantic similarity, while others perform better with semantic relatedness. (Sahlgren and Karlgren, 2008) give an example from the area of text mining. For the analysis of opinions in blogs and discussion forums it is useful to automatically detect synonyms and spelling variants for an interesting term like *recommend*, thereby discovering terms that are used similarly in the given sublanguage, for example *love*, *lurve*, *looove* and *recomend*. To solve this task, measures of semantic similarity are much better suited. On the other hand, to find out what people associate with a target word like *Xbox*, measures of semantic relatedness should be preferred.

Other applications where a strict notion of similarity is more appropriate are automatic thesaurus generation and paraphrasing. In contrast, for word sense disambiguation the semantically related context word *coconut* is as useful as the similar word *tree* to disambiguate between the meanings of *palm*.

As these example applications show it is important to employ a word space with the right type of relations for use with a given application. But while (Rapp, 2002) and especially (Sahlgren, 2006) have investigated the effects of context choice and co-occurrence type on the semantic properties of the resulting word spaces, we are only aware of (Peirsman et al., 2007) to have tested the influence of dimension reduction techniques (namely Random Indexing and frequency cutoffs) on the outcome. The aim of the present paper is to experimentally confirm that the application of other dimension reduction techniques like singular value decomposition (SVD) and corpus preprocessing techniques like lemmatization also have considerable effect on the nature of the resulting word space.

In the next section we present our method for computing distributional similarity, in section 3 we describe three other systems we have chosen for comparison. Section 4 evaluates the performance of the systems against human relatedness judgements and similarities based on WordNet. We report on a series of experiments concerning the size of the input corpus, the choice of context (syntactic vs. window-based), corpus preprocessing and filtering by word frequency. In section 5 we discuss the findings, and in the last section we summarize our contributions.

2 Our Method: DISCO

Our method for computing the distributional similarity between words is called DISCO (*extracting DIStributionally similar words using CO-occurrences*) and works as follows. In a preprocessing step, the corpus at hand is tokenized and highly frequent function words are eliminated. Since we want to keep the method independent from language-specific resources, neither part of speech tagging nor lemmatization are performed, and we use a simple context window of size ± 3 words for counting co-occurrences. Our evaluations showed that it is beneficial to take the exact position within the window into account, as has been done by (Rapp, 1999). This can be seen as a crude approximation of syntactic dependency relations. Instead of syntactic dependency triples like $<$*donut*, OBJ-OF, *eat*$>$ we get triples of the form $<$*donut*, -2, *eat*$>$. Consequently, the features that describe a word's distribution are not just words as in a pure bag-of-words approach, but ordered pairs of word and window position.

Consider the example in table 1. It shows two occurrences of the word *palm* in a context of ± 3 words. When taking the exact window position into account, then *palm* is described by the five different features that result from the two occurrences (we ignore function words), listed on the lower left of the table. The features $<$*, -3, *oil*$>$ and $<$*, +1, *oil*$>$ are distinct and have nothing more in common than $<$*, +3, *hand*$>$ and $<$*, -1, *provides*$>$. If the exact position is not observed, we get only four features (lower right of table 1), since the two occurrences of *oil* can not be distinguished any more. A context that observes the exact window position leads to tighter similarities than a window without exact position. In section 4.4 we evaluate the effect the window-position context bears on the resulting similarities.

Moving the window over our corpus gives us a co-occurrence matrix. Every row of the matrix describes a word, and is also called a "word vector".

-3	-2	-1		$+1$	$+2$	$+3$
oil	into	the	palm	of	his	hand
the	nuts	provides	palm	oil	while	the

<palm, -3, oil>	1	<palm, oil>	2
<palm, +3, hand>	1	<palm, hand>	1
<palm, -2, nuts>	1	<palm, nuts>	1
<palm, -1, provides>	1	<palm, provides>	1
<palm, +1, oil>	1		

Table 1: Example of using window position triples (WPT) as context for counting co-occurrences. WPT features are shown in the 1st column of the lowest row, the bag-of-words features in the 4th column.

