Corpus-Driven Terminology Development: Populating Swedish SNOMED CT with Synonyms Extracted from Electronic Health Records

Aron Henriksson[1], **Maria Skeppstedt**[1], **Maria Kvist**[1,2], **Martin Duneld**[1], **Mike Conway**[3]

[1]Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden
[2]Department of Learning, Informatics, Management and Ethics (LIME), Karolinska Institute, Sweden
[3]Division of Biomedical Informatics, University of California San Diego, USA

Abstract

The various ways in which one can refer to the same clinical concept needs to be accounted for in a semantic resource such as SNOMED CT. Developing terminological resources manually is, however, prohibitively expensive and likely to result in low coverage, especially given the high variability of language use in clinical text. To support this process, distributional methods can be employed in conjunction with a large corpus of electronic health records to extract synonym candidates for clinical terms. In this paper, we exemplify the potential of our proposed method using the Swedish version of SNOMED CT, which currently lacks synonyms. A medical expert inspects two thousand term pairs generated by two semantic spaces – one of which models multiword terms in addition to single words – for one hundred preferred terms of the semantic types *disorder* and *finding*.

1 Introduction

In recent years, the adoption of standardized terminologies for the representation of clinical concepts – and their textual instantiations – has enabled meaning-based retrieval of information from electronic health records (EHRs). By identifying and linking key facts in health records, the ever-growing stores of clinical documentation now available to us can more readily be processed and, ultimately, leveraged to improve the quality of care. SNOMED CT[1] has emerged as the *de facto* international terminology for representing clinical concepts in EHRs and is today used in more than fifty countries, despite only being available in a handful of languages[2]. Translations into several other languages are, however, under way[3]. This translation effort is essential for more widespread integration of SNOMED CT in EHR systems globally.

Translating a comprehensive[4] terminology such as SNOMED CT to an additional language is, however, a massive and expensive undertaking. A substantial part of this process involves enriching the terminology with synonyms in the target language. SNOMED CT has, for instance, recently been translated into Swedish; however, the Swedish version does not as yet contain synonyms. Methods and tools that can accelerate the language porting process in general and the synonym identification task in particular are clearly needed, not only to lower costs but also to increase the coverage of SNOMED CT in clinical text. Methods that can account for real-world language use in the clinical setting, then, as well as to changes over time, are particularly valuable.

This paper evaluates a semi-automatic method for the extraction of synonyms of SNOMED CT preferred terms using models of distributional semantics to induce semantic spaces from a large corpus of clinical text. In contrast to most approaches that exploit the notion of distributional similarity for synonym extraction, this method addresses the key problem of identifying synonymy between terms of varying length: a simple solution is proposed that effectively incorporates the notion of paraphrasing in a distributional framework. The semantic spaces – and, by extension, the method – are evaluated for their ability to extract synonyms of SNOMED CT terms of the semantic types *disorder* and *finding* in Swedish.

[1] `http://www.ihtsdo.org/snomed-ct/`

[2] SNOMED CT is currently available in US English, UK English, Spanish, Danish and Swedish.

[3] `http://www.ihtsdo.org/snomed-ct/snomed-ct0/different-languages/`

[4] SNOMED CT contains more than 300,000 active concepts and over a million relations.

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 36–44,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

2 Background

Synonymy is an aspect of semantics that concerns the fact that concepts can be instantiated using multiple linguistic expressions, or, viewed conversely, that multiple linguistic expressions can refer to the same concept. As synonymous expressions do not necessarily consist of single words, we sometimes speak of paraphrasing rather than synonymy (Androutsopoulos and Malakasiotis, 2010). This variability of language use needs to be accounted for in order to build high-quality natural language processing (NLP) and text mining systems. This is typically achieved by using thesauri or encoding textual instantiations of concepts in a semantic resource, e.g. an ontology. Creating such resources manually is, however, prohibitively expensive and likely to lead to low coverage, especially in the clinical genre where language use variability is exceptionally high (Meystre et al., 2008).

