Testing semantic similarity measures for extracting synonyms from a corpus

Olivier Ferret

CEA, LIST, Vision and Content Engineering Laboratory,
Fontenay-aux-Roses, F-92265, France.
olivier.ferret@cea.fr

Abstract

The definition of lexical semantic similarity measures has been the subject of lots of works for many years. In this article, we focus more specifically on distributional semantic similarity measures. Although several evaluations of this kind of measures were already achieved for determining if they actually catch semantic relatedness, it is still difficult to determine if a measure that performs well in an evaluation framework can be applied more widely with the same success. In the work we present here, we first select a semantic similarity measure by testing a large set of such measures against the WordNet-based Synonymy Test, an extended TOEFL test proposed in (Freitag et al., 2005), and we show that its accuracy is comparable to the accuracy of the best state of the art measures while it has less demanding requirements. Then, we apply this measure for extracting automatically synonyms from a corpus and we evaluate the relevance of this process against two reference resources, WordNet and the Moby thesaurus. Finally, we compare our results in details to those of (Curran and Moens, 2002).

1. Introduction

This article takes place in the field of what is called *lexical semantic similarity* or even more generally *lexical semantic relatedness*. The objective of the work done in this field is to determine how close two words are from a semantic viewpoint and if their similarity is high enough, the type of the semantic relation they share. A part of this work is dedicated to the design of similarity measures that exploit more or less structured sources of knowledge, such as dictionaries or lexical networks (see (Zesch and Gurevych, 2010) for an overview). In this article, we focus more particularly on corpus-based approaches. Most of them rely on the distributional hypothesis, according to which words found in similar contexts tend to have similar meanings (Firth, 1957). Following (Grefenstette, 1994) and (Lin, 1998), this hypothesis is generally implemented by collecting co-occurrences from a large corpus and characterizing each term *T* from the corpus by the vector of its co-occurrents. These co-occurrents, also considered as features, are weighted according to the strength of their link with *T*. Finally, the semantic similarity of two terms is evaluated by applying a similarity measure between their vectors. This perspective was adopted for instance by (Curran and Moens, 2002) and (Weeds, 2003), where a wide set of similarity measures and feature weighting functions were tested.

Some works propose variants of this basic schema but without changing the core principles of the distributional approach. One of these variants is based on a probabilistic viewpoint: each term is characterized by a probability distribution over its co-occurrents and the semantic similarity of two terms is evaluated by a distance between their probability distributions (Weeds, 2003). The application of dimensionality reduction techniques to the co-occurrent vectors covers another set of variants in which the semantic similarity between terms is evaluated in the semantic space resulting from the dimensionality reduction. The *Latent Semantic Analysis* from (Landauer and Dumais, 1997) and the *Random Indexing* from (Salgren, 2006) are the most significant representatives of this trend.

Works about lexical semantic similarity can also be characterized through the way they evaluate the semantic measures they propose. One common way to perform this evaluation is to apply these measures to a set of TOEFL synonym questions, as initially proposed by (Landauer and Dumais, 1997). Each question consists in a headword and a set of 4 words among which a synonym of the headword has to be identified. After the results for the TOEFL questions had reached a high level (Turney et al., 2003), several extensions of this evaluation approach were proposed, either by using questions from similar tests such as the ESL test (Moraliyski and Dias, 2007), building larger sets of questions by relying on a resource such as WordNet (Freitag et al., 2005; Piasecki et al., 2007) or extending the kind of relations covered by the test as with the presence of analogies in the SAT test (Turney, 2008).

Another common way to evaluate semantic measures is to compare their results to a gold standard. Human judgments about the similarity of couples of words are sometimes used as a direct gold standard (Weeds, 2003) but this kind of resources are rare and small. As a consequence, a more indirect evaluation is generally performed (Lin, 1998; Curran and Moens, 2002): the semantic measures to test are used for finding the most similar neighbors of a headword and these neighbors are evaluated against a reference set of synonyms or related words for this headword taken from resources such as WordNet (Miller, 1990) or the Roget's thesaurus (Roget, 1911).

