WORD SPACE MODELS OF SEMANTIC SIMILARITY AND RELATEDNESS

Yves Peirsman

QLVL, University of Leuven & Research Foundation – Flanders

yves.peirsman@arts.kuleuven.be

Abstract. Word Space Models provide a convenient way of modelling word meaning in terms of a word's contexts in a corpus. This paper investigates the influence of the type of context features on the kind of semantic information that the models capture. In particular, we make a distinction between semantic similarity and semantic relatedness. It is shown that the strictness of the context definition correlates with the models' ability to identify semantically similar words: syntactic approaches perform better than bag-of-word models, and small context windows are better than larger ones. For semantic relatedness, however, syntactic features and small context windows are at a clear disadvantage. Second-order bag-of-word models perform below average across the board.

1. Introduction

Word Space Models have become the standard approach to the computational modelling of lexical semantics (Landauer & Dumais 1997, Lin 1998, Schütze 1998, Padó & Lapata 2007). They indeed offer a convenient way of capturing the meaning of a word simply on the basis of the contexts in which it is used in a corpus. In that way, they can retrieve the most similar words for a given target word. Yet, there is no agreement on how context should be defined exactly. Context features vary from sentences or paragraphs to single words, with or without the addition of syntactic relations. While all these features definitely capture *some* semantic information, it is only to be expected that the choice of context definition has an influence on the kind of semantic relatives that the Word Space Models will find.

It is well known that words may be semantically related along a number of dimensions (Cruse 1986). In the NLP literature, similarity takes up a central position, with synonymy as the most obvious example. But there are other types of semantic relations, too. For instance, two words like *doctor* and *hospital* have a clear connection, although they are in no way semantically similar. Recovering this semantic *relatedness* from a corpus may have to proceed along different lines than the modelling of semantic similarity. Specific Word Space Models may thus have a bias towards one or the other of these relations. In the literature, however, the investigation of this semantic behaviour of Word Space Models has only recently come to the fore (Sahlgren 2006, Peirsman, et al. 2007).

In this paper, we investigate eleven Word Space Models, representing three broad classes, with respect to their performance in the fields of semantic similarity and semantic relatedness. It will be shown that there is no such thing as a single best Word Space Model: the ranking of the approaches depends on the type of semantic information we want to find. The paper is structured as follows: in the next section, we will introduce the different context models and the two types of semantic relationship that we investigate. Section 3 then presents the precise setup of our experiments, while section 4 discusses their results. Section 5 wraps up with conclusions and an outlook for future research.

2. Word Space Models

2.1. Competing definitions of context

All Word Space Models of lexical semantics rely on the so-called *distributional hypothesis* (Harris 1954), which claims that words with similar meanings occur in similar contexts. From this hypothesis, it follows that semantic similarity can be modelled in terms of contextual or distributional similarity. This is done by constructing for each target word a so-called *context vector*, which contains the scores of its target word for all possible context features. These scores can be the number of times that the contextual feature co-occurs with the target, or more often, some kind of weighted frequency that captures the statistical link between the target word and that feature. The distributional similarity between two words is then calculated as the similarity between their vectors, on the basis of a function like the cosine. In this way, it is possible to find for each target word the n most distributionally similar words in any given corpus. We call these words the *nearest neighbours* of the target.

Based on the definition of context, it is possible to define a hierarchy of Word Space Models, each with its own kind of contextual features. At the top of the tree we make a distinction between document-based and word-based approaches. Document-based models use sentences, paragraphs or documents as dimensions, and count how often a target word appears in each of these entities in the corpus (Landauer & Dumais 1997, Sahlgren 2006). Word-based models, by contrast, take not the context itself, but features from this context as dimensions. They can be subdivided into syntactic and bag-of-word models. So-called bag-of-word or co-occurrence models take into account all words within a pre-defined distance of the target word (generally with the exception of semantically empty words like articles, etc.), whereas syntactic models consider only those words to which the target is syntactically related. Sometimes the features of such syntactic models consist of these syntactically related words alone (Padó & Lapata 2007), sometimes they are formed by the word plus its relation (Lin 1998). Finally we can distinguish between first-order and second-order approaches. First-order bag-of-word approaches count the context words directly (Levy & Bullinaria 2001), while second-order bag-of-word approaches sum the vectors of these context words. In this last case, the target's context vector thus contains frequency information about the context words of its (first-order) context words (Schütze 1998). Although it is in principle possible to construct second-order syntactic models, to our knowledge no implementation has been presented in the literature.