The matrix size is not $v \times f$ as usual (with v being the number of words for which word vectors are built, f being the number of words used as features), but $v \times f \cdot r$ (r is the window size). The next step is to transform the absolute counts in the matrix fields into more meaningful weights. For this feature weighting we found the measure proposed by (Lin, 1998c), which is based on mutual information, to be optimal:

$$g(w, w', r) = \log \frac{(f(w, r, w') - 0,95) f(*, r, *)}{f(w, r, *) f(*, r, w')} \quad (1)$$

where w and w' stand for words and r for a window position (or a dependency relation, respectively), and f is the frequency of occurrence.

To arrive at a word's distributionally similar words the next step is to compare every word vector with all other word vectors. For vector comparison we use Lin's information theoretic measure ((Lin, 1998a)) as given in equation (2). Because a word vector represents the distribution of a word in the corpus, this vector comparison gives us the words which are used in similar contexts. Put differently, it finds the words that share a maximum number of common co-occurrences. For example, if *bread* co-occurs with *bake*, *eat*, and *crispy*, and *cake* also co-occurs with these three words, then *bread* and *cake* will be distributionally similar. Note that *bread* and *cake* do not need to co-occur themselves a single time to be regarded as similar.

As an example of the outcome, the twelve distributionally most similar words for *palm* are listed here:

> palms (0.1345) coconut (0.1059) olive (0.0870) pine (0.0823) citrus (0.0745) oak (0.0677) mango (0.0652) cocoa (0.0645) banana (0.0627) bananas (0.0623) trees (0.0570) fingers (0.0560)

Such a list of distributionally similar words can in turn be seen as the "second order" word vector of the given word, containing not only the words which occur together with it, but those that occur in similar contexts. We can now compare two words based on their second order word vectors, too. This use of higher-order co-occurrences is to some extent comparable to what is achieved in LSA by singular value decomposition (Kontostathis and Pottenger, 2006).

In conclusion, DISCO provides two different similarity measures: DISCO1, that compares words based on their sets of co-occurring words, and DISCO2, that compares words based on their sets of distributionally similar words (i.e. DISCO2 compares the second order word vectors).

3 Description of the other Systems

LSA. Latent semantic analysis (Landauer and Dumais, 1997) is arguably the most popular variant of word space. Its core step is a dimension reduction technique called singular value decomposition (SVD). SVD computes the least mean square error projection of a matrix onto a lower dimensional matrix. It achieves a kind of generalization by combining columns that represent words with similar meanings. In our experiments we used the LSA implementation accessible at `http://lsa.colorado.edu`.

PMI-IR (pointwise mutual information - information retrieval). (Turney, 2001) presents a method for computing the similarity between arbitrary words that utilizes the WWW search engine AltaVista[1] according to the following for-

[1] `http://www.altavista.com`

$$lin(w,w') = \frac{\sum_{p=1}^{r}\sum_{w_i=1}^{v}\begin{cases} g(w,w_i,p)+g(w',w_i,p) & : \quad g(w,w_i,p)>0 \quad \text{and} \quad g(w',w_i,p)>0 \\ 0 & : \quad \text{else} \end{cases}}{\sum_{p=1}^{r}\sum_{w_i=1}^{v}(g(w,w_i,p)+g(w',w_i,p))} \tag{2}$$

mula, adapted from pointwise mutual information:

$$\textit{PMI-IR}(w_1,w_2) = log\frac{H(w_1 \textit{NEAR}\ w_2)}{H(w_1)H(w_2)} \tag{3}$$

where $H(w)$ is the number of hits the search engine returns for the query w. The more often two words co-occur near each other on a web page, the higher is their PMI-IR score. We computed the PMI-IR similarity values for our evaluation data by querying AltaVista on 4/10/2008.

WordNet::Similarity. WordNet::Similarity (Pedersen et al., 2004) is a Perl module based on WordNet that has been widely used in a variety of natural language processing tasks. It implements three measures of semantic relatedness (namely Hirst-St.Onge (hso), Lesk (lesk) and vector pairs (vp)) and six measures of semantic similarity (Jiang and Conrath (jcn), Leacock and Chodorow (lch), Lin (lin), path length (path), Resnik (res), and Wu and Palmer (wup)). The latter utilize the *is-a* relations in WordNet. Since there are only *is-a* relations between nouns and between verbs in WordNet, the similarity measures cannot be applied to adjectives or across part of speech.

