Expanding a dictionary of marker words for uncertainty and negation using distributional semantics

Alfalahi, Alyaa; Skeppstedt, Maira; Ahlblom, Rickard; Baskalayci, Roza; Henriksson, Aron; Asker, Lars; Paradis, Carita; Kerren, Andreas

Published in:
EMNLP 2015 - 6th International Workshop on Health Text Mining and Information Analysis, LOUHI 2015

2015

Expanding a dictionary of marker words for uncertainty and negation using distributional semantics

Alyaa Alfalahi[1], **Maria Skeppstedt**[2,3,*], **Rickard Ahlbom**[1], **Roza Baskalayci**[1], **Aron Henriksson**[1,*], **Lars Asker**[1], **Carita Paradis**[4], **Andreas Kerren**[3]
[1]DSV, Stockholm University, Stockholm, Sweden [2]Gavagai AB, Stockholm, Sweden
[3]Computer Science Department, Linnaeus University, Växjö, Sweden
[4]Centre for Languages and Literature, Lund University, Lund, Sweden
[*]Corresponding authors: maria@gavagai.se, aronhen@dsv.su.se

Abstract

Approaches to determining the factuality of diagnoses and findings in clinical text tend to rely on dictionaries of marker words for uncertainty and negation. Here, a method for semi-automatically expanding a dictionary of marker words using distributional semantics is presented and evaluated. It is shown that ranking candidates for inclusion according to their proximity to cluster centroids of semantically similar seed words is more successful than ranking them according to proximity to each individual seed word.

1 Introduction

Clinical text, i.e., the narrative sections of health records, has recently received much attention with regards to automatic detection of uncertainty and negation (Uzuner et al., 2011; Velupillai, 2012; Mowery et al., 2014). Methods for automatic detection of which diagnoses and findings are mentioned as negated or uncertain typically rely on a dictionary of marker words, either as a resource for rule-based methods or when constructing features for machine learning (Uzuner et al., 2011). Dictionaries of marker words have previously been constructed by manual annotation or by translation of dictionaries from one language to another (Velupillai et al., 2014). Alternative methods for automating marker word dictionary construction would, however, be useful since manual annotation is time-consuming, and translation results in incomplete dictionaries due to differences between languages in how negation and uncertainty are expressed. The aim of the present study was to explore one such possible method for semi-automatic dictionary expansion: using distributional semantics to extract possible marker words from a large unannotated corpus and, more specifically, attempting to obtain improved performance by applying clustering to the semantic vectors in the resulting semantic space.

Given a dictionary of known uncertainty and negation markers to use as seed words, the task of the system explored here was to rank words not included in the seed dictionary according to their suitability as marker words, with the aim of having good candidates for inclusion in the dictionary among the top-ranked words.

An experiment was carried out to determine if a method whereby words are ranked according to proximity to the centroids of seed word clusters outperforms – in the sense of ranking true marker words higher – a ranking method that instead uses proximity to each individual seed word. The seed words are here represented as vectors comprising word co-occurrence information, created using a model of distributional semantics called random indexing.

2 Background

For the English language, there are a number of large corpora annotated for speculation and negation: bio-medical corpora (Vincze et al., 2008; Uzuner et al., 2011), as well as corpora in other domains (Konstantinova et al., 2012). Systems for detecting negation and speculation are typically constructed by training machine learning models on these corpora (Farkas et al., 2010; Uzuner et al., 2011). For most other languages, there are, however, often only smaller annotated corpora or none at all (Velupillai et al., 2011; Aramaki et al., 2014). In such cases, methods for detecting uncertainty and negation that rely on lexicon/dictionary-matching to lists of marker words for uncertainty or negation are a possible alternative. Such an approach has been shown to perform in line with machine learning methods trained on corpora with fewer training instances (Velupillai et al., 2014; Aramaki et al., 2014).

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 90–96,
Lisbon, Portugal, 17 September 2015.