2.2. Semantic similarity and semantic relatedness

While it is claimed that all Word Space Models capture some kind of semantic information, so far we have only very limited knowledge about the influence of the context definition on the types of semantic relationship that the models find. In this paper we investigate two such types: semantic similarity and semantic relatedness. The first applies to synonyms (e.g., *plane* and *airplane*), hyponyms and hypernyms (e.g., *bird* and *blackbird*) and co-hyponyms (e.g., *blackbird* and *robin*) — two words with a relationship of similarity between the concepts they refer to. *Semantic relatedness*, by contrast, exists between words whose concepts are not necessarily similar, but still related, for instance because they belong to the same script, frame or lexical field. This is true for pairs like *bird* and *beak* or *plane* and *pilot*. Note that it is not possible to draw a clear boundary be-

tween semantic similarity and semantic relatedness. Take the word pair *pepper–salt*, for instance. These two words are clearly semantically similar, since they both refer to spices. At the same time, however, they are also semantically related: not only do they both belong to the lexical fields of food or spices, they also often co-occur together in the phrase *salt and pepper*. Instead of mutually exclusive classes, semantic similarity and relatedness can thus better be thought of as the two ends of a continuum, or two perpendicular axes in a two-dimensional plane.

For many NLP applications, similarity might be the most important relation to model. In typical Query Expansion, for instance, only semantically similar words (synonyms or possibly hyponyms) make for a desired extension of a search query. Similarly, in Question Answering a word in the question should only be matched with semantically similar words in the database where the computer looks for the answer. Semantic similarity, however, is just one way in which words may be related in our mental lexicon, as suggested by psycholinguistic association experiments. According to Aitchinson (2003), the four major types of associations that people give in response to a cue word are, in order of frequency, co-ordination (co-hyponyms like *pepper* and *salt*), collocation (like *salt* and *water*), superordination (hypernyms like *butterfly* and *insect*) and synonymy (like *starved* and *hungry*). A similar observation is made by Schulte im Walde & Melinger (2005). Comparing the results of their German verb association experiment with GermaNet, they note that only 6% of the associations are synonyms, 14% are hypernyms and 16% are hyponyms, while no less than 54% of the associations are unrelated to their cue words in the GermaNet taxonomy. Although part of this can be explained by the incompleteness of the database, such results will be difficult to replicate with models of semantic similarity. After all, these are meant to prefer synonyms over hypernyms and co-hyponyms, and even exclude collocates altogether. The best Word Space Models of semantic similarity may thus not be the best models of relatedness, and vice versa.

Despite the wealth of research into Word Space Models, studies into their semantic characteristics are scarce. Most often one model is applied to a specific computational-linguistic task, and "comparisons between the (...) models have been few and far between in the literature" (Padó & Lapata 2007, p. 166). Sahlgren (2006) is one exception to this rule. Focusing on document-based and first-order bag-of-word models, he showed that the latter are better geared towards the modelling of paradigmatic (similarity) relations, while the former have a clear bias towards syntagmatic relations. Unfortunately, Sahlgren left out a number of popular word space approaches, like those based on syntactic relations or second-order co-occurrences. Peirsman et al. (2007) also included syntactic models, but concentrated on similarity relations only. This article thus sets out to fill these gaps in the literature, by discussing a wide variety of model types from the perspectives of similarity as well as relatedness.

