Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts.

Yves Peirsman[1,2], Kris Heylen[1] and Dirk Speelman[1]

[1] Quantitative Lexicology and Variational Linguistics (QLVL),
University of Leuven, Belgium
[2] Research Foundation – Flanders
yves.peirsman@arts.kuleuven.be
kris.heylen@arts.kuleuven.be
dirk.speelman@arts.kuleuven.be

Abstract. Despite the growing interest in distributional approaches to semantic similarity, the linguistic differences between their results are still unclear. In this paper we use six vector-based techniques to retrieve semantically related nouns from a corpus of Dutch, and investigate them from a computational as well as a linguistic perspective. In particular, we compare the results of a bag-of-word model with a syntactic model, and experiment with different context sizes and means of reducing the dimensionality of the vector space. We find that a full syntactic context model clearly outperforms all other approaches, both in its overall performance as in the proportion of synonyms it discovers.

1 Introduction

The automatic discovery of semantic similarity between nouns on the basis of a corpus has been a hot topic in recent years. It has applications in question answering, information retrieval, thesaurus extraction, parsing, and many other computational-linguistic tasks. Although in theory, electronic resources can be used for the extraction of semantically similar words, these are often incomplete or even unavailable for the language or domain at hand. The most popular automatic approaches are vector-based algorithms that rely on a large corpus to construct a vector with information about the contexts a word occurs in [1, 2]. They are based on the so-called *distributional hypothesis* [3], which states that words that appear in similar contexts will have related meanings. The semantic distance or similarity between two words can then be captured by the distance between their respective vectors.

Those computational-linguistic techniques can also be fruitfully exploited in theoretically oriented research. This paper reports on research carried out within the sem·metrix project, which aims to quantify the differences in word use between different varieties of Dutch. In particular, it investigates what lexemes are used to refer to a large number of concepts and how frequently those occur in the different varieties. In previous research [4], the sets of synonyms for each concept were constructed manually. As a result, the scope of the investigation had to be rather limited. In order to extend this research to a larger number of concepts, the sem·metrix project now makes use of NLP approaches to retrieve semantically related words automatically.

The distributional methods described above allow for many different specific implementations. While their performance on a particular task is being investigated extensively, much less attention goes to the linguistic characteristics of their results. We therefore complement a study of the overall performance of the models with an investigation of the types of semantic relation they discover. With this goal in mind, we varied three of the models' main parameters: (1) the type of context features retrieved from the corpus, (2) the size of the context window and (3) the use of a dimensionality reduction technique.

After a discussion of the setup of our experiments in section 2, section 3 is concerned with the interpretation of the results. Section 4, finally, draws the main conclusions and wraps up with perspectives for future work.

2 Experimental setup

2.1 Types of context features

Broadly speaking, vector-based approaches to semantics witness a competition between two types of context features. First, in a *bag-of-word* or *co-occurrence* model, the vector of a target word records what words occur within a context window of n words on either side of the target. It often contains the frequencies of such co-occurrences, or a figure that expresses if the context word is found more often than expected. Such a bag-of-word model was used by Schütze [1] in his landmark paper in the field.[3] It is the first type of context model that we will investigate.

Second, in a *syntactic model*, the vector of a target word contains information about the words with which the target occurs in a syntactic relationship. This approach was taken by Lin [5] and Curran and Moens [6] for English, and by Van der Plas and Bouma [7] and Van de Cruys [8] for Dutch. Padó and Lapata [2] compared the results of such a syntactic model with a bag-of-word approach, and found the syntactic model to be superior.

These syntactic models still allow much freedom in the type of syntactic relations that are considered. Our implementation took into account eight different types of syntactic dependency relations, in which the target noun could be

- subject of verb v,
- direct object of verb v,
- prepositional complement of verb v introduced by preposition p,
- the head of an adverbial prepositional phrase (PP) of verb v introduced by preposition p,
- modified by adjective a,

[3] It should be noted, however, that Schütze worked with so-called *second-order* co-occurrences, which model a word in terms of the context words of its context words.

- postmodified by a PP with head n, introduced by preposition p,
- modified by an apposition with head n, or
- coordinated with head n.

Each specific instantiation of the variables v, p, a, or n was responsible for a new context feature.

2.2 Size of the context window

In a bag-of-word model, it is also possible to vary the size of the context window around the target word. While a technique like LSA [9] looks at words in the entire paragraph, Schütze [1] defined a context as twenty-five words on either side of the target, and Lund and Burgess [10] worked with fewer than ten words. We experimented with two context sizes: one with five words on either side of the target, one with fifty. For the fifty-word context window, we kept the dimensionality of the vectors relatively low, by only treating the 2,000 most frequent words in the corpus as possible context words, similar to Padó and Lapata [2]. For the five-word window, we experimented with this setup as well as with the full dimensionality.

