Automatic Identification of Synonym Relations in the Dutch Parliament's Thesaurus

Rosa Tsegaye Aga and Christian Wartena and Otto Lange and Nelleke Aders

Abstract For indexing archived documents the Dutch Parliament uses a specialized thesaurus. For good results for full text retrieval and automatic classification it turns out to be important to add more synonyms to the existing thesaurus terms. In the present work we investigate the possibilities to find synonyms for terms of the parliaments thesaurus automatically. We propose to use distributional similarity (DS). In an experiment with pairs of synonyms and non-synonyms we train and test a classifier using distributional similarity and string similarity. Using ten-fold cross validation we were able to classify 75% of the pairs of a set of 6000 word pairs correctly.

Rosa Tsegaye Aga
Hochschule Hannover, Expo Plaza 12, D-30539 Hannover
✉ rosa-tsegaye.aga@hs-hannover.de

Christian Wartena
Hochschule Hannover, Expo Plaza 12, D-30539 Hannover
✉ christian.wartena@hs-hannover.de

Otto Lange
Universiteitsbibliotheek Utrecht, Heidelberglaan 3, Postbus 80124, NL- 3508 TC Utrecht
✉ o.a.lange@uu.nl

Nelleke Aders
Tweede Kamer der Staten-Generaal, Postbus 20018, NL-2500 EA Den Haag
✉ N.aders@tweedekamer.nl

ARCHIVES OF DATA SCIENCE, SERIES A
(ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. 2, No. 1, 2017

DOI 10.5445/KSP/1000058749/23
ISSN 2363-9881

1 Introduction

The Information Service of the Second Chamber of the Dutch parliament (*Tweede Kamer der Staten-Generaal*) archives and indexes documents produced in the parliamentary process and other documents that are possibly relevant to the parliament. For indexing a special thesaurus covering all topics relevant to the society and the parliament is used. The Dutch parliament is investigating additional alternatives to make information available, by full text search, automatic indexing and automatic classification. For good results for full text retrieval and automatic classification, it turns out to be important to add more synonyms to the existing thesaurus terms.

In the present work we investigate the possibilities to find synonyms for terms of the parliaments thesaurus automatically. To do this, we propose to use distributional similarity (DS). DS assigns small distances to terms that are "semantically related". However, this relatedness does not correspond to any traditional semantic relation like synonymy. Terms related by DS might be synonyms, but antonyms or other related words as well. On the other hand, it is unclear as well what the exact semantic relation is between terms that are both labels for the same concept or that stand in a *use/use-for* relation to each other. In order to see whether DS can be used to distinguish between unrelated terms and terms that are related in a thesaurus in this sense, we conducted an experiment with 6000 pairs of terms from the parliament's thesaurus. One half of the pairs consists of related words and the other half of words that are equally distributed over related but not synonymous word pairs and completely unrelated pairs. On these pairs we train and test a classifier that distinguishes pairs of related from pairs of unrelated words.

The remainder of this paper is organized as follows: in Sect. 2 we present the parliament's thesaurus and its role in the information processes of the Dutch parliament. Related work is discussed in Sect. 3. In Sect. 4 we introduce the data used for the experiment. Section 5 describes the experiment and Sect. 6 gives the results. Finally, we reflect on the practical implications of the experiment and discuss further steps in Sect. 7.

2 Information Processes in the Dutch Parliament

Every day, the Dutch parliament produces and processes many documents. In order to make the information available to the users – i.e. the internal users of the Parliament and also the Dutch citizens – documents are enriched with a great number of metadata, which include subject terms from a controlled vocabulary. A parliamentary thesaurus (*Parlementsthesaurus*) has been used for many years as a source for subject indexing of those documents that are produced in the parliamentary process, as well as other documents that are relevant for the parliament such as reports, articles in journals and newspapers, and interviews.