In this article, our overall objective is to extract synonyms for nouns from a corpus by relying on the distributional hypothesis, which starts by selecting an appropriate semantic similarity measure. Although we have seen that many works were done about lexical semantic similarity, it is still difficult to know if their results can be transposed to our problem: most of them are about TOEFL-like tests, which are less difficult tasks than ours; when they come from the evaluation against a gold standard, they are generally given only for a restricted set of words (Curran and Moens, 2002) or the evaluation measure takes into account a larger set of semantically similar words than only synonyms (van der Plas and Bouma, 2004). Hence, in this article, we first report our experiments for finding a semantic similarity measure that performs well on an extended TOEFL test within a set of constraints. Then, we study the results of this measure for extracting synonyms. This is an attempt to have a more global view on semantic similarity, following (Turney, 2008) or (Baroni and Lenci, 2009).

2. Test of semantic similarity measures

2.1. Definition of similarity measures

A semantic similarity measure based on the distributional hypothesis heavily depends on the corpus from which distributional data are taken and the means used for extracting these data. Although corpora for distributional similarity tend to be bigger and bigger, such as in (Pantel et al., 2009), we decided to rely in our case on the AQUAINT-2 corpus, which is a middle-size corpus made of around 380 million words coming from news articles. This choice is motivated by the fact that collecting huge sets of textual data is not always possible for all domains and for all languages.

Concerning the extraction of distributional data, we also chose to use limited means because advanced linguistic tools are not available, or at least freely available, for all languages. While many works, but not all of them, follow (Lin, 1998) and (Curran and Moens, 2002) and rely on a syntactic parser, we only applied lemmatization and selected content words (nouns, verbs and adjectives) as a pre-processing step, which was done by the *TreeTagger* tool (Schmid, 1994). As a consequence, the distributional data associated to each word took the form of a vector of co-occurrents collected by a fixed-size window and not a vector of syntactic co-occurrents based on syntactic dependency relations. We call this vector a *context vector* and we classically refer to its elements, *i.e.* the co-occurrents of the headword, as *features* (f). These features were nouns, verbs and adjectives[1].

Within this framework, we defined a semantic similarity measure between word x and word y through four characteristics:

- a measure to compare the context vectors of x and y;
- a function to weight the significance of the features of a context vector;
- the size of the window used for collecting co-occurrents;
- the threshold applied for discarding low-frequency co-occurrents before building context vectors.

Table 1 shows the context similarity measures and the feature weighting functions we tested as they are defined in (Curran and Moens, 2002) for some of them. The measure proposed by Ehlert (Ehlert, 2003) is a special case: as it is a probabilistic measure, it relies on the probability of features and not on their weight, which means that no weighting function is applied.

2.2. Results and evaluation

As mentioned in the introduction, the selection of the semantic similarity measure we used for synonym extraction was based on an extended TOEFL test and more precisely, on the WordNet-based Synonymy Test (WBST) proposed in (Freitag et al., 2005)[2]. WBST was produced by generating automatically a large set of TOEFL-like questions from the synonyms in WordNet. (Freitag et al., 2005) shows that this test is more difficult than the initial TOEFL test made of 80 questions that was first used in (Landauer and Dumais, 1997). The part of WBST restricted to nouns is made of 9887 questions. All the possible associations between the context similarity measures and the feature weighting functions presented in the previous section were tested with window sizes between 1 and 5 and frequency thresholds between 1 and 5[3]. For each question of the test, the tested similarity measure was computed between the headword and each of the four possible choices. These choices were then sorted according to the decreasing values of their similarity score and the choice with the highest score was taken as a candidate synonym. In the rare cases where no choice could be made from the distributional data (between 3.7 and 6.7% of questions according the measure), a random choice was performed. We classically used the percentage of relevant candidate synonyms as our evaluation measure, which can also be seen as the precision at rank 1 as our similarity measures sorted candidates. Table 2 shows the results of this evaluation.