4 Evaluation

4.1 Data

We built several DISCO word spaces according to the method outlined above. The first word space is based on 300,000 articles from the English Wikipedia[2], amounting to some 267 million tokens. We considered all words with a corpus frequency of at least 100, resulting in a vocabulary size of v =226,000, and used the f =101,000 most frequent words as feature words. This word space is employed in experiments 1 and 2 (sections 4.2 and 4.3).

In experiment 3 (section 4.4) we tested different parameter settings, which meant we had to build a number of word spaces. To limit the computational effort we decided to use a smaller corpus: the British National Corpus which consists of roughly 110 million tokens.

(Finkelstein et al., 2001) prepared a list of 353 noun-noun pairs and employed 16 subjects to estimate their semantic relatedness on a scale from 0 to 10. We use this list as our evaluation data. As seven word pairs contained at least one word that was unknown to WordNet, we deleted them from the list, leaving 346 word pairs for testing.

4.2 Correlation with Human Judgements of Semantic Relatedness

Our first experiment measures the correlation (according to the Pearson correlation coefficient) of the candidate systems with the averaged semantic relatedness scores assigned to the 346 word pairs by the human subjects. Table 2 shows the results. The first two correlation values in the first row of the table are taken from (Finkelstein et al., 2001). Among the systems listed in the first row, DISCO1 shows the lowest correlation with the human judgements, comparable to that of Finkelstein et al.'s vector approach. DISCO2 performs much better, but is still worse than LSA. The best score is achieved by PMI-IR, which is in accordance with other results reported in the literature (Turney, 2001).

The WordNet-based measures (shown in the second row of the table) perform worse, which comes as no surprise for the six measures of similarity, since they are not intended to measure relatedness. But the three measures of relatedness (hso, lesk, and vp) do not perform much better. The best scoring vector pairs measure (vp) only achieves the same score as DISCO1.

4.3 Correlation with WordNet::Similarity

We now take the semantic similarity values produced by the six WordNet similarity measures as gold standard and compare the correlation of the other test systems with these similarities. We assume that the six measures provide a sensible similarity gold standard since they are based exclusively on WordNets IS-A noun hierarchy and do not take into account other lexical relations or associations.

[2] `http://en.wikipedia.org`

		Vector-based	LSA	PMI-IR	DISCO1	DISCO2		
		0.41	0.56	**0.63**	0.39	0.51		
hso	lesk	vp	jcn	lch	lin	path	res	wup
0.35	0.21	0.39	0.23	0.35	0.30	0.38	0.36	0.30

Table 2: Correlation of several systems with the semantic relatedness values assigned by humans.

	jcn	lch	lin	path	res	wup	avg.
PMI-IR	0.14	0.12	0.06	0.15	0.22	0.11	0.13
LSA	0.16	0.26	0.21	0.29	0.28	0.22	0.24
DISCO1	**0.38**	0.39	0.33	**0.45**	0.43	0.33	**0.38**
DISCO2	0.15	**0.40**	**0.39**	0.35	**0.44**	**0.40**	0.36

Table 3: Correlation between WordNet-based semantic similarity and four systems based on word distributions.

In this task, PMI-IR performs worst (cf. table 3), whereas DISCO1 shows the highest correlation on average. The behaviour of the two DISCO measures is difficult to compare, because DISCO1 scores higher than DISCO2 three times, but DISCO2 also scores higher than DISCO1 four times. If we take the averaged score, DISCO1 turns out slightly better. In any case, both DISCOs perform much better than PMI-IR and LSA.

4.4 Effect of different parameter settings and techniques

Our third experiment tests various parameter settings for the DISCO1 measure. As DISCO2, which was meant as a substitute for LSA, performed worse than LSA in the first experiment, we do not further evaluate this measure. Instead, we combine DISCO1 with SVD in the last part of experiment 3.

