Vocabulary Expansion by Semantic Extraction of Medical Terms

Maria Skeppstedt
DSV, Stockholm University
mariask@dsv.su.se

Magnus Ahltorp
Royal Institute of Technology
map@kth.se

Aron Henriksson
DSV, Stockholm University
aronhen@dsv.su.se

Abstract

Automatic methods for vocabulary expansion are valuable in supporting the development of terminological resources. Here, we evaluate two methods based on distributional semantics for extracting terms that belong to a certain semantic category. In a list of 1000 terms extracted from a corpus of Swedish medical text, the best method obtains a recall of 0.53 and 0.88, respectively, for identifying 90 terms that are known to belong to the semantic categories Medical Finding and Pharmaceutical Drug.

1 Introduction

High-coverage terminologies are important for medical text processing systems, such as named entity recognizers and information extractors. Manual terminology development is, however, expensive and time-consuming; it also runs the risk of resulting in insufficiently extensive terminologies and a subsequent negative impact on the recall of systems in which these are used. Methods that can support this process in various ways are thus very valuable.

Given the availability of a large corpus, methods based on distributional semantics – i.e. methods that exploit term co-occurrence patterns – make it possible to determine, in an unsupervised fashion, which terms are semantically related and to what extent. Several studies have demonstrated the potential of these methods in the (bio)medical domain (Cohen and Widdows, 2009), also with clinical corpora for the purpose of semi-automatic medical vocabulary development (Henriksson et al., 2012) and query expansion (Zeng et al., 2012).

Previous applications of distributional semantics for terminology development support and similar tasks have either focused on the extraction of very closely related terms, e.g. synonyms (Landauer and Dumais, 1997; Henriksson et al., 2013), or used features derived from such methods to train named entity recognition systems (Sahlgren and Cöster, 2004; Jonnalagadda et al., 2012). Here, we aim to study more closely the potential of using distributional semantics to extract terms that belong to a specific semantic category of medical terms, which will hopefully contribute to the areas of semi-automatic terminology development and unsupervised feature extraction.

2 Background

Methods for automatic vocabulary extraction can be divided into two main types, depending on whether or not there already exists a terminology (or a set of seed words belonging to predefined semantic categories). With the availability of a terminology in the target domain, as is the case in this study, vocabulary extraction can be seen as a classification task, determining whether an unknown word belongs to a certain semantic category. If there does not yet exist a suitable resource, however, a clustering approach needs to be taken, where clusters constitute candidates for semantic categories. In either case, the vocabulary extraction is based on finding patterns of contexts in which words typically occur (Biemann, 2005).

Semantic (word) spaces, derived from a corpus, represent such context patterns in the form of word co-occurrence information. This representation has been used both for creating clusters of semantically related words (Song et al., 2007) and for determining whether unknown words belong to predefined semantic categories (Widdows, 2003; Curran, 2005). In this study, we use a computationally light-weight version of the semantic space representation called *random indexing* (Kanerva et al., 2000; Karlgren and Sahlgren, 2001; Sahlgren, 2005). Instead of reducing the dimensionality of a word-by-word (or word-by-context) matrix to

make it computationally tractable (which is the approach taken for creating many other types of semantic spaces), a matrix with a smaller dimensionality is created from the beginning. Each word in the corpus is assigned a unique representation in the form of an *index vector* with a dimensionality that is much smaller than the number of unique terms in the corpus. The near-orthogonal index vectors are created by randomly generating very sparse vectors, in which most of the elements are set to 0, while a few (1–2%), randomly selected, elements are set to either $+1$ or -1. Each word is also assigned a *context vector* with the same dimensionality as the index vector, in which all elements are initially set to 0. For every occurrence of a word in the corpus, its context vector is updated by adding the index vectors of the words in the context window (the surrounding words). Different semantic relations can be modelled by varying the size of the context window (Sahlgren, 2006). The resulting semantic space consists of the context vectors, between which, e.g., the cosine similarity can be computed to determine the semantic distance between words.

3 Materials and Methods

The proposed approach essentially requires two resources: a large corpus of medical text and a number of seed terms that belong to the semantic category of interest. To allow the method(s) to be evaluated automatically, additional terms that are known to belong to the same semantic category are also needed. Here, a corpus of Swedish medical text and subsets of the Swedish version of the medical vocabulary MeSH were used.