2.1 Indexing

In the past subject metadata were the only source of information about the subject of documents. Today, all parliamentary documents are full text available and searchable online. In order to make documents accessible in an organized and coherent way, however, subject metadata are still valuable tools. Using subject terms from a controlled vocabulary like a thesaurus may add value to the result of information seeking, as both the fraction of relevant documents that are retrieved (recall) and the fraction of retrieved documents that are relevant (precision) can be enhanced under certain circumstances (see e.g. Tudhope et al (2006)). Furthermore, subject keywords have been found to be valuable tools for enhancing the results of keyword searching in OPACs. Gross and Taylor (2005) found that more than one third of records retrieved by keyword searches would be lost without controlled subject keywords. This finding was recently replicated after the addition of automated enriched metadata like summaries and tables of content (Gross et al, 2015).

In searching for documents on a specific subject it is possible to enter one ore more subject terms from the thesaurus directly, finding only documents that have been indexed with a combination of the subject terms in question. More and more, the thesaurus is also used linked to the search engine in order to enhance the results of the full text search process, e.g. by the use of synonyms. In principle, this also applies to the semantic relations. Including semantic relations in full text search, however, may affect the relevance of the search results. Moreover, subject terms are used in the presentation of information,

e.g. as a basis for word clouds for example indicating areas of interest for parliamentarians.

Subject indexing is done manually by the information officers of the parliament (First and Second Chamber). This, of course, is a rather laborious process. In order to improve the efficiency, the consistency and the quality of indexing, efforts are taken to automate subject indexing. Various forms of language technology are used for this purpose. Until now this has not lead to a useful application of the technology. It is believed this is due to, among others, the breadth of the thesaurus, the large number of subject terms and the specific features of parliamentary documents.

Automatically adding more synonyms, especially frequently used words in the parliamentary context, to the thesaurus is one direction to improve the usefulness of the thesaurus for new applications.

2.2 The Parliament's Thesaurus

The origin of the current thesaurus lies in the 1980s, when first steps were taken to develop a controlled vocabulary for the Dutch parliament. Nowadays, the Parlement's thesaurus is a large polyhierarchical thesaurus that has been built around a number of main subjects or policy areas. A wide range of policy areas is covered, from health care and education to environmental planning and agriculture. The thesaurus consists of > 4000 descriptors and > 6000 non-descriptors, along with their semantic relations, synonyms and definitions (scope notes). Semantic relations include hierarchical relations (broader, more general, and narrower, more specific concepts), and also concepts that are otherwise related ("associative" relations).

Maintenance of the thesaurus is done by a thesaurus manager in consultation with a number of thesaurus editors: Information officers who are specialized in one or more policy areas and who are using the thesaurus in indexing. The nature of this process enhances the substantive quality of the thesaurus, but also makes it less dynamic than would be desirable in the present time, as it usually takes some time to include new subject areas and new terminology. Therefore, the information office of the parliament is looking for a way to improve the dynamic properties of the thesaurus while maintaining the substantive quality.

3 Related Work

Distributional similarity (DS) has been widely studied to solve many different tasks related to the meaning of words, and has become an established method to find similar words. As Harris (1954) states in his distributional hypothesis, the degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear. DS has been studied intensively for over two decades. A systematic overview of the different approaches is e.g. given by Turney and Pantel (2010) and Saif and Hirst (2012). Detailed studies of the influence of different choices in the implementation of distributional similarity are given by Bullinaria and Levy (2007, 2012) and Kiela and Clark (2014).

Construction of thesauri is often mentioned as a possible application of DS (Crouch, 1990; Crouch and Yang, 1992; Curran and Moens, 2002). Nevertheless, there are only very few studies that concretely investigate the problem of inserting new terms into an existing thesaurus.

Witschel (2005) uses DS to find the right insertion position for new terms in a hierarchically organized taxonomy. Starting at the root, the taxonomy is traveled downwards as long as one child of the current node is more similar to the new concept than all its sibling nodes. However, even in a very small taxonomy this method did not give very good results.

Meusel et al (2010) use a web search engine and Hearst-patterns to find hyponyms and synonyms for new words in an existing thesaurus. In order to reduce the number of queries that have to be issued, the method is only applied to the 100 thesaurus terms with the highest distributional similarity to the new term. They tested their method on two different thesauri. For the two-way classification between synonyms and non-synonyms they got an accuracy of 98% and 85% respectively. For the more difficult two-way classification between synonyms and hyponyms they got an accuracy of 71% and 68% respectively.