3. Experimental setup

We investigate three classes of Word Space Models, for a total of eleven approaches: five first-order bag-of-word models, five second-order bag-of-word models and one syntactic model. Our corpus is the 300 million word Twente Nieuws Corpus of Dutch newspaper articles, collected at the University of Twente and parsed by the Alpino parser at the University of Groningen. As our test set, we selected from this corpus the 10,000 most

frequent nouns. For each of these, we had all models retrieve the 100 most similar neighbours from the 9,999 remaining nouns in the set.

The bag-of-word models, both first-order and second-order, varied the size of the context window they took into account — 1, 3, 5, 10 or 20 words to either side of the target — for a total of ten models. Sentence boundaries were ignored; article boundaries were not. The syntactic model considered eight different types of syntactic dependency relations, in which the target word could be (1) the subject of verb v, (2) the direct object of verb v, (3) a prepositional complement of verb v introduced by preposition p, (4) the head of an adverbial prepositional phrase (PP) of verb v introduced by preposition p, (5) modified by adjective a, (6) postmodified by a PP with head n introduced by preposition p, (7) modified by an apposition with head n, or (8) coordinated with head n. Each specific instantiation of the variables v, p, a, or n was responsible for a new context feature.

The other parameter settings were shared by all eleven models:

- *Dimensionality:* For all approaches, we used the 2,000 most frequent contextual features in the corpus as dimensions. This is a simple but common way of reducing the otherwise huge dimensionality of the vectors, which leads to state-of-the-art results, particularly for the syntactic model (Levy & Bullinaria 2001, Padó & Lapata 2007). For the syntactic model these dimensions are the 2,000 most frequent syntactic features, like `subj_of_fly`. For the bag-of-word models, they are formed by the 2,000 most frequent words in the corpus. Function words and other semantically empty words were excluded a priori on the basis of a stop list.

- *Frequency cut-off:* Depending on the context size, we established a cut-off value n, so that the models ignored those features that occurred together with the target fewer than n times. For context size 3, this cut-off was fixed at 3, for the larger context sizes it lay at 5. The syntactic model and the bag-of-word model with context size 1 did not use a cut-off, since it led to data sparseness.

- *Frequency weighting:* As is usual in the literature, the context vectors of the target words did not contain the simple frequencies of the features. Instead, they listed the point-wise mutual information between each feature and the target word. This measure expresses whether the two occur together more or less often in the corpus than we expect on the basis of their individual relative frequencies.

- *Similarity measure:* Finally, the distributional similarity between two target words was measured by the cosine between their context vectors.

4. Results

4.1. Semantic similarity

We evaluated the ability of our models to find semantically similar words on the basis of a comparison with Dutch EuroWordNet (Vossen 1998). This lexical database contains more than 34,000 sets of noun synonyms and the relations that exist between them. Two evaluation measures were applied. First, we focused on the general ability of our models to capture semantic similarity. Then we looked into the distribution of four more specific similarity relations.

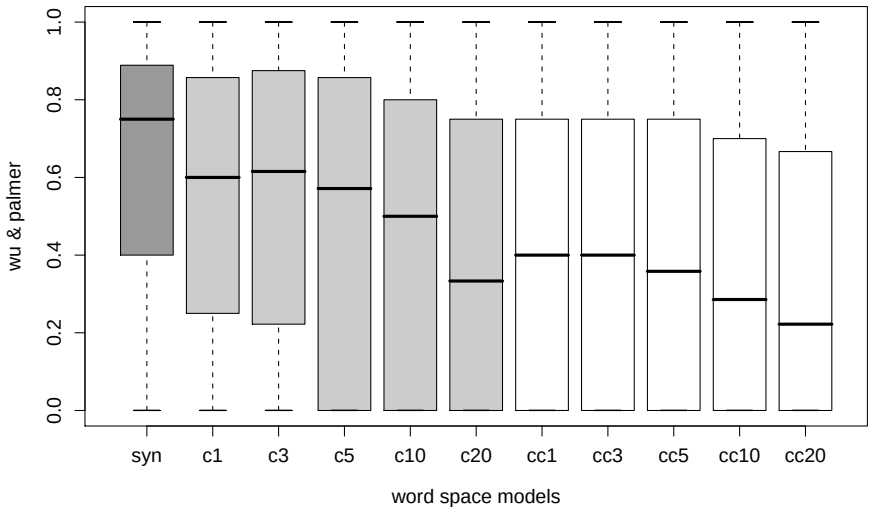

Figure 1: Wu & Palmer similarity scores between target and nearest neighbour.
syn: syntactic model, cn: first-order bag-of-words, cc$n$: second-order bag-of-words
n: context size (number of words on either side of target)