3.1 Semantic Spaces of Medical Text

Semantic spaces were induced from a Swedish medical corpus: Läkartidningen, which is the Journal of the Swedish Medical Association (Kokkinakis, 2012) and contains articles on, for instance, new scientific findings in medicine, pharmaceutical studies and health-economic evaluations. Editions from the years 1996–2005 were used, as these have been made available for research, albeit with the sentences given in a random order. The corpus was preprocessed by (white-space) tokenising and lower-casing the text. Since the sentence order is scrambled, a document break was inserted between sentences to ensure that co-occurrence information is not collected across sentences in the construction of the semantic spaces. The corpus was not lemmatised, as inflected forms of medical terms may also be relevant candidates for vocabulary expansion. The corpus contains 21 447 900 tokens and 444 601 unique terms.

Random indexing was applied to induce 1000-dimensional semantic spaces[1] from variants of this corpus. The semantic spaces were evaluated in two steps: (1) in a development phase, where context window size was optimised separately for each of the two semantic categories (Medical Finding and Pharmaceutical Drug) and for each of the two proposed methods, and (2) in a final evaluation phase, where the best-performing semantic spaces, in terms of recall, were evaluated on unseen data. The context window sizes 1+1, 2+2, 4+4 and 50+50 were evaluated in the development phase. The 50+50 window size is, in effect, a sentence-level context definition since the sentence delimiters ensure that context information from adjacent sentences is ignored.

3.2 Semantic Term Extraction

Two computationally efficient methods for vocabulary expansion using random indexing were devised and evaluated: *Term Replacement (TermRep)* and *Cosine Addition (CosAdd)*.

In the first method, *TermRep*, the corpus was modified before the semantic spaces were created. All occurrences of a set of seed terms that belong to a given semantic category were replaced by a common string denoting that category. This can be seen as an aggressive form of term normalisation and entails that each semantic category is assigned a single context vector, which is populated with the index vectors of terms that co-occur with all lexical instantiations of that semantic category. The string that represents the semantic category of interest was then given as a query term to the semantic space, resulting in a ranked list of distributionally similar terms, presumably some of which belong to the same semantic category.

In the second method, *CosAdd*, the semantic spaces were created with the unmodified corpus. Each term in the set of seed terms was instead used as a query term, resulting in one ranked list per seed term, containing the cosine similarity between this seed term and every other word in the

[1] 10 non-zero elements (i.e., 1%) were assigned to the index vectors. When populating the context vectors, increasingly less weight was assigned to index vectors as the distance from the target term increases.

corpus. The ranked lists of the seed terms were then merged into a single ranked list per semantic category. The merge was performed by summing the cosine similarity scores.

A certain number of observations of a term is required for its context vector to be accurately positioned. Words occurring fewer than 50 times were therefore not included as seed terms; they were also excluded from the lists of candidate terms.

3.3 Medical Terminology and Evaluation

The medical terminology was here employed for two purposes: (1) as a set of seed terms for a given semantic category and (2) as a reference standard for evaluating the two proposed methods.

The Swedish version of MeSH[2] (Karolinska Institutet, 2012), a controlled vocabulary for indexing life science literature, was here used for these purposes. For the semantic category Medical Finding, terms that belong to the Swedish MeSH categories *Disease or syndrome* and *Sign or symptom*[3] were used; for the semantic category Pharmaceutical Drug, the MeSH category *Pharmacologic substance* was used.

MeSH terms occurring fewer than 50 times were excluded as seed terms (as mentioned above), as well as reference standard terms. Multiword terms were also excluded, as current models of distributional semantics perform better on unigram terms (Henriksson et al., 2013). When rare and multiword terms had been removed, 309 terms that belong to Medical Finding and 181 terms that belong to Pharmaceutical Drug remained. In order to enable a fairer comparison between the two semantic categories, 181 Medical Finding terms – identical to the number of Pharmaceutical Drug terms – were randomly selected.

The terms used in the evaluation for each semantic category were divided into two stratified, equally large groups, a *development set* and an *evaluation set*, in which the strata consisted of terms with similar frequencies in the corpus. In the development phase, the terms in the development set were used for optimising context window size. In the evaluation phase, all terms were used: the terms in the development set were treated as seed terms, which, in a real-world scenario, would be known and already included in the terminology; the terms in the evaluation set were ones that, in a real-world scenario, we would like to add to the terminology.