Overviews of methods used in the more general problem of automatic thesaurus construction are given e.g. by Biemann (2005) and Drumond and Girardi (2008).

4 Data

In this section we present both the test set we have created and used, and the corpus used to construct the context vectors for all terms.

4.1 Word pairs

Since it is both unclear how the relation between descriptors and non-descriptors in a thesaurus has to be described in traditional semantic terms and what type of semantic relation corresponds to distributional similarity, we developed a new test collection for this type of relation. The goal of the test collection is to see whether distributional similarity can be used to describe the relation between non-descriptors and descriptors, also known as the *use/use-for* relation, in a thesaurus.

Our test set consists of 6000 pairs of words. Half of the pairs are two words that are both a label (descriptor or non-descriptor) for the same concept. These pairs have been sampled randomly. For the negative pairs we have included a balanced quantity of easy and hard pairs, i.e. pairs that are hard to distinguish from positive pairs. To do so, we randomly selected 500 pairs of words that are labels for directly related concepts. Any type of relation specified in the thesaurus was used for this. Next we selected 500 pairs of labels for concepts that are related by one intermediate concept; for the next 500 pairs we took concepts that have two intermediate concepts on the shortest path between each other in the thesaurus. The remaining pairs were found in the same way for a longer thesaurus distance each time.

Some examples of positive and negative pairs are given in Table 1.

4.2 Corpus

Distributional similarity between two words is basically a similarity in the distribution of those words in a large corpus. We have compiled a corpus with texts that are quite characteristic for the texts that are annotated in the Dutch parliament. Thus we expect that the meaning and the use of the words in the corpus is similar to that intended in the thesaurus.

Table 1 Example of pairs of labels for the same and for different concepts and the number of intermediate concepts tn the tarliaments thesaurus.

term 1	term 2	intermediate nodes	synonymous
volksgezondheid (*public health*)	gezondheid (*health*)	0	+
regelgeving (*regulations*)	wetsvoorstel (*bill, draft law*)	0	+
krijgsraad (*court-martial*)	militair strafprocesrecht (*military criminal procedure*)	0	+
schilderij (*painting*)	beeldende kunst (*visual arts*)	1	-
volksuniversiteit (*adult education center*)	rijksuniversiteit (*state university*)	2	-
zwijgplicht (*confidentiality*)	politierechter (*magistrate of a police court*)	3	-

As the base for our corpus we have collected texts from `bestanden.officielebekendmakingen.nl`, the site with all official publications from the Dutch government, from the years 2010, 2011 and 2012. This site partially overlaps with the archived material from the parliament. Due to server and connection time outs the corpus does not contain all documents from the aforementioned years. The raw corpus consists of 88,8 million words. Since many documents start or end with exactly the same formulations, we kept only unique sentences. This results in a corpus of 47 million words.

For the computation of the distributional contexts of each word we lemmatized the complete corpus (using the TreeTagger with the parameter files included in the distribution) and removed all stop words. A corpus of 40 million lemmata now remains.

5 Experiment

In this section we describe a simple experiment to test whether we can train a classifier on the set of positive and negative examples. As features we use just one type of distributional similarity and string similarity. Other features we tested, did not improve the results.

5.1 Feature Vectors

First we have to construct feature vectors for each word. Though we can use a lot of different types of information about the context each word appears in, simple co-occurrence yields in most cases best results (Bullinaria and Levy, 2007, 2012; Kiela and Clark, 2014). Thus the features of each word are the co-occurrence values for each other word in a window of two words to the left and two words to the right. However, to limit the number of features, we use only words in a mid-frequency range. We have used 200 occurrences as a lower and $1 \cdot 10^6$ occurrences in our corpus as an upper bound. This gave us a total number of 11 080 context features for each word. The value of each feature c for each word w is the Positive Pointwise Mutual Information (PPMI) between w and c. The PPMI between w and c is defined as:

$$ppmi(w,c) = \max\left(\log\frac{p(w|c)}{p(w)}, 0\right). \quad (1)$$

Finally, we have used the cosine between the vectors of context features as the similarity between the words.