2.1 Synonym Extraction

As a result, the task of extracting synonyms – and other semantic relations – has long been a central challenge in the NLP research community, not least in the biomedical (Cohen and Hersh, 2005) and clinical (Meystre et al., 2008) domains. A wide range of techniques has been proposed for relation extraction in general and synonym extraction in particular – lexico-syntactic patterns (Hearst, 1992), distributional semantics (Dumais and Landauer, 1997) and graph-based models (Blondel et al., 2004) – from a variety of sources, including dictionaries (Blondel et al., 2004), linked data such as Wikipedia (Nakayama et al., 2007), as well as both monolingual (Hindle, 1990) and multilingual (van der Plas and Tiedemann, 2006) corpora. In recent years, ensemble methods have been applied to obtain better performance on the synonym extraction task, combining models from different families (Peirsman and Geeraerts, 2009), with different parameter settings (Henriksson et al., 2012) and induced from different data sources (Wu and Zhou, 2003).

In the context of biomedicine, the goal has often been to extract synonyms of gene and protein names from the biomedical literature (Yu and Agichtein, 2003; Cohen et al., 2005; McCrae and Collier, 2008). In the clinical domain, Conway and Chapman (2012) used a rule-based approach to generate potential synonyms from the BioPortal ontology web service, verifying candidate synonyms against a large clinical corpus. Zeng et al. (2012) used three query expansion methods for information retrieval of clinical documents and found that a model of distributional semantics – LDA-based topic modeling – generated the best synonyms. Henriksson et al. (2012) combined models of distributional semantics – random indexing and random permutation – to extract synonym candidates for Swedish MeSH[5] terms and possible abbreviation-definition pairs. In the context of SNOMED CT, distributional methods have been applied to capture synonymous relations between terms of varying length: 16-24% of English SNOMED CT synonyms present in a large clinical corpus were successfully identified in a list of twenty suggestions (Henriksson et al., 2013).

2.2 Distributional Semantics

Models of distributional semantics (see Cohen and Widdows (2009) for an overview of methods and their application in the biomedical domain) were initially motivated by the inability of the vector space model to account for synonymy, which had a negative impact on recall in information retrieval systems (Deerwester et al., 1990). The theoretical foundation underpinning such models of semantics is the *distributional hypothesis* (Harris, 1954), according to which words with similar meanings tend to appear in similar contexts. By exploiting the availability of large corpora, the meaning of terms can be modeled based on their distribution in different contexts. An estimate of the semantic relatedness between terms can then be quantified, thereby, in some sense, rendering semantics computable.

An obvious application of distributional semantics is the extraction of semantic relations between terms, such as synonymy, hyp(o/er)nymy and co-hyponymy (Panchenko, 2013). As synonyms are interchangeable in some contexts – and thus have similar distributional profiles – synonymy is certainly a semantic relation that should be captured. However, since hyp(o/er)nyms and co-hyponyms – in fact, even antonyms – are also likely to have similar distributional profiles, such semantic relations will be extracted too.

Many models of distributional semantics differ in how context vectors, representing term

[5] **Me**dical **S**ubject **H**eadings (`http://www.nlm.nih.gov/mesh`).

meaning, are constructed. They are typically derived from a term-context matrix that contains the (weighted, normalized) frequency with which terms occur in different contexts. Partly due to the intractability of working with such high-dimensional data, it is projected into a lower-dimensional (semantic) space, while approximately preserving the relative distances between data points. Methods that rely on computationally expensive dimensionality reduction techniques suffer from scalability issues.

Random Indexing

Random indexing (RI) (Kanerva et al., 2000) is a scalable and computationally efficient alternative in which explicit dimensionality reduction is avoided: a lower dimensionality d is instead chosen *a priori* as a model parameter and the d-dimensional context vectors are then constructed incrementally. Each unique term in the corpus is assigned a static index vector, consisting of zeros and a small number of randomly placed 1s and -1s[6]. Each term is also assigned an initially empty context vector, which is incrementally updated by adding the index vectors of the surrounding words within a sliding window, weighted by their distance to the target term. The semantic relatedness between two terms is then estimated by calculating, for instance, the cosine similarity between their context vectors.