In the previous experiments a 267 million token corpus from the English Wikipedia was used, in the following we use a smaller corpus, namely the British National Corpus, which consists of only about 110 million tokens, i.e. has only 40% of the size of the Wikipedia corpus.

The reduced size of the input data has a noticeable effect on the computation of semantic relatedness (first row in table 4). While in the previous experiments DISCO1 achieved a correlation of 0.39 with the Finkelstein gold standard for semantic relatedness (abbreviated as *finkel353* in table 4), the same method now only scores 0.34 on the same task, which constitutes a decrease by 12.8%.

To quantify the effect of corpus size on semantic similarity we compute the correlation with WordNet::Similarity's Resnik measure from experiment 2 (*res* in table 4). As one can see from tables 3 and 4, the reduced size of the corpus has no negative effect on semantic similarity: the correlation stands at 0.43.

	finkel353	res
DISCO1 WPT	0.34	**0.43**
DISCO1 without WPT	0.32	0.12
DISCO1 WPT lemmatized	**0.36**	0.41
DISCO1 dependency	**0.36**	0.39

Table 4: Experiment 3: Correlation between DISCO1 and two gold standards for different parameter settings.

To quantify the benefit of our poor man's dependency triples – the window position triples (WPT) as explained in section 2 – we built a word space with a simple bag-of-words window as context. The size of the window remains the same (three words on either side of the target word), but the position inside the window is not observed any more. The result is shown in the second row of table 4. The correlation with the semantic relatedness gold standard drops from 0.34 to 0.32 (-5.9%). The correlation with the similarity reference crashes down by 72.1% from 0.43 to 0.12.

Next we lemmatized the corpus before applying DISCO using the well known Tree Tagger (Schmid, 1994). While lemmatization has a positive effect on semantic relatedness (cf. the third row in table 4) it has an almost equally strong negative effect on semantic similarity.

In the next part of experiment 3 we ran the Minipar (Lin, 1998b) robust dependency parser over

f	finkel353	res
101,000	0.34	0.43
50,000	0.37	0.43
20,000	0.40	0.45
10,000	**0.41**	**0.46**
5,000	0.40	0.43
1,000	0.38	0.43
500	0.36	0.33

Table 5: Frequency cutoff: Correlation of DISCO1 with the two gold standards for different quantities of feature words.

our corpus to extract syntactic dependency triples. This increases the correlation with the semantic relatedness gold standard from 0.34 to 0.36 (last row in table 4). That is, robust parsing has the same effect as lemmatization. Since Minipar automatically does lemmatization, we can conclude that syntactic dependency triples are no better than our window position triples.

Surprisingly, the correlation with the semantic similarity gold standard drops from 0.43 to 0.39 (-9.3%). We hypothesize that this might be the effect of noise produced by the parser.

Recall from section 2 that the size of the co-occurrence matrix is given by $v \times f \cdot r$ with v being the number of vocabulary items for which word vectors are collected, f being the number of feature words (the words that are used to populate the word vectors), and r being the window size. As stated in section 4.1, for all experiments so far we chose $f = 101,000$, i.e. we used the 101,000 most frequent words in the corpus as feature words. We will now systematically decrease this parameter. The effect of this adjustment can be seen in table 5. As the number of feature words decreases, the correlation with both gold standards increases, peaking at $f = 10,000$. For f lower than $1,000$, the performance of semantic similarity drops sharply, whereas semantic relatedness seems to suffer relatively less from such a dramatic decrease of the number of features. Note that for the optimal setting of this parameter the performance for semantic relatedness is now even better than with the much bigger corpus from the previous experiments (0.41 as compared to 0.39 in table 2). The same holds for the correlation with the semantic similarity gold standard (0.46 vs. 0.43, cf. table 3).