5.2 String similarity

Many labels are just spelling variants of the preferred label. This type of similarity can be easily captured with n-gram overlap or with edit distance. In the following we use trigram overlap, since that turned out to be the most effective similarity measure. Also no combination of different string similarity measures was better than just the trigram overlap.

For a word or string $w = w_0 w_1 \dots w_n$ we define the set of trigrams as $tg(w) = \bigcup_{i=0}^{n-2}\{w_i w_{i+1} w_{i+2}\}$. For two words w^1, w^2 we define the trigram overlap as the Jaccard coefficient of their sets of trigrams: $\text{overlap}(w^1, w^2) = \frac{|tg(w^1) \cap tg(w^2)|}{|tg(w^1) \cup tg(w^2)|}$.

5.3 Experimental Setup

For each pair of words we have two features: the cosine of the context vectors and the trigram overlap. We train a Support Vector Machine (SVM) that classifies the pairs of words into related and unrelated pairs.

We used LIBSVM to learn the model and classify the word pairs represented by the two features. The hyper-parameters of the model have been tuned using grid search. To find the best C parameter value, we investigated the numbers in between 0 and 20 in steps of 0.05.

We have used 10-fold cross validation with stratified sampling for evaluation.

6 Results

Average results from tenfold cross validation are given in Table 2. Both the distributional similarity and the trigram overlap are useful features for the classification of words as being labels of the same thesaurus concept or of different thesaurus concepts. Using both features also gives better results than using one of both features. The reached accuracy of 0.75 is clearly better than e.g. a majority classifier, that would assign each pair to one of both classes, but still far from perfect.

The experiments carried out by Meusel et al (2010) are quite similar to our experiment. Since they use different thesauri and different test sets, it is not really possible to compare the results. Nevertheless, the results they give for the binary classification of synonyms vs. non-related terms are much better than our results. However, they use random word pairs for the non-synonyms, whereas we deliberately selected difficult pairs, including hyponyms, co-hyponyms etc. In fact Meusel et al. also tested the classification of such difficult pairs in their classification of synonyms versus hyponyms. For this task their results are slightly worse than our results.

Table 2 Accuracy results of classification with ten-fold cross validation of 6000 pairs of labels for the Dutch parliament's thesaurus. Half of the pairs consist of labels of the same concept, half of labels from different concepts.

Features	Accuracy
cosine	0.69
trigram overlap	0.72
both features	0.75

7 Conclusion and Future Work

We have shown that distributed similarity can be used to model the relation between concept labels in a traditional thesaurus. In addition we used string similarity and trained a classifier to combine both types of features. For a data set of 6000 related and unrelated word pairs that we constructed for this task, the classifier could classify about 75% of the pairs correctly.

Despite the fact that the proposed features are useful for the considered task, the classification is still far from perfect and either other features have to be found or the distributional similarity has to be improved massively. Moreover, first experiments show, that the results do not carry over to a practical situation in which there are many candidate words. In such a scenario, we would have a large number of possible thesaurus terms. Each new term would be proposed as new alternative label for the concept with the labels that are most similar to the label under consideration. In this situation, however, the fraction of spelling variants is much smaller than in our test collection. This makes the string similarity a less useful feature. Furthermore, the number of word pairs that have a high trigram overlap just by accident, seems to increase with a growing number of words.

For future work we will pursue three different directions. First we will work on improving the distributional similarity. E.g. using the cosine for the similarity of the context vectors might not be a good choice in the given situation (Weeds et al, 2004; Wartena, 2013, 2014). Furthermore, we will integrate the use of Hearst patterns.

A second interesting question is the influence of the corpus used to construct the context vectors. While the influence of corpus size on distributional similarity has been studied quite well, little is known on the influence of the text selection. For the current task we have the possibility to compare a neutral corpus with a specialized corpus within the same domain as the test set.