For a dictionary-matching approach, extensive dictionaries of marker words are, however, required, and to build such a resource manually can also be prohibitively expensive. An alternative to creating a dictionary of marker words manually is to use automatic methods for creating lists of candidate words to include in the dictionary. For semi-automatically creating vocabulary resources of other types than marker words, there are a number of previous studies wherein various methods are used. Those that rely on terms being explicitly defined in the text (Hearst, 1992; Yu and Agichtein, 2003; Cohen et al., 2005; McCrae and Collier, 2008; Neelakantan and Collins, 2014) are unlikely to be successful for negation and uncertainty terms. Term extraction methods that measure similarity between words according to how frequently they occur in similar contexts (Lin, 1998), on the other hand, might be more suitable. Such distributional semantic properties are often represented by spatial models, i.e., given a geometric representation in the form of a vector space (Cohen and Widdows, 2009), and there are examples in which such spatial models have been used for vocabulary expansion (Zhang and Elhadad, 2013; Skeppstedt et al., 2013; Henriksson et al., 2014), as well as for related tasks (Jonnalagadda et al., 2012), in the bio-medical domain.

Random indexing is a computationally lightweight method for producing spatial models of distributional semantics (Kanerva et al., 2000; Sahlgren, 2006). Random indexing requires two types of vectors: index vectors, which are used only for semantic space construction, and context vectors, which represent the meaning of words and collectively make up the resulting semantic space. Each unique word w_j in the corpus vocabulary W is assigned an index vector $\vec{w}_j^i$ and a context vector $\vec{w}_j^c$ of dimensionality d. The index vectors are static representations of contexts (here, these are unique words) that are approximately uncorrelated to each other, which is achieved by creating very sparse vectors that are randomly assigned a small number of non-zero elements (1s and -1s). A $\vec{w}_j^c$ – containing the distributional profile of the word w_j – is then the (weighted) sum of all the index vectors of the words with which w_j co-occurs within a (typically symmetric) window of a certain size. Spatial proximity between two context vectors is taken to indicate the semantic similarity between the two words they represent. The context vectors can also be further analysed, for instance by applying different kinds of clustering (Rosell et al., 2009; Pyysalo et al., 2013).

3 Method

The conducted experiment consisted of the following steps: 1) constructing a semantic space with random indexing; 2) applying hierarchical clustering to context vectors representing seed words; 3) for different levels in the cluster tree, producing a ranked list of the words in the corpus according to their proximity to the centroids of the constructed clusters; 4) evaluating the recall of the top-ranked words in the produced lists against a reference standard.

1) A semantic space was constructed with random indexing on a freely available subset (years 1996–2005) of the *Läkartidningen* (Journal of the Swedish Medical Association) corpus (Kokkinakis, 2012). This subset contains 21,447,900 tokens and 444,601 unique terms. In order also to allow inflected forms of marker words to be captured, the corpus was not lemmatised. 1,000-dimensional vectors were used in a context window of two preceding and two following words and double weight was given to the two words closest to the target word. Since the sentences in the corpus appear in a randomised order, no context windows were allowed to cross sentence boundaries.

2) Single-linkage agglomerative hierarchical clustering (Sibson, 1973) was applied to the context vectors representing the seed words. A tree-formed cluster hierarchy was thereby created, with progressively larger clusters, starting from clusters in which each seed word formed its own cluster (cluster level *0* on the x-axis in Figure 1), until all seed words collectively formed a single cluster (cluster level *79* on the x-axis in Figure 1).