Random Permutation

Random permutation (RP) (Sahlgren et al., 2008) is a modification of RI that attempts to take into account term order information by simply *permuting* (i.e. shifting) the index vectors according to their direction and distance from the target term before they are added to the context vector. RP has been shown to outperform RI on the synonym part of the TOEFL[7] test.

Model Parameters

The model parameters need to be configured for the task that the semantic space is to be used for. For instance, with a document-level context definition, *syntagmatic* relations are modeled, i.e. terms that belong to the same topic (<*car, motor, race*>), whereas, with a sliding window context definition, *paradigmatic* relations are modeled (<*car, automobile, vehicle*>) (Sahlgren, 2006). Synonymy is an instance of a paradigmatic relation.

The dimensionality has also been shown to be potentially very important, especially when the size of the vocabulary and the number of contexts[8] are large (Henriksson and Hassel, 2013).

3 Materials and Methods

The task of semi-automatically identifying synonyms of SNOMED CT preferred terms is here approached by, first, statistically identifying multiword terms in the data and treating them as compounds; then, performing a distributional analysis of a preprocessed clinical corpus to induce a semantic term space; and, finally, extracting the semantically most similar terms for each preferred term of interest.

The experimental setup can be broken down into the following steps: (1) data preparation, (2) term recognition, (3) model parameter tuning and (4) evaluation. Semantic spaces are induced with different parameter configurations on two dataset variants: one with unigram terms only and one that also includes multiword terms. The model parameters are tuned using MeSH, which contains synonyms for Swedish. The best parameter settings for each of the two dataset variants are then employed in the final evaluation, where a medical expert inspects one hundred term lists extracted for SNOMED CT preferred terms belonging to the semantic types *disorder* and *finding*.

3.1 Data Preparation

The data used to induce the semantic spaces is extracted from the Stockholm EPR Corpus (Dalianis et al., 2009), which contains Swedish health records from the Karolinska University Hospital in Stockholm[9]. The subset (∼33 million tokens) used in these experiments comprises all forms of text-based records – i.e., clinical notes – from a large variety of clinical practices. The documents in the corpus are initially preprocessed by simply lowercasing tokens and removing punctuation and digits. Lemmatization is not performed, as we want to be able to capture morphological

[6] By generating sparse vectors of a sufficiently high dimensionality in this way, the context representations will be *nearly* orthogonal.

[7] Test Of English as a Foreign Language

[8] The vocabulary size and the number of contexts are equivalent when employing a window context definition.

[9] This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/834-31/5.

variants of terms; stop-word filtering is not performed, as traditional stop words – for instance, high-frequency function words – could potentially be constituents of multiword terms.

3.2 Term Recognition

Multiword terms are extracted statistically from the corpus using the C-value statistic (Frantzi and Ananiadou, 1996; Frantzi et al., 2000). This technique has been used successfully for term recognition in the biomedical domain, largely due to its ability to handle nested terms (Zhang et al., 2008). Using the C-value statistic for term recognition first requires a list of candidate terms, for which the C-value can then be calculated. Here, this is simply produced by extracting n-grams – unigrams, bigrams and trigrams – from the corpus with TEXT-NSP (Banerjee and Pedersen, 2003). The statistic is based on term frequency and term length (number of words); if a candidate term is part of a longer candidate term (as will be the case for practically all unigram and bigram terms), the number and frequency of those longer terms are also taken into account (Figure 1).

In order to improve the quality of the extracted terms, a number of filtering rules is applied to the generated term list: terms that begin and/or end with certain words, e.g. prepositions and articles, are removed. The term list – ranked according to C-value – is further modified by giving priority to terms of particular interest, e.g. SNOMED CT *disorder* and *finding* preferred terms: these are moved to the top of the list, regardless of their C-value. As a result, the statistical foundation on which the distributional method bases its semantic representation will effectively be strengthened.