The performance using different window sizes was measured using 10-fold cross-validation on the data in the *development set*. The 91 terms that belong to Medical Finding and the 91 terms that belong to Pharmaceutical Drug were divided into ten folds. That is, for each fold, approximately 82 terms were used as query terms – or, in the *TermRep* case, replaced by a common identifier in the corpus – and approximately 9 terms were expected to be retrieved, effectively making up the reference standard. Recall was measured as the proportion of expected terms that were found in a list of retrieved terms. Recall at different cut-off values (from 50 to 1000, with a step size of 50) were calculated. The semantic spaces with the highest average recall values were selected and used in the evaluation phase. This means that the semantic spaces were not optimised for a specific cut-off value, rendering the cut-off value a flexible parameter in the final evaluation.

In the evaluation phase, the primary evaluation was conducted in the form of a fully automatic evaluation of recall against the *evaluation set*. To determine to what extent retrieved terms belong to the expected semantic category, despite not being present in the reference standard, a semi-automatic evaluation of precision among the 500 top-ranked terms was also performed. Retrieved terms classified as Finding or Drug in MeSH or FASS (2012) were automatically classified as correct or incorrect (assuming that a known Finding can never be a Drug and vice versa). The remaining terms were manually classified by a single annotator as belonging to the category or not.

4 Results

Averaging the recall measurements for the 20 cut-off values yielded the results shown in Table 1. There were no large differences between window sizes, but the best recall (for both methods) was obtained with a context window of 2+2 for Medical Finding and 1+1 for Pharmaceutical Drug. Semantic spaces induced with these window sizes were therefore used in the final evaluation.

The ability of the two methods to extract the expected terms in the evaluation set is shown in Figure 1. For Medical Finding there was no large difference between the two methods, whereas *Cosine*

[2] **ME**dical **S**ubject **H**eadings: `http://www.nlm.nih.gov/mesh`

[3] As there is a rather fine distinction between these two subcategories, they were merged into a single category.

Window Size	1+1	2+2	4+4	50+50
	Medical Finding			
CosAdd	0.372	**0.389**	0.384	0.382
TermRep	0.357	**0.368**	0.361	0.360
	Pharmaceutical Drug			
CosAdd	**0.567**	0.516	0.502	0.501
TermRep	**0.409**	0.386	0.375	0.371

Table 1: Average recall values over 20 different cut-offs (top 50 – top 1000) on development data.

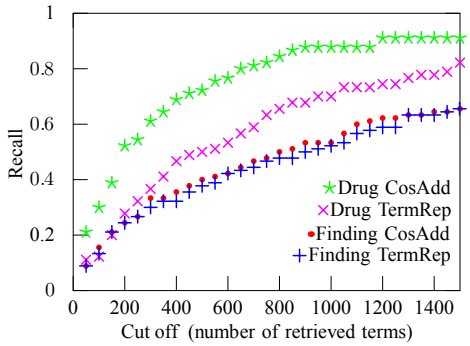

Figure 1: Recall values for different cut-offs

Addition outperformed *Terminology Replacement* for Pharmaceutical Drug. Both methods obtained better recall for extracting Drug terms than Finding terms. The overlap of retrieved terms for the two methods was 83% for Finding and 76% for Drug (top 1000). For the *CosAdd* method, precision was also evaluated, with better results for Finding than for Drug (0.80 vs. 0.64 for top 50 and 0.68 vs. 0.47 for top 100, Figure 2).

5 Discussion

Two computationally light-weight methods for automatic vocabulary expansion have been studied.

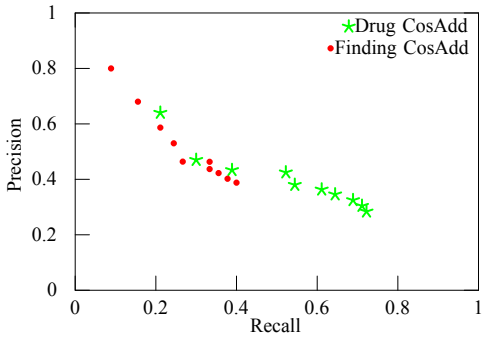

Figure 2: Precision (partially based on manual classification) vs. recall (automatically measured against the reference standard), cut-off 50–500.