The first thing to notice is that for almost all our context similarity measures, the best results are obtained with a window size and a frequency threshold equal to 1. Moreover, we can observe that the accuracy of similarity measures tends to decrease while the frequency threshold and the window size increase[4]. This means that semantic similarity is preferably characterized by very short range co-occurrents among which only a weak selection has to be performed for discarding co-occurrences that are the most likely to be present only by chance[5]. The second main thing to notice is that the *Cosine* measure with *Pointwise Mutual Information* and the *Ehlert* measure have good results, which agrees the findings of (Freitag et al., 2005). However, (Freitag et al., 2005) had found that *Ehlert* outperforms *Cosine* while we found the opposite. More precisely, our best accuracy for *Cosine* is equal to its best accuracy (without supervised optimization) for *Ehlert*. Moreover, its measures had been defined with a one-billion word corpus, hence much larger than ours, and the frequency of the WBST nouns in their corpus was as least 1000 while we only discarded words with frequency lower than 11.

This evaluation also shows that measures such as *Jaccard*, *Dice*† or *Lin*, whose precision is high for extracting similar words according to (Curran and Moens, 2002), have close accuracy values that are significantly lower than *Cosine* or

[1] More precisely, only the words whose frequency is strictly higher than 10 are kept, both in context vectors and for headwords.

[2] Available at the following address: http://www.cs.cmu.edu/~dayne/wbst-nanews.tar.gz.

[3] Frequency must be higher or equal to the threshold.

[4] There are some rare exceptions, which mainly concern the Jaccard† measure.

[5] A frequency threshold equal to 1 discards around half co-occurrences.

Context similarity measure		**Feature weighting function**	
Cosine	$\frac{\sum_i wgt(x_i) \cdot wgt(y_i)}{\sqrt{\sum_j wgt(x_j)^2 \cdot \sum_j wgt(y_j)^2}}$	Pointwise Mutual Information (pmi)	$\log(\frac{p(x,f)}{p(x) \cdot p(f)})$
Jaccard	$\frac{\sum_i min(wgt(x_i), wgt(y_i))}{\sum_j max(wgt(x_j), wgt(y_j))}$	T-test	$\frac{p(x,f) - p(x) \cdot p(f)}{\sqrt{p(x) \cdot p(f)}}$
Jaccard†	$\frac{\sum_i min(wgt(x_i), wgt(y_i))}{\sum_i max(wgt(x_i), wgt(y_i))}$	Tf.Idf	$N(x,f) \cdot \log(\frac{N_x}{N_{x,f}})$
Dice	$\frac{2 \cdot \sum_i min(wgt(x_i), wgt(y_i))}{\sum_j wgt(x_j) + \sum_j wgt(y_j)}$		
Dice†	$\frac{2 \cdot \sum_i min(wgt(x_i), wgt(y_i))}{\sum_i wgt(x_i) + wgt(y_i)}$		
Lin	$\frac{\sum_i wgt(x_i) + wgt(y_i)}{\sum_j wgt(x_j) + \sum_j wgt(y_j)}$		

Table 1: Tested similarity measures for contexts and weighting functions for features[6]