3) For each cluster level (0 to 79), a ranked list of all words in the corpus (except those used as seed words) was produced. The words were ranked according to the Euclidean distance between their length-normalised context vector and their most closely located cluster centroid (also length-normalised). That is, the word with the context vector that was closest to any of the centroid vectors achieved the highest ranking, the word with the context vector that was second closest to any of the centroid vectors was ranked as number two on the list, and so on. For cluster level

Cluster level 0	Cluster level 40	Cluster level 79
misstänka (suspect)	risken (the risk)	barnet (the child)
sannolikt (likely)	analys (analysis)	folk (people)
angeläget (pressing)	påvisats (proven)	arbetsgivaren (the employer)
rimligt (reasonable)	acceptera (accept)	så (so)
förmodligen (probably)	riskerar (risks)	uppdraget (the assignment)
tycker (think)	registrering (registration)	personalen (the staff)
kontrollera (check)	använda (use)	verksamhetscheferna (the business managers)
hävda (assert)	läran (doctrine)	medlet (the agent)
kartlägga (survey)	kommer (come)	läkarna (the doctors)
värdera (estimate)	kunskapen (knowledge)	landstingen (the counties)

Table 1: Top 10 words retrieved for a randomly selected seed word sampling (among the 500 re-samplings used in the experiment. The top 10 words for cluster level 0, 40 and 79 are shown).

0, in which each seed word formed its own cluster, the centroids were composed of the context vectors for the seed words, and the words were thus ranked according to their proximity to any of the seed words.

4) As a final step, the method was evaluated using an existing, freely available, dictionary of Swedish marker words for uncertainty and negation. This dictionary was developed through translation of English marker words and through manual annotation of clinical text (Velupillai et al., 2014). Markers in the dictionary were used as seed words as well as for evaluation data.

The dictionary was filtered by removing multi-word terms, since the constructed semantic space only contains single-word terms. In addition, words occurring fewer than 50 times in the corpus were removed, since a certain number of observations of a word is required for its context vector to be modeled reliably in semantic space. The performed filtering resulted in a set of 161 marker words for uncertainty and negation. The vocabulary used is shown in Figure 3.

This set of vocabulary terms was used in the evaluation by randomly splitting it into two equally large subsets: one set of seed words and one set of words to use as reference standard. The set of seed words represents words that, in a real-world scenario, would be included in an existing, but incomplete, dictionary of marker words, and the reference standard represents words that should be included as top-ranked candidates by the evaluated system. The performance of the system was evaluated through a standard information retrieval measure, i.e., by calculating recall (for the n top-ranked candidates) of the produced list against the words in the reference standard. Recall was calculated for up to top 5,000 candidate words (from top 100 with a step size of 100). Candidate list precision for the automatic evaluation is not reported, as this is separated only by a constant from recall, and would therefore show the same pattern with respect to cluster sizes.

To make the results less dependent on which terms were used as seed words and which were used as reference standard words, the experiment was repeated 500 times, each time with a new random split of the 161 words in the dictionary into a seed words set and reference standard set. The final results were achieved by averaging the achieved recall results.

Table 1 shows an example of the top 10 candidates retrieved for one randomly selected seed sample among the 500 evaluated re-samplings. In this short list, and for this sample, there are better candidates for cluster level 0 than for the other cluster levels.

4 Results and Discussion

As can be seen in Figure 1, results achieved with a moderate cluster level (20–40) were better than those achieved when proximity to each individual seed word was used as the ranking method (level 0). When the clusters grew larger (cluster level $>$ 50), however, recall started to decrease, and using proximity to the centroid of a cluster containing all seed words resulted in much lower recall than when using proximity to each individual seed word, indicating that there are important differences in the usage of marker words. As a method for ranking the words in the corpus, it was thus better to use proximity to the centroid of a