The general performance of the models was quantified by the average Wu and Palmer score between a target word and its single nearest neighbour (Wu & Palmer 1994). This Wu and Palmer score is a popular way of measuring the semantic similarity between two words, based on their depth and their distance from each other in a taxonomic structure like EuroWordNet. If either the target or its nearest neighbour were not present in the database, the pair was simply ignored. In order to make the results perfectly comparable across models, we restricted the results to the 4183 target words with a nearest neighbour in EuroWordNet for all models. The resulting Wu and Palmer scores are given in Figure 1.

Figure 1 shows a clear decrease in Wu and Palmer score as the definition of context becomes less strict. A Friedman test indeed confirms the influence of the type of Word Space Model on performance (Friedman chi-squared = 3541.575, df = 10, p-value $<$ $.001$). The syntactic model achieves the highest average similarity score by far, followed by the first-order bag-of-word models and finally the second-order bag-of-word models. Moreover, small contexts appear to model semantic similarity better than large ones. A test of multiple comparisons after Friedman showed that the differences between all pairs of models are indeed statistically significant at the $.05$ level, except for those between context sizes 1 and 3 (both first-order and second-order) and that between the first-order model with context size 20 and the second-order model with context size 5.

Of course, this general similarity score does not give any information about what specific type of similarity relation the models find. We therefore defined four taxonomic similarity relations, again with EuroWordNet as a gold standard. *Synonyms* were defined as words in the same synonym set as the target word, *hypernyms* as words exactly one node above the target, *hyponyms* those one node below and *co-hyponyms* as words one node below any of the target's hypernyms. Together, these relations make up the target's *EuroWordNet environment*. Note that our strict definition of these relationships does not

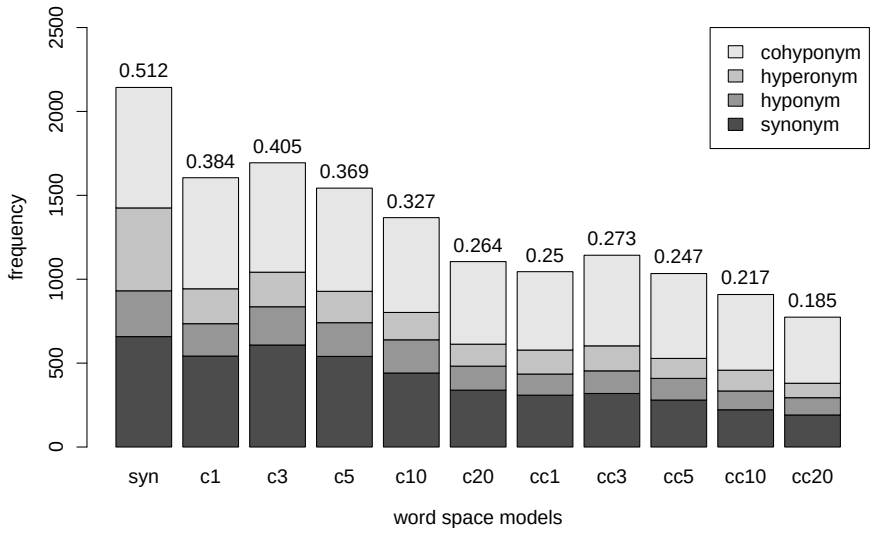

Figure 2: Distribution of semantic similarity relations for all models.

allow for more than one or two steps in the tree, and thus disregards possible hypernyms or hyponyms that are more than one step away from the target. This approach ensures the reliability of our gold standard, but constitutes a test that a relatively low percentage of nearest neighbours will pass. Figure 2 shows how the single nearest neighbours of our target words are distributed over the four similarity relations. Again we restricted ourselves to the 4183 target words with a neighbour in EuroWordNet for all models.