The frequency cutoff at $f = 10,000$ lead to a considerable reduction of the size of our co-occurrence matrix which enabled us to apply the singular value decomposition to it. We used SVDLIBC[3] to reduce the matrix to its 300 principal components (i.e. we reduced the matrix size from $v \times 10,000 \cdot r$ to $v \times 300$). The result is shown in table 6. The use of SVD significantly increases the correlation with the relatedness gold standard, whereas it decreases the correlation with all six similarity measures.

5 Discussion

In the first experiment (see section 4.2) we found that PMI-IR scored best at the task of computing semantic relatedness, outperforming LSA and even more DISCO. The most interesting result of experiment 1 was that DISCO2 scored much better than DISCO1. Since the only difference between the two measures is the use of second order co-occurrences by DISCO2, we can conclude that for computing semantic relatedness higher-order co-occurrences can substitute for SVD – not fully, but at least to a certain degree.

We also observed that the three WordNet-based measures of semantic relatedness performed quite badly. The reason for this is unclear.

Experiment 2 (section 4.3) evaluated the correlation of different methods with semantic similarities produced by WordNet::Similarity. It was shown that DISCO1 scored much better in this task than PMI-IR and LSA. Moreover, the higher-order co-occurrences of DISCO2 did not seem to have a consistent positive effect. From this result we can conclude that singular value decomposition and higher-order co-occurrences increase the performance when computing semantic relatedness, but they do not help in computing semantic similarity. This conclusion is confirmed by the last part of experiment 3 (section 4.4), where we combined DISCO1 with SVD, leading to a significant performance increase for the relatedness gold standard, but to a decrease for all six similarity measures.

The poor performance of PMI-IR in the second experiment can be explained by the type of co-occurrence it is based on. While DISCO1 compares words based on their collocation sets, thereby finding words that are used similarly, PMI-IR's similarities *are* collocations. Therefore it rather produces very loose word similarities, i.e.

[3] `http://tedlab.mit.edu/~dr/SVDLIBC/`

	finkel353	jcn	lch	lin	path	res	wup
DISCO1-10K	0.41	**0.62**	**0.52**	**0.50**	**0.52**	**0.46**	**0.47**
DISCO1-10K-SVD	**0.55**	0.46	0.37	0.41	0.39	0.38	0.35

Table 6: Performance of DISCO1 after frequency cutoff at f = 10,000 with and without singular value decomposition (SVD)

words that are topically similar.

Experiment 3 (section 4.4) suggests that measures of relatedness highly profit from more input data. This is confirmed by the finding of experiment 1 that PMI-IR outperforms LSA, despite the fact that both methods use co-occurrence in a short piece of text as context. While LSA additionally employs SVD, there is nothing in PMI-IR that would explain its strong performance except the huge size of the corpus it is based on (the web).

Experiment 3 also confirms that the recording of the position within the context window has an enormous positive effect on computing semantic similarity, while the effect on semantic relatedness is less significant. This could be expected from the discussion of the relevant literature in section 1, where distributional similarity is explicitly defined by the use of a strict context that pays attention to syntactic features like word order. Our experiments indicate that any method which "blurs" the context (bag-of-words window, lemmatization, SVD) decreases the quality of semantic similarity. Instead, a "naked" approach based on indirect co-occurences should be chosen. This finding is in line with (Peirsman et al., 2007) who state that "severely reducing the dimensionality of the word vectors leads to a retrieval of more loosely related words." One should presume that consequently a syntactic context would score best, since this is the strictest imaginable context. Therefore, it is a bit surprising that the use of Minipar did not lead to an improvement. (Rapp, 2004) seems sceptical about the advantages of syntactic dependency triples over simple window approaches and assumes that the employment of a part-of-speech tagger will result in the same performance as the use of a parser. This hypothesis is confirmed by our results. (Grefenstette, 1996) and recently (Padó and Lapata, 2007) and (Peirsman et al., 2007) compared syntactic and window based approaches, and found that syntactic contexts performed superior. However, they used bag-of-words windows without taking into account the position inside the window. We propose that our window position triples should be rather seen as a syntactic context and not as a bag-of-words context. Yet we believe that for languages with a less strict word order than English (like for example Czech) syntactic dependency triples will outperform our window position triples.