The term list is then used to perform exact string matching on the entire corpus: multiword terms with a higher C-value than their constituents are concatenated. We thereby treat multiword terms as separate (term) types with distinct distributions in the data, different from those of their constituents.

3.3 Model Parameter Tuning

Term spaces with different parameter configurations are induced from the two dataset variants: one containing only unigram terms (*Unigram Word Spaces*) and one containing also multiword terms (*Multiword Term Spaces*). The following model parameters are tuned:

- Distributional Model: Random indexing (RI) vs. Random permutation (RP)
- Context Window Size: 2+2, 4+4, 8+8 surrounding terms (*left+right* of the target term)
- Dimensionality: 1000, 2000, 3000

As the Swedish version of SNOMED CT currently does not contain synonyms, it cannot be used to perform the parameter tuning automatically. This is instead done with the Swedish version of MeSH, which is one of the very few standard terminologies that contains synonyms for medical terms in Swedish. However, as the optimal parameter configurations for capturing synonymy are not necessarily identical for all semantic types, the parameter tuning is performed by evaluating the semantic spaces for their ability to identify synonyms of MeSH terms that belong to the categories *Disease or Syndrome* and *Sign or Symptom*. These particular categories are simply chosen as they, to a reasonable extent, seem to correspond to the SNOMED CT semantic types studied in this paper, namely *Disorder* and *Finding*. Only synonym pairs that appear at least fifty times in each of the dataset variants are included (155 for *Unigram Word Spaces* and 123 for *Multiword Term Spaces*), as the statistical foundation for terms that only occur rarely in the data may not be sufficiently solid. In these *Multiword Term Spaces*, the MeSH terms – but not the synonyms – are given precedence in the term list. A term is provided as input to a semantic space and the twenty semantically most similar terms are output, provided that they also appear at least fifty times in the data. Recall Top 20 is calculated for each input term: *what proportion of the MeSH synonyms are identified in a list of twenty suggestions?* Since each synonym pair must appear at least fifty times in the corresponding dataset variant, it should be duly noted that the optimization sets will not be identical, which in turn means that the results of the *Unigram Word Spaces* and the *Multiword Term Spaces* are not directly comparable. The optimal parameter configuration, then, may be different when also multiword terms are modeled.

3.4 Evaluation

The optimal parameter configuration for each dataset variant is employed in the final evaluation. In this *Multiword Term Space*, the SNOMED CT

$$C\text{-}value(a) = \begin{cases} \log_2 |a| \cdot f(a) & \text{if a is not nested} \\ \log_2 |a| \cdot (f(a) - \frac{1}{P(Ta)} \sum_{b \epsilon Ta} f(b)) & \text{otherwise} \end{cases}$$

a	= candidate term	Ta	= set of extracted candidate terms that contain a
b	= longer candidate terms	$P(Ta)$	= number of candidate terms in Ta
$f(a)$	= term frequency of a	$f(b)$	= term frequency of longer candidate term b
$\lvert a \rvert$	= length of candidate term (number of words)		

Figure 1: *C-Value Formula*. The formula for calculating C-value of candidate terms.

preferred terms of interest, rather than the MeSH terms, are prioritized in the term list. The semantic spaces – and, in effect, the method – are primarily evaluated for their ability to identify synonyms of SNOMED CT preferred terms, in this case of concepts that belong to the semantic types *disorder* and *finding*. The need to identify synonyms for these semantic types is clear, as it has been shown that the coverage of SNOMED CT for mentions of disorders (38%) and, in particular, findings (23%) in Swedish clinical text is low (Skeppstedt et al., 2012). Since the Swedish version of SNOMED CT currently lacks synonyms, the evaluation reasonably needs to be manual, as there is no reference standard. One option, then, could be to choose a random sample of preferred terms to use in the evaluation. A potential drawback of such a(n) (unguided) selection is that many concepts in the English version of SNOMED CT do not have any synonymous terms, which might lead to evaluators spending valuable time looking for something which does not exist. An alternative approach, which is assumed here, is to inspect concepts that have many synonyms in the English version of SNOMED CT. The fact that some concepts have many textual instantiations in one language does not necessarily imply that they also have many textual instantiations in another language. This, however, seems to be the case when comparing the English and Swedish versions of MeSH: terms[10] that have the most synonyms in the English version tend to have at least one synonym in the Swedish version to a larger extent than a random selection of terms (60% and 62% of the terms in the Swedish version have at least one synonym when looking at the top 100 and top 50 terms with the most synonyms in the English version, compared to 41% overall in the Swedish version).