window size		1			3			5		
frequency threshold		1	3	5	1	3	5	1	3	5
cosine	pmi	**71.6**	69.7	67.6	65.7	63.7	62.8	62.5	60.6	59.4
	t-test	**68.9**	66.7	65.0	65.4	64.6	63.8	63.3	62.9	62.0
	tf.idf	**64.0**	63.1	62.0	63.3	62.9	62.5	62.6	62.4	61.7
ehlert	–	**70.2**	68.5	66.2	68.9	67.2	65.9	66.9	65.9	64.4
jaccard	pmi	**64.8**	63.0	61.7	57.1	55.0	54.1	54.6	52.6	51.3
	t-test	**68.1**	65.8	63.9	61.3	58.8	57.7	58.4	55.9	54.6
	tf.idf	**54.2**	53.9	53.6	49.7	49.6	49.3	48.0	47.9	47.4
dice	pmi	**64.8**	63.0	61.7	57.1	55.0	54.1	54.6	52.6	51.3
	t-test	**68.1**	65.8	63.9	61.3	58.8	57.7	58.4	55.9	54.6
	tf.idf	**54.2**	53.9	53.6	49.7	49.6	49.3	48.0	47.9	47.4
lin	pmi	**65.6**	63.5	61.7	57.0	54.6	53.6	54.2	52.1	51.1
	t-test	**67.3**	65.3	63.3	61.0	59.5	58.9	58.5	57.3	55.9
	tf.idf	**60.6**	59.6	58.3	57.9	56.6	55.9	56.6	54.9	53.9
dice†	pmi	**65.0**	63.2	61.5	58.7	57.5	57.0	56.5	55.9	55.3
	t-test	**66.0**	64.3	62.3	59.7	57.9	57.0	57.5	56.0	55.1
	tf.idf	51.6	52.3	**52.7**	48.4	47.9	48.3	47.2	47.2	46.6
jaccard†	pmi	**56.1**	54.7	53.2	54.3	54.3	53.4	54.0	54.3	53
	t-test	39.6	37.9	38.2	46.7	43.7	42.2	**48.1**	45.7	43.0
	tf.idf	35.3	34.3	34.4	40.2	38.1	37.3	**41.4**	39.7	38.4

Table 2: Evaluation of semantic similarity measures

Ehlert's accuracies. For these measures, *T-test* is the best weighting function, which is compatible with (Curran and Moens, 2002), while *Tf.idf* is the worst. *Jaccard*† is clearly the worst choice as a context similarity measure. Finally, our best measure compares favorably with (Broda et al., 2009), which uses the nouns of WBST for evaluation as in our case but relies on syntactic co-occurrences collected from the British National Corpus, a 100 million word corpus. For nouns with frequency > 10, its best accuracy is equal to 68.04%.

3. Applying a lexical similarity measure for extracting synonyms and similar words

3.1. Principles

Results from the previous section show that we have built a distributional semantic similarity measure that performs at least as well as state of the art measures on a standard benchmark for evaluating semantic similarity. We now examine in this section to what extent this measure can be used to extract synonyms and similar words.

Our extraction process is simple: the possible synonyms of a word are found by retrieving its N nearest neighbors according to our similarity measure. In our case, the retrieval process only consists in applying the similarity measure between the target word and all the other words of the considered vocabulary with the same part-of-speech. Finally, all

[6] i: index on shared features of x and y; j: index on all features of x or y; $N(x, c)$: frequency of c as a co-occurrent of x; N_x: number of words; $N_{x,c}$: number of words having c as co-occurrent.

freq.	ref.	# words	#target syno.	#found syno.	R-prec.	MAP	P@1	P@5	P@10	P@100
> 10	W	10,473	29,947	7,374	0.082	0.098	0.117	0.051	0.034	0.007
(all)	M	9,216	460,923	43,950	0.067	0.032	0.241	0.164	0.130	0.048
# 14670	WM	12,243	473,833	46,656	0.077	0.056	0.225	0.140	0.108	0.038
> 1000	W	3,690	13,509	3,826	**0.111**	**0.125**	0.171	0.077	0.051	0.010
# 4378	M	3,732	258,836	29,426	0.102	0.049	**0.413**	**0.280**	**0.219**	**0.079**
	WM	4,164	263,216	30,375	0.110	0.065	**0.413**	0.268	0.208	0.073
$100 <$	W	3,732	9,562	2,733	0.104	**0.125**	0.136	0.058	0.037	0.007
≤ 1000	M	3,306	136,467	12,664	0.064	0.031	0.187	0.131	0.104	0.038
# 5175	WM	4,392	140,750	13,844	0.092	0.073	0.209	0.123	0.093	0.031
≤ 100	W	3,051	6,876	815	0.021	0.033	0.026	0.012	0.009	0.003
# 5117	M	2,178	65,620	1,860	0.012	0.005	0.025	0.015	0.015	0.008
	WM	3,687	69,867	2,437	0.021	0.024	0.033	0.017	0.015	0.007