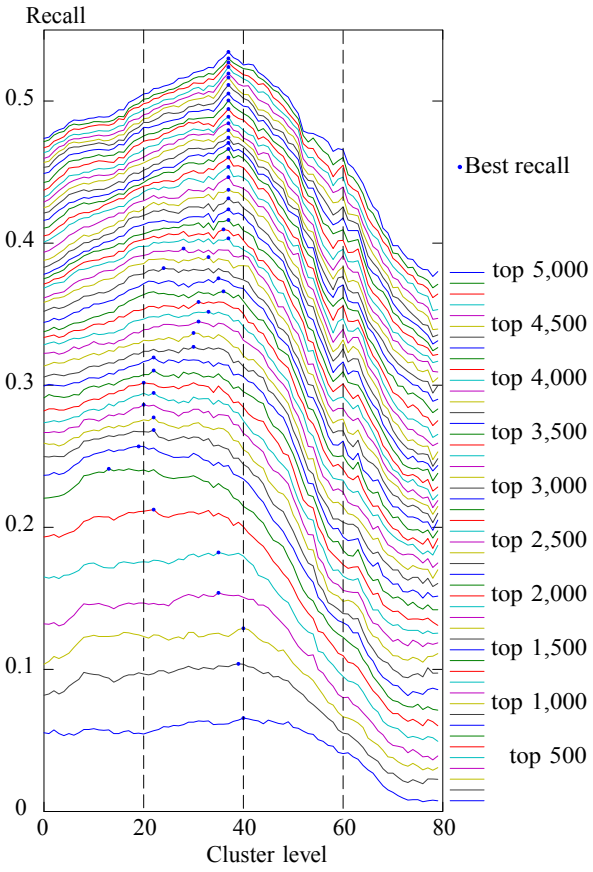

Figure 1: Recall for different levels of clustering. Cluster level 0 means that each seed word forms its own cluster. The higher the cluster level, the larger the clusters created. Cluster level 79 means that all seed words form one large cluster.

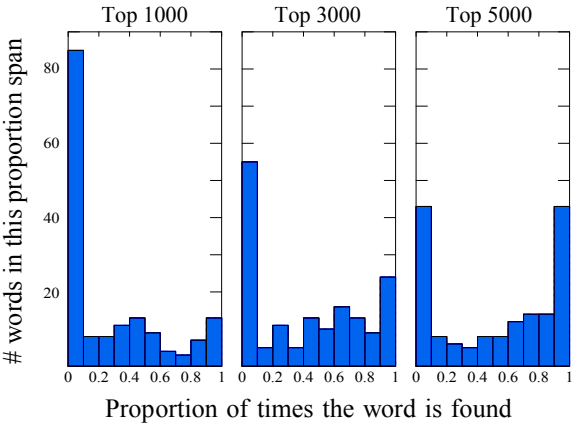

Figure 2: Histogram over the proportion of times a word is found when used as a reference standard word. The first stack shows the number of words that are found between 0% and 10% of the times they are used in the reference standard. The second stack shows the number of words found between 10% and 20% of the times, and so on. The statistics are shown for top 1,000, 3,000 and 5,000 candidates (using the cluster level optimal for top 3,000).

number of semantically similar words than to use proximity to each individual word. When using large clusters of seed words, however, distributionally dissimilar words, e.g., *förnekar (denies)* and *möjlig (possible)*, were clustered together, which decreased recall.

Recall is shown in Figure 1 from among the top 100 best candidates up to among the top 5,000 best candidates (with a step size of 100). The improvement that is achieved with a larger number of candidate words slowly levels out with an increasing number of candidates. The average result among the top 5,000 best candidates was a recall of just above 50%. A possible reason for these relatively low recall scores could be that the dictionary of marker words for uncertainty and negation contains many semantic outliers, i.e., words that do not occur in contexts similar to the other words in the list. The statistics shown in Figure 2 support this theory. The first stack in each of the three histograms, which shows the number of words that are very rarely found, is large in all three histograms. This indicates that regardless of which seed words are used, there is a large number of words that are never or very rarely found. It might, therefore, be the case that methods based on distributional semantics cannot be used for constructing a complete dictionary of negation and uncertainty markers, as such a dictionary includes semantic outliers, although the methods are useful for expanding a dictionary with typical marker words. Figure 3 shows the vocabulary used and how often a word was retrieved among the top 1,000 candidates when used as evaluation data.