Not surprisingly, the number of retrieved similarity relations mirrors the general Wu and Palmer similarity score. Again the syntactic model performs best: 51.2% of its single nearest neighbours that occur in EuroWordNet are situated in the environment of the target word. This precision drops to between 40.5% and 26.4% for the first-order bag-of-word methods and even lower for the second-order models. As above, the performance of the models seems to depend on the strictness of their context definition. The stricter they view context — i.e., syntactic context rather than a bag of words, smaller context windows rather than large ones — the more examples of semantic similarity they find. This pattern remains unchanged when a larger number of nearest neighbours is taken into account.

With one exception, the distribution of the four relations is comparable across the different models. Co-hyponyms figure most prominently among the nearest neighbours, followed by synonyms, hypernyms and hyponyms. Only the syntactic model finds an unexpectedly high number of hypernyms. This can probably be explained by the way syntactic relations are typically inherited in a taxonomy: all characteristics of a (prototypical) concept (*can fly*, for instance) also apply to its hypernyms, so that these are often most similar in terms of syntactic distribution in a corpus.

4.2. Semantic relatedness

The results in the previous section do not necessarily express the overall quality of the investigated Word Space Models. It is possible that the models that scored relatively badly

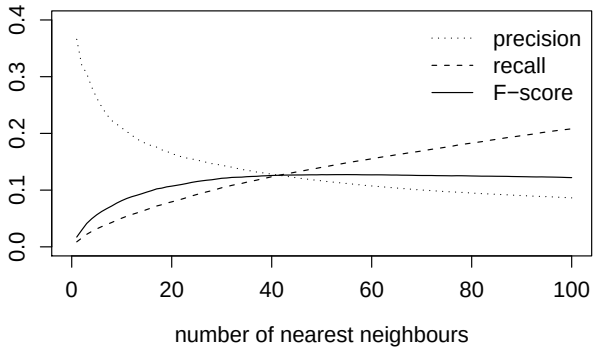

Figure 3: Evolution of the precision, recall and F-score of the first-order bag-of-word model with context size 5 in its retrieval of associations.

in the similarity experiments are simply biased towards a different kind of semantic relation. In this second round of experiments we therefore turn our attention from semantic similarity to semantic relatedness.

For this task, we relied on a psycholinguistic experiment of human associations, described in De Deyne & Storms (in press). In this experiment, participants were asked to list three different word associations for 1,424 cue words. Each word was presented to at least 82 participants, resulting in a total of 381,909 responses. For instance, *aap* ('monkey') triggered the response *zoo* ('zoo') 27 times, *aarde* ('earth') prompted *planeet* ('planet') 14 times and *bikini* ('bikini') elicited *vakantie* ('holiday') 6 times. These examples show that this experiment taps into a different kind of semantic relationship than the previous one. Note that at this moment, we ignore the fact that association strength is often asymmetric (Michelbacher, et al. 2007).

In order to make the results comparable to those in section 4.1., we reduced the data set to those cue words and associations that belong to the 10,000 most frequent nouns in our corpus. This gave a gold standard of 768 cue words with a total of 31,862 different cue–association pairs. When these associations are checked against EuroWordNet, we indeed find that only 8% belong to the EuroWordNet environment of their target word. 9% of these are synonyms, 19% are hypernyms, 16% are hyponyms and 56% are cohyponyms.

We evaluated the Word Space Models against this gold standard by counting the number of associations that they find as the nearest neighbours to the cue words. If we consider just one nearest neighbour, the results already show a considerable difference from the previous experiments. As the top chart in Figure 4 indicates, the syntactic model still performs best, with 340 associations (a precision of .443), followed by the first-order and then the second-order bag-of-word models. However, within the bag-of-word models, the ideal context size has changed. The first-order bag-of-word models with context sizes 10 and 20 have 299 and 293 associations among their single nearest neighbours, respectively. For 768 targets, this gives precision values of .389 and .382. Then we find context sizes 5 ($n = 281$, $P = .366$), 3 ($n = 269$, $P = .350$) and 1 ($n = 228$, $P = .297$). Larger contexts thus outperform their smaller competitors here. Note that the two best models share only

90 correct predictions, which indicates that they have different preferences among the associations. A look at the data suggests that the syntactic model indeed picks out those associations that are also semantically similar to their target word, while the first-order bag-of-word models with large contexts cover collocational relatedness better. With the second-order models, finally, context size 3 seems optimal.