Table 3: Evaluation of synonym extraction

these words are sorted according to their similarity value and only the first N, which is equal to 100 in our experiments, of them are kept[7]. As we use the *Cosine* measure for evaluating the semantic similarity of words, we could use techniques such the ones described in (Bayardo et al., 2007) to face the scalability problem of our basic approach for retrieving the nearest neighbors of a word. (Pantel et al., 2009) also addresses this problem for huge sets of data.

3.2. Results and evaluation

Table 3 shows the results of the application of the best similarity measure of the previous section to the extraction of synonyms and similar words. Two well-known resources were taken as reference: WordNet, more precisely its version 3.0, and the Moby thesaurus (Ward, 1996). As we focus on the ability of a semantic similarity measure to extract reliable synonyms more than on the coverage of these resources, we filtered these two references by removing from them all the words that weren't part of the set of mono-term nouns of the AQUAINT 2 corpus for which our distributional data were collected. We also built a third reference (WM) by merging the data coming from WordNet (W) and the Moby thesaurus (M).

In distributional approaches, the frequency of words related to the size of the corpus is an important factor. Hence, we give our results globally but also for three ranges of frequencies that split our vocabulary into roughly equal parts (see first column of Table 3): high frequency nouns (frequency > 1000), middle frequency nouns ($100 <$ frequency ≤ 1000) and low frequency nouns ($10 <$ frequency ≤ 100). The third column of Table 3 gives for each resource the number of words for which the evaluation was actually performed. This number is lower than the number of nouns of the first column as some nouns of the AQUAINT 2 corpus have no entry in our resources. The fourth column corresponds to the number of synonyms and similar words in our reference resources that have to be found for the nouns of the AQUAINT 2 corpus while the fifth column gives the number of synonyms and similar words that were actually found among the first 100 semantic neighbors of each target word of our distributional base. As these neighbors are ranked according to their similarity value with their target word, the evaluation measures can be taken from the Information Retrieval field by replacing documents with synonyms and queries with target words (see the three last columns of Table 3). The R-precision (R-prec.) is the precision after the first R neighbors were retrieved, R being the number of reference synonyms; the Mean Average Precision (MAP) is the average of the precision value after a reference synonym is found; precision at different cut-offs is given for the 1, 5, 10 and 100 first neighbors.

The results of Table 3 are globally low in spite of the good results on the WBST test of the similarity measure we have used. This weakness concerns both the recall of synonyms (around 25% for WordNet and 10% for the Moby thesaurus) and their rank among semantic neighbors (see R-precision, MAP and P@1,5,10,100). This observation goes beyond our particular experiments as the similarity measure we relied on is not specific to our framework. However, the situation is somewhat different depending on the frequency range of target words: the best results are obtained for high-frequency words and evaluation measures significantly decrease for words whose frequency is less than 100 occurrences. More globally, the ability for a distributional approach to catch the semantic relatedness of words seems to be closely correlated with the frequency of these words in the corpus from which distributional data are collected. While this is an argument in favor of the use of larger and larger corpora, as illustrated by (Pantel et al., 2009), it doesn't invalidate the idea that rare words may have a different distributional behavior that should be taken into account specifically.

Table 3 also shows that the characteristics of the reference resources has a significant impact on results. WordNet provides a restricted number of synonyms for each noun (2.8 on average) while the Moby thesaurus contains for each entry a larger number of synonyms and similar words (50 on average). This difference directly explains that the precision at rank 1, for words whose frequency is higher than

[7] It was performed approximately in 4 hours on 48 cores of a cluster.