It should be noted that the used list of marker words has been constructed using clinical text and has the aim of being used for clinical text, while this study was carried out on medical journal text. The used medical corpus has the advantage of being freely available, in contrast to large clinical corpora, which are only rarely available for research, and it also makes it possible for anyone to repeat the experiments carried out in this study. As there are many differences between medical journal text and clinical text (Smith et al., 2014), some marker words might be used in other contexts in clinical text than in medical journal text,

övertygande(convincing):0.0 överväga(consider):0.82 övervägas(considered):0.0 aldrig(never):0.0 alternativ(option):0.0 alternativa(alternative):0.0
alternativt(alternatively):0.43 angående(relating):0.0 anse(deem):0.97 ansetts(considered):0.0 antagits(been guessed):0.0 antas(is-guessed):0.21 antingen(either):0.0
antogs(was guessed):0.0 antydan(hint):0.0 antyder(implies):0.98 antytt(hinted):0.55 avfärda(dismiss):0.0 avfärdar(dismisses):0.0 beaktande(regard):0.0
bedömning(assessment):0.47 betänka(reports):0.48 borta(gone):0.0 differentialdiagnos(differential-diagnosis):0.0 ej(not):0.0 eventuell(possible):0.3 eventuella(any):0.0
eventuellt(optionally):0.0 förefaller(appears):0.4 föreslå(propose):0.92 föreslår(proposes):0.16 föreslagit(proposed):0.55 förmoda(surmise):0.8 förmodad(putative):0.0
förmodade(putative):0.0 förmodas(believed):0.28 förmodligen(probably):0.76 förneka(deny):0.97 förnekar(denies):0.08 förslagsvis(tentatively):0.4 fråga(issue):0.0
frågan(the-issue):0.06 frågeställning(issue):0.0 frågeställningen(the-issue):0.0 framstår(stands):0.66 framträder(stands):0.0 fri(free):0.0 fria(free):0.0 funderingar(speculations):0.0
granskning(review):0.33 indicerat(indicated):0.0 indikation(indication):0.31 indikationen(the-indication):0.59 indikationer(indications):0.27
indikationerna(the-indications):0.61 indikera(indicate):0.0 indikerar(indicates):0.97 indikerat(indicated):0.43 inga(no):0.0 ingen(no):0.22 ingenting(nothing):0.02
inget(no):0.21 inte(not):0.0 känna(feel):0.0 kanske(maybe):0.74 löst(solved):0.0 liknade(similar):0.02 liknar(resembles):0.3 märka(notice):0.96
möjlig(possible):0.13 möjliga(possible):0.0 möjligen(possibly):0.14 möjligheten(possibility):0.91 möjligt(possible):0.02 möjligtvis(possibly):0.14
misstänka(suspect):0.9 misstänker(suspect):0.57 misstänkt(suspect):0.0 misstänkta(suspects):0.0 misstankar(suspicions):0.64 misstanke(suspicion):0.36
misstanken(suspicion):0.58 negativ(negative):0.13 negativa(negative):0.0 negativt(negative):0.0 nog(probably):0.19 observerades(observed):0.0 observerats(observed):0.0
och/eller(and/or):0.0 oklar(unclear):0.53 oklart(unclear):0.45 oroande(worrying):0.05 osäker(unsure):0.52 osäkerhet(uncertainty):0.0 osäkert(uncertain):0.35
osannolik(improbable):0.0 osannolikt(improbable):0.51 otroligt(incredible):0.0 otydliga(unclear):0.02 påstår(states):1.0 preliminär(provisional):0.0
preliminärt(preliminary):0.0 protokoll(protocol):0.0 protokollet(protocol):0.0 representerar(represents):0.0 rimligtvis(reasonably):0.0 saknar(lack):0.0 saknas(missing):0.0
sannolik(probable):0.47 sannolika(probable):0.4 sannolikheten(probability):0.18 sannolikt(likely):0.81 sett(seen):0.0 stödjer(supports):0.03 svårbedömd(hard-to-assess):0.42
svårtolkade(difficult-to-interpret):0.02 syns(visible):0.0 tendens(tendency):0.0 tendenser(trends):0.0 tolka(interprete):0.98 tolkades(was-interpreted):0.59
tolkar(interpretes):0.0 tolkas(interpreted):0.0 tolkats(interpreted):0.05 torde(should):0.36 tro(believe):0.91 trodde(thought):0.83 trolig(probable):0.34
troliga(probable):0.22 troligen(probably):0.81 troligt(likely):0.22 troligtvis(probably):0.71 tror(think):0.03 tros(believed):0.0 trott(imagined):0.0 tveksam(passable):0.0
tveksamhet(hesitancy):0.0 tveksamt(doubtful):0.14 tycker(think):0.06 tycks(appears):0.46 tydligen(apparently):0.36 undersökning(study):0.37 uppenbarligen(obviously):0.41
uppleva(experience):0.9 upplevd(perceived):0.0 upplevdes(perceived):0.64 upplever(experiencing):0.0 utan(without):0.0 uteslöt(excluded):0.0
utesluta(exclude):0.87 uteslutas(excluded):0.0 utesluter(excludes):0.0 uteslutet(precluded):0.01 uteslutit(excluded):0.0 uteslutits(excluded):0.0
utreda(investigate):0.91 utredning(investigation):0.47 utvärdering(evaluation):0.47 varken(neither):0.0 verkar(seems):0.3 visa(show):0.94