When we consider one nearest neighbour, the models cannot find more than 768 associations, and recall thus stays extremely low. We therefore increased the number of nearest neighbours from 1 to 100 and calculated the precision, recall and F-score at each step. By way of example, Figure 3 plots the evolution of these values for the best-performing model. The bottom bar chart in Figure 4, then, shows the maximum F-score of all the models. The syntactic approach has lost its lead, which suggests that it is able to model only a small number of associations well — probably those that also score highly on similarity. Instead it is now the first-order bag-of-word model with context size 5 that outclasses all others, with an F-score of .127 ($P = .112$, $R = .148$) at 55 neighbours. Extending the context window to 10 words brings the F-score down to .122 ($P = .102$, $R = .150$, 61 neighbours); reducing the window to 3 words takes it to .120 ($P = .104$, $R = .143$, 57 neighbours). Next, we have the bag-of-word model with context size 20 ($F = .115$, $P = .102$, $R = .133$, 54 neighbours) and only then the syntactic model ($F = .111$, $P = .102$ $R = .123$, 50 neighbours). Large contexts now score slightly worse than intermediate ones, which probably strike the best balance between similarity relations and collocational links. Second-order models never attain an F-score above .10, and neither do the smallest context windows, which are thus clearly biased towards similarity.

4.3. Discussion

In part, our experiments have confirmed earlier results in the literature. For instance, Sahlgren (2006) already noted that with first-order bag-of-word models, larger contexts score better in his association experiment, while smaller contexts score better in the synonymy test. Peirsman et al. (2007) found even better results for a syntactic model in Dutch, at least with respect to semantic similarity evaluated against EuroWordNet. Both findings are borne out by our experiments.

At the same time, our results add some new insights to these earlier observations. We have shown that the syntactic model and the bag-of-word models with context size 1 are most biased towards semantic similarity. The syntactic model scored best in our first round of experiments, while the results of the bag-of-word models with context size 1 were either not statistically different from or better than those of models with larger context windows. When it came to the discovery of semantic associations, however, context size 1 proved the least advisable choice, and the syntactic model was outperformed by all first-order bag-of-word models with an intermediate or large context window. Second-order bag-of-word models scored below average in both experiments. They probably only show their power when data sparseness is an issue, as with Word Sense Discrimination (Schütze 1998) or with corpora smaller than ours.

5. Conclusions and future research

In this paper, we investigated the influence of the context definition on the ability of several Word Space Models to capture two kinds of semantic information — semantic