For the two dataset variants, we thus select 25 SNOMED CT preferred terms for each semantic type – *disorder* and *finding* – that (1) have the most synonyms in the English version and (2) occur at least fifty times in the data. In total, fifty terms are input to the *Unigram Word Space* and another fifty terms (potentially with some overlap) are input to the *Multiword Term Space*. A medical expert inspects the twenty semantically most similar terms for each input term. Synonymy is here the primary semantic relation of interest, but the semantic spaces are also evaluated for their ability, or tendency, to extract other semantic term relations: hypernyms or hyponyms, co-hyponyms, antonyms, as well as *disorder-finding* relations.

4 Results

The term recognition and concatenation of multiword terms naturally affect some properties of the dataset variants, such as the vocabulary size (number of types) and the type-token ratio. The *Unigram Word Space* contains 381,553 types and an average of 86.54 tokens/type, while the *Multiword Term Space* contains 2,223,953 types and an average of 9.72 tokens/type. This, in turn, may have an effect on which parameter configuration is 'optimal' for the synonym extraction task. In fact, this seems to be the case when tuning the parameters for the two dataset variants. For the *Unigram Word Spaces*, random indexing with a sliding context window of 8+8 terms and a dimensionality of 2000 seems to work best, whereas for the *Multiword Term Spaces*, random permutation with a sliding window context of 4+4 terms and a dimensionality of 3000 works better (Table 1).

When these parameter configurations are applied to the SNOMED CT terms, a total of 40 synonyms are extracted by the *Unigram Word Space* and 33 synonyms by the *Multiword Term Space* (Table 2). On average, 0.80 and 0.66 synonyms are extracted per preferred term, respectively. The number of identified synonyms per input term varies significantly: for some, none; for others, up to ten. Other semantic relations are also extracted

[10] These calculations are based on MeSH terms that belong to the categories *Disease or Syndrome* and *Sign or Symptom*.

	Unigram Word Spaces						**Multiword Term Spaces**					
	RI			**RP**			**RI**			**RP**		
Sliding Window →	**2+2**	**4+4**	**8+8**	**2+2**	**4+4**	**8+8**	**2+2**	**4+4**	**8+8**	**2+2**	**4+4**	**8+8**
1000 dimensions	0.43	0.47	0.48	0.41	0.45	0.42	0.21	0.25	0.26	0.25	0.26	0.24
2000 dimensions	0.43	0.48	**0.49**	0.48	0.48	0.43	0.21	0.24	0.25	0.25	0.25	0.24
3000 dimensions	0.44	0.47	0.48	0.46	0.45	0.43	0.22	0.24	0.24	0.23	**0.27**	0.25

Table 1: *Model Parameter Tuning*. Results, reported as recall top 20, for MeSH synonyms that appear at least 50 times in each of the dataset variants (unigram vs. multiword). Random indexing (RI) and Random permutation (RP) term spaces were built with different context window sizes (2+2, 4+4, 8+8 surrounding terms) and dimensionality (1000, 2000, 3000).

by the semantic spaces: mainly co-hyponyms, but also hypernyms and hyponyms, antonyms and *disorder-finding* relations. The *Unigram Word Space* extracts, on average, 0.52 hypernyms or hyponyms, 1.8 co-hyponyms, 0.1 antonyms and 0.34 *disorder-finding* relations. The *Multiword Term Space* extracts, on average, 0.16 hypernyms or hyponyms, 1.1 co-hyponyms, 0.14 antonyms and 0.66 *disorder-finding* relations. In general, more of the above semantic relations are extracted by the *Unigram Word Space* than by the *Multiword Term Space* (178 vs. 136). It is, however, interesting to note that almost twice as many *disorder-finding* relations are extracted by the latter compared to the former. Of course, none of the relations extracted by the *Unigram Word Space* involve a multiword term; on the other hand, more than half (around 57%) of the relations extracted by the *Multiword Term Space* involve at least one multiword term.