Figure 3: The vocabulary used for the experiments, displayed in a font size corresponding to how often a word, when included in the evaluation data, was retrieved among the top 1,000 candidates. Words displayed in black were retrieved in less than 10% of the times they were included in the evaluation data.

and there might be fewer semantic outliers if the experiments were to be repeated using a clinical corpus.

There were also 54 negation and uncertainty markers in the used dictionary that were excluded from the study since they occurred fewer than 50 times in the corpus. The existence of these words, which were mainly inflected forms, abbreviations and a few misspellings that are unusual outside of the clinical language, e.g., *beaktandes (taking into consideration)*, *alt (alternatively)*, *diffdiagnos (differential diagnosis)*, is also a reason for why the experiment should be repeated with a clinical corpus. Multi-word terms formed an even larger proportion of the terms excluded from the negation and uncertainty dictionary when constructing the vocabulary used in the experiments (376 terms). There are previous studies in which multi-word negation and uncertainty markers have been constructed from single-word markers (Velupillai et al., 2014), but an alternative could be to directly model multi-word terms in semantic space (Henriksson et al., 2013a; Henriksson et al., 2013b).

A manual evaluation of a Swedish uncertainty and negation marker candidate list, produced with the methods of this study, could also be carried out in order to determine to what extent it is possible to obtain words not yet included in the dictionary using this method. The dictionary used for evaluation was, however, obtained by translation of English marker words and by extracting markers from clinical text in which 2,500 diagnostic statements had been annotated (Velupillai et al., 2014). It could, therefore, be difficult to retrieve standard language single-word terms for negation and uncertainty not already included in this dictionary. There might, however, still be a need to add abbreviated forms and multi-word terms. The methods evaluated here could also be applied to other languages, for which resources of marker words for negation and uncertainty, used in medical text, have not yet been constructed.

5 Conclusion

It was shown that proximity to the centroid of a number of semantically similar seed words was a more successful method for ranking the words in the corpus as candidates for negation and uncertainty markers than to use proximity to each individual seed word as the ranking method. However, many of the marked words used in the evaluation were never, or very rarely, ranked highly on the candidate list, regardless of which seed words were used.

Acknowledgements

This work was partly funded through the project StaViCTA by the framework grant "the Digitized Society Past, Present, and Future" with No. 2012-5659 from the Swedish Research Council (Vetenskapsrådet) and partly 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. The authors would also like to direct thanks to the reviewers for valuable comments.