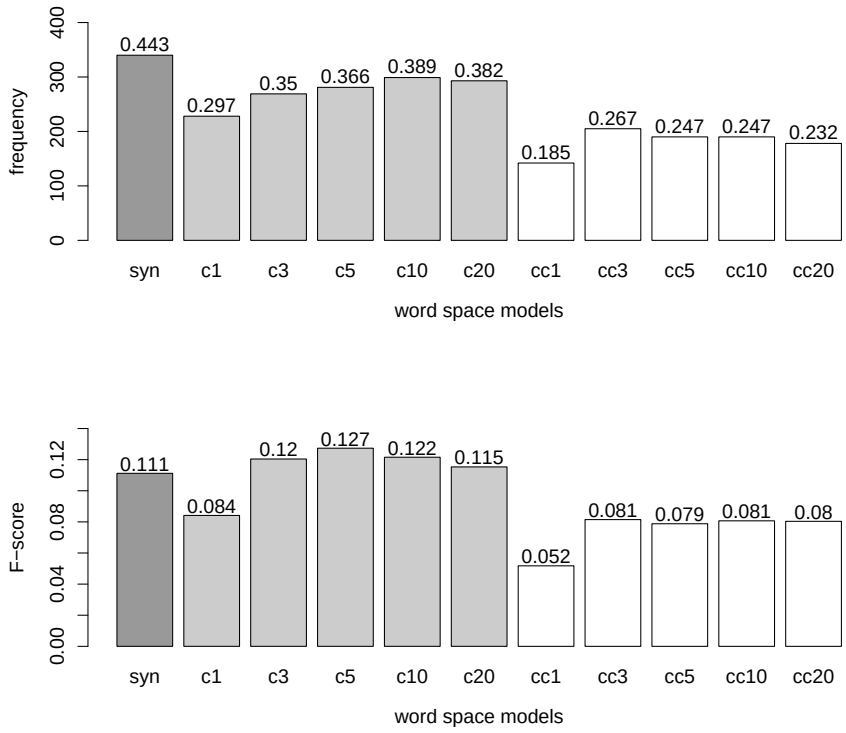

Figure 4: Frequency of associations among single nearest neighbours (top) and maximal F-scores for all models (bottom).

similarity and semantic relatedness. We studied a total of eleven Word Space Models: one syntactic approach and ten bag-of-word models with context sizes 1, 3, 5, 10 and 20, first-order as well as second-order. Both for semantic similarity and semantic relatedness, first-order models clearly beat their second-order competitors. However, while syntactic models gave the best results for semantic similarity, first-order bag-of-word approaches with intermediate to large context windows fared better in the retrieval of associated words.

In the short term, we aim to extend the repository of Word Space Models that we are investigating — document-based models and second-order syntactic models are particularly high on our list. In the longer term, we will try and determine if the differences we observed in the modelling of semantic relations between word types also play a role in Word Sense Discrimination. In this task, all contexts of a word are clustered in order to automatically find the multiple senses of that word. Given the results here, we suspect that different kinds of polysemy or homonymy may not demand the same context definitions.

References

J. Aitchinson (2003). *Words in the Mind. An Introduction to the Mental Lexicon*. Oxford: Blackwell.

D. A. Cruse (1986). *Lexical Semantics*. London: Cambridge University Press.

S. De Deyne & G. Storms (in press). 'Word Associations: Norms for 1,424 Dutch Words in a Continuous Task'. *Behaviour Research Methods* .

Z. Harris (1954). 'Distributional Structure'. *Word* **10**(23):146–162.

T. K. Landauer & S. T. Dumais (1997). 'A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge'. *Psychological Review* **104**:211–240.

J. P. Levy & J. A. Bullinaria (2001). 'Learning Lexical Properties from Word Usage Patterns: Which Context Words Should Be Used'. In R. French & J. Sougne (eds.), *Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop*, pp. 273–282. London: Springer.

D. Lin (1998). 'Automatic Retrieval and Clustering of Similar Words'. In *Proceedings of COLING-ACL98*, pp. 768–774, Montreal, Canada.

L. Michelbacher, et al. (2007). 'Asymmetric Association Measures'. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-07)*, Borovets, Bulgaria.

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

Y. Peirsman, et al. (2007). 'Finding Semantically Related Words in Dutch. Co-occurrences versus Syntactic Contexts'. In *Proceedings of the CoSMO Workshop*, pp. 9–16, Roskilde, Denmark.

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 University.

S. Schulte im Walde & A. Melinger (2005). 'Identifying Semantic Relations and Functional Properties of Human Verb Associations'. In *Proceedings of the joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 612–619, Vancouver, Canada.

H. Schütze (1998). 'Automatic Word Sense Discrimination'. *Computational Linguistics* **24**(1):97–124.

P. Vossen (ed.) (1998). *EuroWordNet: a Multilingual Database with Lexical Semantic Networks for European Languages*. Dordrecht: Kluwer.

Z. Wu & M. Palmer (1994). 'Verb Semantics and Lexical Selection'. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pp. 133–138, Las Cruces, NM.