Finally, we will work on more realistic scenarios, like extracting candidate terms from the corpus, assign these terms to the most likely concept and evaluate manually.

References

Biemann C (2005) Ontology learning from text - a survey of methods. In: LDV forum, vol 20, pp 75–93

Bullinaria JA, Levy JP (2007) Extracting semantic representations from word co-occurrence statistics: A computational study. Behaviour Research Methods 39(3):510–526, DOI 10.3758/BF03193020

Bullinaria JA, Levy JP (2012) Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists, Stemming and SVD. Behaviour Research Methods 44(3):890–907, DOI 10.3758/s13428-011-0183-8

Crouch CJ (1990) An approach to the automatic construction of global thesauri. Information Processing & Management 5:629–640, DOI 10.1016/0306-4573(90)90106-C

Crouch CJ, Yang B (1992) Experiments in automatic statistical thesaurus construction. In: Belkin N, Ingwersen P, Pejtersen AM (eds) Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, SIGIR '92, pp 77–88, DOI 10.1145/133160.133180

Curran JR, Moens M (2002) Improvements in automatic thesaurus extraction. In: Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition - Volume 9, Association for Computational Linguistics, Stroudsburg, ULA '02, pp 59–66, DOI 10.3115/1118627.1118635

Drumond L, Girardi R (2008) A survey of ontology learning procedures. In: de Freitas FLG, Stuckenschmidt H, Pinto HS, Malucelli A, Corcho Ó (eds) Proceedings of the 3rd Workshop on Ontologies and their Applications, Salvador, Bahia, Brazil, October 26, 2008, CEUR-WS.org, vol 427

Gross T, Taylor AG (2005) What have we got to lose? the effect of controlled vocabulary on keyword searching results. College & Research Libraries 66(3):212–230, DOI 10.5860/crl.66.3.212

Gross T, Taylor AG, Joudry DN (2015) Still a lot to lose: The role of controlled vocabulary in keyword searching. Cataloging & Classification Quarterly 53(1):1–39, DOI 10.1080/01639374.2014.917447

Harris ZS (1954) Distributional structure. Word 10(23):146–162, DOI 10.1080/00437956.1954.11659520

Kiela D, Clark S (2014) A Systematic Study of Semantic Vector Space Model Parameters. In: 2nd Workshop on Continuous Vector Space Models and

their Compositionality (CVSC), Association for Computational Linguistics, Gothenburg, pp 21–30

Meusel R, Niepert M, Eckert K, Stuckenschmidt H (2010) Thesaurus extension using web search engines. The Role of Digital Libraries in a Time of Global Change pp 198–207, DOI 10.1007/978-3-642-13654-2_24

Saif M, Hirst G (2012) Distributional Measures of Semantic Distance: A Survey. arXiv preprint/12031858

Tudhope D, Binding C, Blocks D, Cunliffe D (2006) Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation 62(4):509–533, DOI 10.1108/00220410610673873

Turney PD, Pantel P (2010) From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research 37:141–188, DOI 10.1613/jair.2934

Wartena C (2013) Hsh: Estimating semantic similarity of words and short phrases with frequency normalized distance measures. In: Manandhar S, Yuret D (eds) Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, Atlanta, pp 48–52, URL `http://www.aclweb.org/anthology/S13-2008`

Wartena C (2014) On the effect of word frequency on distributional similarity. In: Ruppenhofer J, Faaß G (eds) Proceedings of the 12th Edition of the Konvens Conference, Hildesheim, Germany, October 8-10, 2014, Universitätsbibliothek Hildesheim, pp 1–10

Weeds J, Weir DJ, McCarthy D (2004) Characterising Measures of Lexical Distributional Similarity. In: Proceedings of the 20th International Conference on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, COLING '04, DOI 10.3115/1220355.1220501

Witschel HF (2005) Using decision trees and text mining techniques for extending taxonomies. In: Proceedings of the Workshop on Learning and Extending Lexical Ontologies by Using Machine Learning Methods, ACM, New York, pp 61–68