References

Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. 2014. Overview of the NTCIR-11 MedNLP-2 task. In *Proceedings of NTCIR-11*.

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

Aaron Cohen, William R. Hersh, Christopher Dubay, and Kent Spackman. 2005. Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts. *BMC Bioinformatics*, 6(1):103.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 1–12.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING 1992*, pages 539–545.

Aron Henriksson, Mike Conway, Martin Duneld, and Wendy W. Chapman. 2013a. Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. In *Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA 2013)*, Washington DC, USA.

Aron Henriksson, Maria Skeppstedt, Maria Kvist, Martin Duneld, and Mike Conway. 2013b. Corpus-Driven Terminology Development: Populating Swedish SNOMED CT with Synonyms Extracted from Electronic Health Records. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 36–44, Sofia, Bulgaria. Association for Computational Linguistics.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius, and Martin Duneld. 2014. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *J Biomed Semantics*, 5(1):6.

Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. 2012. Enhancing clinical concept extraction with distributional semantics. *J Biomed Inform*, 45(1):129–40, Feb.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In L. R. Gleitman and A. K. Joshi, editors, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, Mahwah, NJ.

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*.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğanur, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics–Volume 2*, ACL '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

John McCrae and Nigel Collier. 2008. Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC Bioinformatics*, 9:159.

Danielle Mowery, Sumithra Velupillai, Brett R. South, Lee Christensen, David Martinez, Liadh Kelly, Lorraine Goeuriot, Noemie Elhadad, Sameer Pradhan, Guergana Savova, and Chapman Wendy W. 2014. Task 2: ShARe/CLEF eHealth Evaluation Lab 2014. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF2014 Working Notes*, volume 1180, pages 31–42. CEUR-WS, September.

Arvind Neelakantan and Michael Collins. 2014. Learning dictionaries for named entity recognition using minimal supervision. In Gosse Bouma and Yannick Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 452–461. The Association for Computer Linguistics.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of Languages in Biology and Medicine*.

Magnus Rosell, Martin Hassel, and Viggo Kann. 2009. Global evaluation of random indexing through Swedish word clustering compared to the people's dictionary of synonyms. In *Proceedings of Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Doctoral thesis, Stockholm University.

R. Sibson. 1973. SLINK: an optimally efficient algorithm for the single-link cluster method. *Comput. J.*, 16(1):30–34.

Maria Skeppstedt, Magnus Ahltorp, and Aron Henriksson. 2013. Vocabulary expansion by semantic extraction of medical terms. In *Proceedings of Languages in Biology and Medicine (LBM)*, Tokyo, Japan, December.

Kelly Smith, Beata Megyesi, Sumithra Velupillai, and Maria Kvist. 2014. Professional language in Swedish clinical text: Linguistic characterization and comparative studies. *Nordic Journal of Linguistics*, 37:297–323.

Özlem. Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556.

Sumithra Velupillai, Hercules Dalianis, and Maria Kvist. 2011. Factuality Levels of Diagnoses in Swedish Clinical Text. In A. Moen, S. K. Andersen, J. Aarts, and P. Hurlen, editors, *Proc. XXIII International Conference of the European Federation for Medical Informatics (User Centred Networked Health Care)*, pages 559–563, Oslo, August. IOS Press.

Sumithra Velupillai, Maria Skeppstedt, Maria Kvist, Danielle Mowery, Brian E Chapman, Hercules Dalianis, and Wendy W Chapman. 2014. Cue-based assertion classification for swedish clinical text–developing a lexicon for pycontextswe. *Artif Intell Med*, 61(3):137–44, Jul.

Sumithra Velupillai. 2012. *Shades of Certainty – Annotation and Classification of Swedish Medical Records*. Doctoral thesis, Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden, April.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra1, and János Csirik. 2008. The BioScope Corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Hong Yu and Eugene Agichtein. 2003. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, 1(19):340–349.

Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, 46(6):1088–1098. Special Section: Social Media Environments.