Another interesting finding of experiment 3 resulted from the application of a frequency filter. We found that limiting the size of the co-occurrence matrix to the 10,000 most frequent feature words yielded the highest performance for both semantic similarity and relatedness.

6 Conclusion

In the present paper we have reported on several experiments regarding the influence of dimension reduction techniques, corpus size, and choice of context on the semantic properties of the resulting word spaces.

For future work we propose to carry out application-centered evaluations in order to confirm the practical relevance of the similarity–relatedness distinction put forth in this paper.

DISCO is freely available for research purposes at `http://www.linguatools.de/disco_en.html`.

References

A. Budanitsky and G. Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1).

P. Cimiano, A. Hotho, and S. Staab. 2005. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *Journal of Artificial Intelligence Research*, 24:305–339.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2001. Placing search in context: the concept revisited. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 406–414, New York, NY, USA. ACM.

M. Geffet and I. Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proc.*

of the 43rd Annual Meeting of the ACL, pages 107–114.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer, Dordrecht.

G. Grefenstette. 1996. Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. In B. Boguraev and J. Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 205 – 216. MIT Press, Cambridge, MA.

Z.S. Harris. 1968. *Mathematical Structures of Language*. Interscience Publishers, New York.

A. Kilgarriff and C. Yallop. 2000. What's in a thesaurus? In *Proceedings of the Second Conference on Language Resources and Evaluation*, pages 1371–1379, Athens.

A. Kilgarriff. 2003. Thesauruses for Natural Language Processing. In *Proceedings of Natural Language Processing and Knowledge Engineering (NLPKE)*, Beijing.

A. Kontostathis and W.M. Pottenger. 2006. A framework for understanding latent semantic indexing (LSI) performance. *Information Processing and Management*, 42(1):56–73, January.

T.K. Landauer and S.T. Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

D. Lin. 1998a. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING-ACL 1998*, Montreal.

D. Lin. 1998b. Dependency-based Evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*, Granada, Spain.

D. Lin. 1998c. Extracting Collocations from Text Corpora. In *Workshop on Computational Terminology*, pages 57–63, Montreal, Kanada.

S. Mohammad and G. Hirst. 2005. Distributional Measures as Proxies for Semantic Relatedness. Unpublished.

C. Müller, I. Gurevych, and M. Mühlhäuser. 2007. Integrating Semantic Knowledge into Text Similarity and Information Retrieval. In *Proceedings of the First IEEE International Conference on Semantic Computing*, pages 57–63, Montreal, Kanada.

S. Padó and M. Lapata. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

S. Patwardhan, S. Banerjee, and T. Pedersen. 2007. UMND1: Unsupervised Word Sense Disambiguation Using Contextual Semantic Relatedness. In *SemEval-2007: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 390–393, Prague, Czech Republic, June.

T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04)*, pages 38–41, Boston, MA., May.

Y. Peirsman, K. Heylen, and D. Speelman. 2007. Finding semantically related words in Dutch. Cooccurrences versus syntactic contexts. In *Workshop on Contextual Information in Semantic Space Models (CoSMo 2007)*, Roskilde.

R. Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of ACL*, pages 519–526.

R. Rapp. 2002. The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. In *Proceedings of COLING-02*, Taipeh.

R. Rapp. 2004. A Freely Available Automatically Generated Thesaurus of Related Words. In *Proceedings of LREC 2004*, pages 395–398.

M. Sahlgren and J. Karlgren. 2008. Buzz Monitoring in Word Space. In *European Conference on Intelligence and Security Informatics (EuroISI 2008)*. Esbjerg, Denmark.

M. Sahlgren. 2001. Vector-based Semantic Analysis: Representing Word Meanings Based on Random Labels. In Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli, editors, *The Acquisition and Representation of Word Meaning*. Kluwer Academic Publishers.

M. Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm.

H. Schmid. 1994. Probabilistic Part-of-speech Tagging Using Decision Trees. *International Conference on New Methods in Language Processing*.

P.D. Turney. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proc. of the Twelth European Conference on Machine Learning*, pages 491–502.

P.D. Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 905–912, Manchester, UK.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206