Seed terms were modelled as if they would form two separate clusters in the semantic space: one for Medical Finding and one for Pharmaceutical Drug. When applying the replacement method, we are in effect searching for new words that are close to a weighted centroid of the cluster. The weighting emerges from the fact that the effect of each seed term on the resulting centroid context vector is directly proportional to the frequency of the seed term in the corpus. This makes the method vulnerable to frequent seed terms that are atypical for the semantic category, which might explain the lower results with this method for Pharmaceutical Drug, as, for instance, *alcohol* was the second most frequent seed term. With the addition method, on the other hand, each seed term is given equal weight, and new words are deemed equally typical to the semantic category irrespective of the frequency of the seed term to which they are close. This means that employing a low frequency threshold for which seed terms to include might drastically lower the results, as there is a weak statistical foundation for the position of the context vectors of the many low-frequent terms.

6 Conclusion and Future Work

The best performing method was able to extract 53% of the 90 expected Medical Findings and 88% of the 90 expected Pharmaceutical Drugs among the top 1000 retrieved terms, showing its potential as a useful component in a semi-automatic vocabulary expansion process. Future work should, however, include a comparison between the approaches evaluated here and previous approaches, for their ability to retrieve expected terms and also for their computational efficiency.

Moreover, modelling a MeSH category as one cluster in the created semantic space is most likely an over-simplification. There might be a number of sub-clusters within each of the two categories Finding and Drug – sub-clusters that are positioned at large distances from each other in the semantic space. Words not part of these sub-clusters, but close to two or more clusters, will then receive a high ranking with the methods applied here, even though they ought to be ranked lower than words close to the centroids of the sub-clusters. As the next step, we will therefore attempt to cluster the seed terms into sub-clusters and apply the distance measures of this study to rank the similarity of unknown words to these sub-clusters.

Acknowledgments

We are very grateful to Rafal Rzepka and Shiho Kitajima for their valuable feedback on the study. We would also like to thank the three reviewers for many good comments.

This work was partly supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection (ref. no. IIS11-0053) at Stockholm University, Sweden.

References

Chris Biemann. 2005. Ontology learning from text: A survey of methods. In Alexander Mehler, editor, *Themenschwerpunkt Korpuslinguistik, GLDV-Journal for Computational Linguistics and Language Technology*. Gesellschaft für Linguistische Datenverarbeitung e. V. (GLDV).

Trevor Cohen and Dominic Widdows. 2009. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390 – 405.

James R. Curran. 2005. Supersense tagging of unknown nouns using semantic similarity. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 26–33, Stroudsburg, PA, USA. Association for Computational Linguistics.

FASS. 2012. Fass.se. http://www.fass.se, Accessed 2012-08-27 08-27.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Ann-Marie Eklund, Vidas Daudaravičius, and Martin Hassel. 2012. Synonym Extraction of Medical Terms from Clinical Text Using Combinations of Word Space Models. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine (SMBM)*.

Aron Henriksson, Maria Skeppstedt, Maria Kvist, Martin Duneld, and Mike Conway. 2013. Corpus-Driven Terminology Development: Populating Swedish SNOMED CT with Synonyms Extracted from Electronic Health Records. In *Proceedings of BioNLP*. Association for Computational Linguistics.

Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. 2012. Enhancing clinical concept extraction with distributional semantics. *Journal of Biomedical Informatics*, 45(1):129–140.

Pentti Kanerva, Jan Kristofersson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of 22nd Annual Conference of the Cognitive Science Society*, page 1036.

Jussi Karlgren and Magnus Sahlgren. 2001. From words to understanding. *Foundations of Real-World Intelligence*, pages 294–308.

Karolinska Institutet. 2012. Hur man använder den svenska MeSHen (In Swedish, translated as: How to use the Swedish MeSH). http://mesh.kib.ki.se/swemesh/manual_se.html. Accessed 2012-03-10.

Dimitrios Kokkinakis. 2012. The journal of the Swedish medical association - a corpus resource for biomedical text mining in Swedish. In *The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop. Turkey*.

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*, pages 211–240.

Magnus Sahlgren and Rickard Cöster. 2004. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 487. Association for Computational Linguistics.

Magnus Sahlgren. 2005. An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, volume 5.

Magnus 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, PhD thesis, Stockholm University.

Dawei Song, Guihong Cao, Peter D. Bruza, and Raymond Lau. 2007. Concept induction via fuzzy c-means clustering in a high-dimensional semantic space. In J. Valente de Oliveira and W. Pedrycz, editors, *Advances in Fuzzy Clustering and its Applications*, pages 393–403. John Wiley & Sons, Chichester.

Dominic Widdows. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 197–204, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qing T Zeng, Doug Redd, Thomas Rindflesch, and Jonathan Nebeker. 2012. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. In *AMIA Annual Symposium Proceedings*, pages 1050–1059.