2.3 Random Indexing

The feature vectors that can be obtained from a 300 million word corpus are massive, and the resulting computation cost is obviously very high. A number of techniques have been developed to deal with this computational inefficiency. One of those is Random Indexing (RI) [11]. RI may be less popular than Singular Value Decomposition (SVD) [12], but it has the advantage of bypassing the construction of an enormous co-occurrency matrix. While SVD reduces the dimensionality of the matrix after it has been made, RI does this during its construction.

Random Indexing creates a so-called *index vector* for each contextual feature, whose length is much smaller than the total number of contextual features. This vector contains a large number of 0s, and a small number of randomly distributed +1s and −1s. The context vector of a word is then constructed by summing the index vectors of all its contextual features. During this process, each index vector can optionally be weighted according to the statistical behaviour of its feature and the current target [13]. For our experiments, however, we implemented basic Random Indexing, which simply weights each index vector by the frequency of its feature in the context of the target word. We used index vectors of length 1,800 with eight non-zero elements.

2.4 Other parameters and evaluation

The rest of the setup was left identical for all our experiments. As our data we used the parsed and lemmatized Twente Nieuws Corpus (TwNC), a 300 million word corpus of Dutch newspaper articles from 1999 to 2002. Our word vectors did not contain the frequency of the context features, but their point-wise mutual information (PMI) with the target word. This statistic quantifies if the combination *(target word, context feature)* appears more often than expected on the basis of their individual frequencies in the corpus. Informative context features thus receive a higher value. The usefulness of this statistic was demonstrated by Van der Plas and Bouma [7]. Semantically empty words were automatically ignored (on the basis of a stoplist), as were co-occurrences with a frequency of less than five. For the syntactic features, we did not use a frequency cutoff.

For the evaluation of the algorithms, we randomly sampled 1,000 nouns from the corpus, making sure that they had an absolute frequency of at least 50 and also appeared in Dutch EuroWordNet. The former requirement was meant to sidestep data sparseness, while the latter ensured the possibility of automatic evaluation. Twenty-four of the 1,000 sampled words turned out not to have any features with a frequency of five or more for at least one of our models. This brought the final number of test words down to 976. Like Schütze [1], we used the cosine measure to find for each word its ten most similar words in the corpus. Here, too, we only looked at words with an absolute frequency of at least fifty.

3 Results and discussion

We compared a total of six models. First, we looked at the overall performance of each of the approaches, by investigating if the ten words that they returned are indeed semantically related to their targets. Second, we homed in on the semantic relations that the algorithms discover — synonyms, hypernyms, hyponyms or co-hyponyms — in order to find whether a certain model has a preference for a specific syntactic relation. Both types of evaluation used Dutch EuroWordNet [14] as a gold standard.[4]

3.1 Performance

For each of our 976 target words, the algorithm returned the ten most similar words in the corpus. On the basis of Dutch EuroWordNet, we computed the Wu & Palmer similarity score between each of these words and its target. This score captures the similarity between two words w_1 and w_2 on the basis of their relation in a lexical hierarchy [15]. In particular, it finds their lowest shared hypernym, h_l, and divides twice the depth of this hypernym in the hierarchy by the sum of the depths of w_1 and w_2:

$$s_{WP}(w_1, w_2) = \frac{2 \times depth(h_l)}{len(w_1, h_l) + len(w_2, h_l) + 2 \times depth(h_l)} \tag{1}$$

[4] We should add a word of caution here, since the coverage of Dutch EuroWordNet is not as high as that of its English counterpart. On average, EuroWordNet contains between 5.7 and 7.4 of the ten most related words that the algorithm finds. Despite this coverage problem, EuroWordNet is still the most common and straightforward means of evaluation for distributional algorithms.

	average Wu & Palmer scores		
	syntactic model	bag-of-word model	
		5-word window	50-word window
all features	0.48	0.34	—
Random Indexing	0.31	0.26	—
2,000 most frequent features	—	0.28	0.23

Table 1. Average Wu & Palmer similarity score for the ten most related words.