Both semantic spaces identify more synonyms of preferred terms that belong to the semantic type *finding* than *disorder* (in total 56 vs. 39). The same holds true for hyp(er/o)nyms and co-hypnoyms; however, the converse is true for antonyms and *disorder-finding* relations.

5 Discussion

The results demonstrate that it is indeed possible to extract synonyms of medical terms by performing a distributional analysis of a large corpus of clinical text – unigram-unigram relations, as well as unigram-multiword and multiword-unigram relations. It is also clear, however, that other semantically related terms share distributional profiles to a similar degree as synonymous terms. The predominance of the other semantic relations, except for antonymy, in the term lists can reasonably be explained by the simple fact that there exist more hypernyms, hyponyms, co-hyponyms and *disorder-finding* relations than synonyms (or antonyms).

It is also evident that more semantic relations, and indeed more synonyms, are extracted by the *Unigram Word Space* than the *Multiword Term Space*. Again, it is important to underline that the results cannot be compared without due qualification since the evaluation sets are not identical: the *Unigram Word Space* does not contain any multiword terms, for instance. The ability to model multiword terms in a distributional framework and to handle semantic composition – i.e., how meaning is, and sometimes is not, composed by the meaning of its constituents – has long been an endeavor in the NLP research community (Sag et al., 2002; Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011; Mitchell, 2011). Treating multiword terms as compound tokens is a simple and rather straightforward approach, which also makes intuitive sense: rather than treat individual words as clearly delineated bearers of meaning, identify *semantic units* – regardless of term length – and model their distributional profiles. Unfortunately, there are problems with this approach. First, the attendant increase in vocabulary size entails a lower tokens-type ratio, which in turn means that the statistical foundation for terms will weaken. In this case, the average token-type ratio decreased from 86.54 to 9.72. This approach therefore requires access to a sufficiently large corpus. Second, the inflation in vocabulary size entails a corresponding increase in the number of vectors in the semantic space. This not only requires more memory; to ensure that the crucial *near-orthogonality* property[11] of RI-based models is maintained, the dimensionality has to be suffi-

[11] Random indexing assumes that the index vectors – representing distinct contexts – are *nearly* orthogonal.

	Unigram Word Space		**Multiword Term Space**	
	DISORDER	*FINDING*	*DISORDER*	*FINDING*
Synonyms				
sum	18	22	16	17
average	0.72	0.88	0.64	0.68
≥ 1 / preferred term	12	12	8	6
involves mwe	-	-	10	13
Hyp(er/o)nyms				
sum	12	14	4	4
average	0.48	0.56	0.16	0.16
≥ 1 / preferred term	6	8	4	3
involves mwe	-	-	3	3
Co-hyponyms				
sum	34	56	22	33
average	1.36	2.24	0.88	1.32
≥ 1 / preferred term	14	17	10	13
involves mwe	-	-	19	15
Antonyms				
sum	3	2	4	3
average	0.12	0.08	0.16	0.12
≥ 1 / preferred term	3	2	3	3
involves mwe	-	-	0	1
Disorder-Finding				
sum	11	6	28	5
average	0.44	0.24	1.12	0.2
≥ 1 / preferred term	6	5	12	5
involves mwe	-	-	11	2

Table 2: *Evaluation Results*. The types of semantic relations extracted among the twenty most semantically similar terms of 25 *DISORDER* and 25 *FINDING* SNOMED CT preferred terms from each semantic space. *Sum* is the total number of identified relevant terms. *Average* is the average number of relevant terms per preferred term. *≥ 1 / preferred term* is the number of preferred terms for which at least one relevant term is identified. *Involves mwe* is the number of relevant relations where either the preferred term or the relevant term is a multiword expression.