1000, is equal to 0.413 for the Moby thesaurus while it is only equal to 0.171 for WordNet.

3.3. Discussion

As a reference evaluation framework doesn't exist for the extraction of synonyms by distributional methods, the comparison of our results with already existing works faces some difficulties. The main one is the lack of consensus about the type of the target relations to find. The extraction of synonyms such as those of WordNet is a difficult task because their low number (see previous Section) requires an extraction method with a very high precision for having acceptable results. As a consequence, the type of the reference relations goes generally beyond synonymy and is extended to the notion of similar words, which is supposed to account for semantic relatedness. A part of the relations of the Moby thesaurus can be put into this category in our case. (van der Plas and Bouma, 2004) followed a similar trend: although it relied on the Dutch EuroWordNet, it made use for evaluation of a WordNet similarity measure that also took into account the hierarchy of hypernyms. (Pantel et al., 2009) is another variant: it evaluated its results against *Entity Sets*, which gathered entities that were not only similar but more generally analogous.

(Curran and Moens, 2002) is more directly comparable to our work. It tested a large number of similarity measures based on syntactic co-occurrences by using them for extracting semantic neighbors. The evaluation of this extraction was done against the fusion of three thesauri: the Macquarie (Bernard, 1990), Roget's and Moby thesauri. It focused on 70 nouns randomly chosen from WordNet such that they were representative of WordNet's nouns in terms of frequency, number of senses, specificity (depth in the hierarchy of WordNet) and domains. Among all tested measures, the best results were obtained by the pair *Dice*† + *T-test*, with 0.76 as precision at rank 1, 0.52 at rank 5 and 0.45 at rank 10 for 70 nouns while our best precision is 0.413 at rank 1, 0.280 at rank 5 and 0.219 at rank 10 for 3,732 nouns. Apart from the fact that our test set is much larger than (Curran and Moens, 2002)'s one, the gold standards are partly different in the two cases, which can have a significant influence on results as we pointed it out in the previous section. For our 3,732 nouns, the Moby thesaurus provides 69 synonyms on average while 331 synonyms are available for each of the 70 nouns of (Curran and Moens, 2002)[8]. Moreover, we can observe that the recall rate is different for the two evaluations, equal to 8.3% for (Curran and Moens, 2002) and to 11.4% in our case. Even if the difference in the average number of relations for each entry in the two reference resources has an impact that it is difficult to estimate, this comparison suggests that using syntactic co-occurrences is a way to increase precision while graphical co-occurrences are more interesting for favoring recall.

4. Conclusion and future work

In this article, we have first presented our experiments for selecting the similarity measure based on the distributional paradigm that is the most likely to catch the semantic relatedness of words. This selection relied on an extended version of a TOEFL test, which is a classical way to evaluate semantic similarity measures. We then applied this selected measure for extracting automatically synonyms from a corpus and we evaluated the resulting set of candidate synonyms against two complementary resources: WordNet and the Moby thesaurus. Although the results of this evaluation are coherent with the state of the art, they show that results about semantic similarity for tasks such as TOEFL-like tests must be considered with caution when they are transposed to more difficult tasks such as finding synonyms. From our viewpoint, they represent a starting point for studying more precisely the kind of relations that are covered by distributional semantic similarity measures.

The most straightforward extension to this work is to substitute syntactic co-occurrences for graphical co-occurrences to determine if the use of syntactic features leads to increase precision, as it is suggested by our analysis of the results of (Curran and Moens, 2002). Futhermore, we would like to test methods for improving the quality of our distributional data, as those proposed in (Zhitomirsky-Geffet and Dagan, 2009) or (Broda et al., 2009), and extending them by taking into account new criteria such as words senses coming from a word sense discrimination method (Ferret, 2004). Finally, we plan to make publicly available A2ST[9], the similarity thesaurus we have built from the AQUAINT 2 corpus, similarly to the similarity thesaurus of Dekang Lin (Lin, 1998).