If a word in our output did not occur in EuroWordNet, it was simply ignored. If it appeared several times, the maximum Wu & Palmer score was taken.[5]

Table 1 sums up the average Wu & Palmer scores for the six models. These results show three important patterns. First, the syntactic context model performs better than the bag-of-word approach. Syntactic context features thus allow us to find words that are more closely related to the target word. Second, drastically reducing the dimensionality of the vector space (either by Random Indexing or by a cutoff at the 2,000 most frequent words) brings down performance considerably. This indicates that the algorithm had better take into account as much information contained in the corpus as possible. Third, with 2,000 features, the fifty-word context model performs less well than the five-word one. We believe this is the case because the fifty-word model finds more *loose* semantic associations (e.g. *car* – *road*), while the five-word model discovers more *tight* semantic relations (e.g. *car* – *vehicle*). This hypothesis follows from the fact that a narrow context will contain a higher proportion of words that are also syntactically related to the target word, and hence, more linguistic information about the target. This deserves more careful investigation, however.

In order to see if the full syntactic and bag-of-word models score well on the same words, we calculated the correlation between the average Wu & Palmer scores for the best two models. In order to make the figures more robust, the rest of this section looks only at target words with five or more of their ten nearest neighbours in EuroWordNet (552 targets). Spearman's correlation statistic between the results was 69.9%. This suggests that the difficulty of a word is indeed fairly independent of the type of context features that the algorithm uses.

Still, there are differences between the models' behaviour. In particular, the full syntactic context model gives a correlation of 38.8% between the average Wu & Palmer similarity of the ten nearest neighbours and the depth of the target in the EuroWordNet hierarchy.[6] For the bag-of-word model, the same statistic is only 28.3%. Both models thus work better for words deeper in the hierarchy, but the performance of the bag-of-word model decays less with decreasing depth.

[5] In the case of a polysemous word, we are only interested in the meaning that is most related to the target word.

[6] If a word appeared several times in EuroWordNet, we used its minimum depth.

3.2 Semantic relations

The ultimate goal of our project is to retrieve words that have a specific semantic relation to the target word. We would thus like to gain more insight in the types of semantic relations that the algorithms find. We therefore checked against EuroWordNet what relation, if any, each of the 9760 retrieved words has with its target word. We took the following semantic relations into account:[7]

- synonymy: the retrieved word co-occurs in a synset together with the target.
- hyponymy: the retrieved word occurs in a synset that is a direct daughter of (one of) the target's synset(s).
- hypernymy: the retrieved word occurs in a synset that is the direct mother of (one of) the target's synset(s).
- cohyponymy: the retrieved word occurs in a synset that is a direct daughter of one of the target's hypernym synsets as defined above.

Table 2 gives the absolute and relative frequencies of the relations found with the different models, summing over all the target words. Because of the limited coverage of EuroWordNet, not all retrieved top ten similar words could be evaluated against the thesaurus. The percentages of related words are therefore relative to the number of retrieved words present in EuroWordNet (last column). As above, the syntactic context model outperforms all other approaches, both in the overall proportion of semantically related words (31.4%) as in the percentage of retrieved words with a specific semantic relation (e.g. 6.3% synonyms[8]). The other relations between the context models mirror those in the previous section.

Moreover, a chi-square test shows that there is an interaction between the type of context model and the frequency of the semantic relations ($\chi^2 = 49.65$, $df = 15$, $p < 0.001$). Table 3 gives the differences between the observed frequencies of the semantic relations and their expected frequencies. It shows that the full syntactic model returns significantly more synonyms and hyponyms than expected under independence, but far fewer cohyponyms. Interestingly, all models with a reduced dimensionality find fewer synonyms than expected, but more cohyponyms. In three of the four cases, this number of retrieved cohyponyms is significantly higher than expected. Severely reducing the dimensionality of the word vectors thus leads to a retrieval of more loosely related words. One of the reasons for this finding may lie in the informativeness of relatively low-frequent context features that are highly correlated with the occurrence of the target word. While these features receive a high PMI value in the full models, they are weighted by their frequency in the Random Indexing setup, or even simply

[7] Since our system did not disambiguate between the different senses of a word, all synsets in which a retrieved word or a target word appeared were taken into consideration. When a word had several relations with the target, only the closest relation was added to the tally, with synonymy > hyponymy > hypernymy > cohyponymy.