ciently large in relation to the number of contexts (represented by index vectors). In the *Multiword Term Space* the vocabulary size is over two million (compared to less than 400,000 in the *Unigram Word Space*). A dimensionality of 3000 is likely insufficient to ensure that each term type has an initial distinct and uncorrelated representation. In the evaluation, there were several examples where two groups of terms – semantically homogenous within each group, but semantically heterogenous across groups – co-existed in the same term list: these 'topics' had seemingly collapsed into the same subspace. Despite these problems, it should be recognized that the *Multiword Term Space* is, in fact, able to retrieve 23 synonymous relations that involve at least one multiword term. The *Unigram Word Space* cannot retrieve any such relations.

The ability to extract high-quality terms would seem to be an important prerequisite for this approach to modeling multiword terms in a distributional framework. However, despite employing a rather simple means of extracting terms – without using any syntactic information – the terms that actually appeared in the lists of semantically related terms were mostly reasonable. This perhaps indicates that the term recognition task does not need to be perfect: terms of interest, of course, need to be identified, but some noise in the form of bad terms might be acceptable. A weakness of the term recognition part is, however, that too many terms were identified, which in turn led to the aforementioned inflation in vocabulary size.

Limiting the number of multiword terms in the initial term list – for instance by extracting syntactic phrases as candidate terms – could provide a possible solution to this problem.

Overall, more synonyms were identified for the semantic type *finding* than for *disorder*. One possible explanation for this could be that there are more ways of describing a finding than a disorder – not all semantic types can be assumed to have the same number of synonyms. The same holds true for all other semantic relations except for *disorder-finding*, where disorders generated a much larger number of distributionally similar findings than vice versa. This could perhaps also be explained by the possible higher number of synonyms for *finding* than *disorder*.

When this method was evaluated using the English version of SNOMED CT, 16-24% of known synonyms were identified (Henriksson et al., 2013). In this case, however, we extracted synonym candidates for terms that may or may not have synonyms. This is thus a scenario that more closely resembles how this method would actually be used in a real-life setting to populate a terminology with synonyms. Although the comparison with MeSH showed that terms with many synonyms in English also tend to have at least one synonym in Swedish, approximately 40% of them did not have any synonyms. It is thus not certain that the terms used in this evaluation all have at least one synonym, which was also noted by the evaluator in this study.

6 Conclusions

In this study, we have demonstrated a method that could potentially be used to expedite the language porting process of terminologies such as SNOMED CT. With access to a large corpus of clinical text in the target language and an initial set of terms, this language-independent method is able to extract and present candidate synonyms to the lexicographer, thereby providing valuable support for semi-automatic terminology development. A means to model multiword terms in a distributional framework is an important feature of the method and is crucial for the synonym extraction task.

Acknowledgments

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, and partly funded by the Stockholm University Academic Initiative through the Interlock project. Finally, we would like to thank the reviewers for their constructive feedback.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Satanjeev Banerjee and Ted Pedersen. 2003. The Design, Implementation, and Use of the Ngram Statistic Package. In *Proceedings of CICLing*, pages 370–381.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In *Proceedings of EMNLP*, pages 1183–1193.

Vincent D. Blondel, Anahí Gajardo, Maureen Heymans, Pierre Senellart, and Paul Van Dooren. 2004. A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching. *SIAM Review*, 46(4):647–666.

Aaron M. Cohen and William R. Hersh. 2005. A Survey of Current Work in Biomedical Text Mining. *Briefings in Bioinformatics*, 6(1):57–71.

Trevor Cohen and Dominic Widdows. 2009. Empirical Distributional Semantics: Methods and Biomedical Applications. *J Biomed Inform*, 42(2):390–405.

AM Cohen, WR Hersh, C Dubay, and K Spackman. 2005. Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts. *BMC Bioinformatics*, 6(1):103.

Mike Conway and Wendy W. Chapman. 2012. Discovering Lexical Instantiations of Clinical Concepts using Web Services, WordNet and Corpus Resources. In *AMIA Fall Symposium*, page 1604.

Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. 2009. The Stockholm EPR Corpus: Characteristics and Some Initial Findings. In *Proceedings of ISHIMR*, pages 243–249.

Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Susan T. Dumais and Thomas K. Landauer. 1997. A Solution to Plato's Problem: The Latent Semantic

Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

Katerina Frantzi and Sophia Ananiadou. 1996. Extracting Nested Collocations. In *Proceedings of COLING*, pages 41–46.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic Recognition of Multi-Word Terms: The C-value/NC-value Method. *International Journal on Digital Libraries*, 3(2):115–130.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental Support for a Categorical Compositional Distributional Model of Meaning. In *Proceedings of EMNLP*, pages 1394–1404.

Zellig S. Harris. 1954. Distributional Structure. *Word*, 10:146–162.

Marti Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of COLING*, pages 539–545.

Aron Henriksson and Martin Hassel. 2013. Optimizing the Dimensionality of Clinical Term Spaces for Improved Diagnosis Coding Support. In *Proceedings of Louhi*.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Ann-Marie Eklund, and Vidas Daudaravicius. 2012. Synonym Extraction of Medical Terms from Clinical Text Using Combinations of Word Space Models. In *Proceedings of SMBM*, pages 10–17.

Aron Henriksson, Mike Conway, Martin Duneld, and Wendy W. Chapman. 2013. Identifying Synonymy between SNOMED Clinical Terms of Varying Length Using Distributional Analysis of Electronic Health Records. In *AMIA Annual Symposium (submitted)*.

Donald Hindle. 1990. Noun Classification from Predicate-Argument Structures. In *Proceedings of ACL*, pages 268–275.

Pentti Kanerva, Jan Kristofersson, and Anders Holst. 2000. Random Indexing of Text Samples for Latent Semantic Analysis. In *Proceedings CogSci*, page 1036.

John McCrae and Nigel Collier. 2008. Synonym Set Extraction from the Biomedical Literature by Lexical Pattern Discovery. *BMC Bioinformatics*, 9(1):159.

Stéphane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler, John F. Hurdle, et al. 2008. Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *Yearb Med Inform*, 35:128–44.

Jeffrey Mitchell. 2011. *Composition in Distributional Models of Semantics*. Ph.D. thesis, University of Edinburgh.

Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2007. Wikipedia Mining for an Association Web Thesaurus Construction. In *Proceedings of WISE*, pages 322–334.

Alexander Panchenko. 2013. *Similarity Measures for Semantic Relation Extraction*. Ph.D. thesis, PhD thesis, Université catholique de Louvain & Bauman Moscow State Technical University.

Yves Peirsman and Dirk Geeraerts. 2009. Predicting Strong Associations on the Basis of Corpus Data. In *Proceedings of EACL*, pages 648–656.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of CICLing*, pages 1–15.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a Means to Encode Order in Word Space. In *Proceedings of CogSci*, pages 1300–1305.

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.

Maria Skeppstedt, Maria Kvist, and Hercules Dalianis. 2012. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. In *Proceedings of LREC*, pages 1250–1257.

Lonneke van der Plas and Jörg Tiedemann. 2006. Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity. In *Proceedings of COLING/ACL*, pages 866–873.

Hua Wu and Ming Zhou. 2003. Optimizing Synonym Extraction Using Monolingual and Bilingual Resources. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 72–79.

Hong Yu and Eugene Agichtein. 2003. Extracting Synonymous Gene and Protein Terms from Biological Literature. *Bioinformatics*, 19(suppl 1):i340–i349.

Qing T Zeng, Doug Redd, Thomas Rindflesch, and Jonathan Nebeker. 2012. Synonym, Topic Model and Predicate-Based Query Expansion for Retrieving Clinical Documents. In *Proceedings AMIA Annual Symposium*, pages 1050–9.

Ziqi Zhang, José Iria, Christopher Brewster, and Fabio Ciravegna. 2008. A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of LREC*.