5. Acknowledgements

This work was partly supported by the Jean-Luc Lagardère Foundation (http://www.fondation-jeanluclagardere.com).

6. References

Marco Baroni and Alessandro Lenci. 2009. One distributional memory, many semantic spaces. In *EACL 2009 Workshop on Geometrical Models of Natural Language Semantics*, pages 1–8, Athens, Greece, March.

Roberto J. Bayardo, Yiming Ma, and Ramakrishnan Srikant. 2007. Scaling up all pairs similarity search. In *16th international conference on World Wide Web (WWW'07)*, pages 131–140, Banff, Alberta, Canada. ACM.

John R. L. Bernard. 1990. *The Macquarie Encyclopedic Thesaurus*. The Macquarie Library, Sydney, Australia.

Bartosz Broda, Maciej Piasecki, and Stan Szpakowicz. 2009. Rank-based transformation in measuring semantic relatedness. In *22nd Canadian Conference on Artificial Intelligence (Canadian Conference on AI)*, pages 187–190.

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 59–66, Philadelphia, USA.

Bert Ehlert. 2003. Making accurate lexical semantic similarity judgments using wordcontext co-occurrence statistics. Master's thesis, University of California, San Diego, USA.

[8] This difference shows that the Macquarie thesaurus is far much richer than the Moby thesaurus and WordNet.

[9] A2ST stands for AQUAINT 2 Similarity Thesaurus.

Olivier Ferret. 2004. Discovering word senses from a network of lexical cooccurrences. In *20th International Conference on Computational Linguistics (COLING 2004)*, pages 1326–1332, Geneva, Switzerland.

John R. Firth, 1957. *Studies in Linguistic Analysis*, chapter A synopsis of linguistic theory 1930-1955, pages 1–32. Blackwell, Oxford.

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Ninth Conference on Computational Natural Language Learning (CoNLL)*, pages 25–32, Ann Arbor, Michigan, USA.

Gregory Grefenstette. 1994. *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

Thomas K. Landauer and Susan 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.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (ACL-COLING’98)*, pages 768–774, Montréal, Canada.

George A. Miller. 1990. WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4).

Rumen Moraliyski and Gaël Dias. 2007. One sense per discourse for synonym detection. In *5th International Conference Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria.

Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, and Vishnu Vyas. 2009. Web-scale distributional similarity and entity set expansion. In *2009 Conference on Empirical Methods in Natural Language Processing*, pages 938–947, Singapore, August.

Maciej Piasecki, StanisBaw Szpakowicz, and Bartosz Broda. 2007. Extended similarity test for the evaluation of semantic similarity functions. In *Language Technology Conference (LTC)*.

Peter Roget. 1911. *Thesaurus of English words and phrases*. Longmans, Green and Co., London, UK.

Magnus Salgren. 2006. *The Word-space model*. Ph.D. thesis, Stockholm University.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*.

Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. In *4th International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pages 482–489, Borovets, Bulgaria.

Peter D. Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and association. In *COLING 2008*, pages 905–912.

Lonneke van der Plas and Gosse Bouma. 2004. Syntactic contexts for finding semantically related words. In Ton van der Wouden, Michaela Poß, Hilke Reckman, and Crit Cremers, editors, *Computational Linguistics in the Netherlands 2004, Selected Papers from the Fifteenth CLIN Meeting*, Leiden, Netherlands.

Grady Ward. 1996. Moby thesaurus. Moby Project.

Julie Weeds. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, Department of Informatics, University of Sussex.

Torsten Zesch and Iryna Gurevych. 2010. Wisdom of crowds versus wisdom of linguists - measuring the semantic relatdness of words. *Natural Language Engineering*, 16(1):25–59.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35(3):435–461, september.