[8] The percentage of synonyms is rather low for all systems, but this might be partly due to the cut-off at ten most similar words. Not many words have ten synonyms, while the number of potential (co)hyponyms is often far larger.

	syno.	hypo.	hyper.	cohyp.	all 4	in EWN
syn	392 (6.3)	249 (4.0)	262 (4.2)	1050 (16.9)	1953 (31.4)	6215
bow, 5w	236 (4.2)	150 (2.7)	156 (2.8)	686 (12.2)	1228 (21.8)	5645
syn, RI	154 (2.1)	94 (1.3)	141 (1.9)	624 (8.6)	1013 (13.9)	7275
bow, 5w, 2000	144 (2.5)	106 (1.8)	99 (1.7)	561 (9.8)	910 (15.9)	5739
bow, 5w, RI	135 (2.4)	75 (1.3)	118 (2.1)	467 (8.3)	795 (14.2)	5598
bow, 50w, 2000	106 (1.6)	60 (0.9)	76 (1.1)	416 (6.2)	658 (9.8)	6682

Table 2. Relations found among the ten most related words, if present in EuroWordNet, summed over all target words (percentage of row totals between brackets).

syn, bow:	syntactic/bag-of-word context model
5w,50w:	5/50-word context window
RI:	Random Indexing
2000:	2,000 most frequent words in corpus as context words

	syno.	hypo.	hyper.	cohyp.
syn	**44.41**	**30.38**	8.23	**-83.02**
bow,5w	17.44	12.54	-3.56	-26.42
syn,RI	**-26.29**	**-19.40**	9.37	**36.31**
bow,5w,2000	-17.96	4.13	**-19.24**	**33.07**
bow,5w,RI	-6.49	-13.99	14.70	5.79
bow,50w,2000	-11.11	-13.66	-9.50	**34.27**

Table 3. Differences between observed frequencies and expected frequencies on the basis of a chi-square test for Table 2. Significant differences are shown in bold.

ignored when only the 2,000 most frequent corpus words are taken into account. When the algorithm is to find tight semantic relations like synonymy, the statistical relationship between a target word and each of its features may thus play an important role.

4 Conclusions and future work

In this paper, we have compared six distributional approaches that find the most similar words for any given target on the basis of a corpus. In particular, we have contrasted the impact of three parameters: (1) the type of context feature (co-occurrences vs. syntactic features), (2) the size of the context window (five vs. fifty context words) and (3) the reduction of the dimensionality, either through Random Indexing or a frequency cutoff.

The full syntactic context model outperformed all other combinations, both in overall performance as in the number of synonyms it finds. Drastically reducing the dimensionality of the vector space brings down performance substantially, as does enlarging the context window of our bag-of-word model from five to fifty words. An analysis of the retrieved semantic relations showed that the full models are more sensitive to synonymy relations, while those with reduced dimensionality are biased towards co-hyponyms.

In the future, we would like to experiment with more parameter settings and more advanced context models. In particular, we would like to investigate second-order co-occurrences [1] or indirect syntactic relations [2]. For all of these possible extensions, we are particularly interested in the linguistic implications of the different models: the types of words they are suited for, and the kind of semantic relations that they discover.

References

1. Schütze, H.: Automatic word sense discrimination. Computational Linguistics **24**(1) (1998) 97–124
2. Padó, S., Lapata, M.: Dependency-based construction of semantic space models. Computational Linguistics **33**(2) (2007) 161–199
3. Harris, Z., ed.: Mathematical Structures of Language. New York: Wiley (1968)
4. Geeraerts, D., Grondelaers, S., Speelman, D.: Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen. Meertens Instituut, Amsterdam (1999)
5. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of COLING-ACL98, Montreal, Canada (1998) 768–774
6. Curran, J.R., Moens, M.: Improvements in automatic thesaurus extraction. In: Proceedings of the Workshop on Unsupervised Lexical Acquisition (SIGLEX), Philadelphia, PA, USA (2002)
7. Van der Plas, L., Bouma, G.: Syntactic contexts for finding semantically related words. In: Proceedings of Computational Linguistics in the Netherlands 15. (2005) 173–186
8. Van de Cruys, T.: The application of Singular Value Decomposition to Dutch noun-adjective matrices. In: Actes de la 13e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), Leuven, Belgium (2006) 767–772
9. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review **104** (1997) 211–240
10. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments and Computers **27** (1996) 203–208
11. Kanerva, P., Kristoferson, J., Holst, A.: Random Indexing of text samples for Latent Semantic Analysis. In: Proceedings of the 22nd Annual Conference of the Cognitive Science Society, Philadelphia, PA, USA (2000) 1036
12. Golub, G.H., Van Loan, C.F.: Matrix Computations. London: The Johns Hopkins University Press (1989)
13. Gorman, J., Curran, J.R.: Random indexing using statistical weight functions. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), Sydney, Australia (2006) 457–464
14. Vossen, P., ed.: EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer, Dordrecht (1998)
15. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd Annual meeting of the Association for Computational Linguistics, Las Cruces, NM (1994) 133–138