Identifying Synonymy between SNOMED Clinical Terms of Varying Length Using Distributional Analysis of Electronic Health Records

Aron Henriksson, MS[1], Mike Conway, PhD[2], Martin Duneld, PhD[1], Wendy W. Chapman, PhD[3]
[1] Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden
[2] Division of Behavioral Medicine, Department of Family & Preventive Medicine, University of California, San Diego, USA
[3] Division of Biomedical Informatics, Department of Medicine, University of California, San Diego, USA

Abstract

Medical terminologies and ontologies are important tools for natural language processing of health record narratives. To account for the variability of language use, synonyms need to be stored in a semantic resource as textual instantiations of a concept. Developing such resources manually is, however, prohibitively expensive and likely to result in low coverage. To facilitate and expedite the process of lexical resource development, distributional analysis of large corpora provides a powerful data-driven means of (semi-)automatically identifying semantic relations, including synonymy, between terms. In this paper, we demonstrate how distributional analysis of a large corpus of electronic health records – the MIMIC-II database – can be employed to extract synonyms of SNOMED CT preferred terms. A distinctive feature of our method is its ability to identify synonymous relations between terms of varying length.

Introduction and Motivation

Terminological and ontological standards are an important and integral part of workflow and standards in clinical care. In particular, SNOMED CT[1] has become the *de facto* standard for the representation of clinical concepts in Electronic Health Records. However, SNOMED CT is currently only available in British and American English, Spanish, Danish, and Swedish (with translations into French and Lithuanian in process)[2]. In order to accelerate the adoption of SNOMED CT (and by extension, Electronic Health Records) internationally, it is clear that the development of new methods and tools to expedite the language porting process is of vital importance.

This paper presents and evaluates a semi-automatic – and language agnostic – method for the extraction of synonyms of SNOMED CT preferred terms using distributional similarity techniques in conjunction with a large corpus of clinical text (the MIMIC-II database[3]). Key applications of the technique include:

1. Expediting SNOMED CT language porting efforts using semi-automatic identification of synonyms for preferred terms
2. Augmenting the current English versions of SNOMED CT with additional synonyms

In comparison to current rule-based synonym extraction techniques, our proposed method has two major advantages:

1. As the method uses statistical techniques (i.e. distributional similarity methods), it is agnostic with respect to language. That is, in order to identify new synonyms, all that is required is a clinical corpus of sufficient size in the target language
2. Unlike most approaches that use distributional similarity, our proposed method addresses the problem of identifying synonymy between terms of varying length – a key limitation in traditional distributional similarity approaches

In this paper, we begin by presenting some relevant background literature (including describing related work in distributional similarity and synonym extraction); then we describe the materials and methods used in this research (in

particular corpus resources and software tools), before going on to set out the results of our analysis. We conclude the paper with a discussion of our results and a short conclusion.

Background

In this section, we will first describe the structure of SNOMED CT, before going on to discuss relevant work related to synonym extraction. Finally, we will present some opportunities and challenges associated with using distributional similarity methods for synonym extraction.

SNOMED CT

In recent years, SNOMED CT has become the *de facto* terminological standard for representing clinical concepts in Electronic Health Records[4]. SNOMED CT's scope includes *clinical findings*, *procedures*, *body structures*, and *social contexts* linked together through relationships (the most important of which is the hierarchical *IS_A* relationship). There are more than 300,000 active concepts in SNOMED CT and over a million relations[1]. Each concept consists of a:

1. *Concept ID*: A unique numerical identifier
2. *Fully Specified Name*: An unambiguous string used to name a concept
3. *Preferred Term*: A common phrase or word used by clinicians to name a concept. Each concept has precisely one preferred term in a given language. In contrast to the fully specified name, the preferred term is not necessarily unique and can be a synonym or preferred name for a different concept
4. *Synonym*: A term that can be used as an acceptable alternative to the preferred term. A concept can have zero or more synonyms

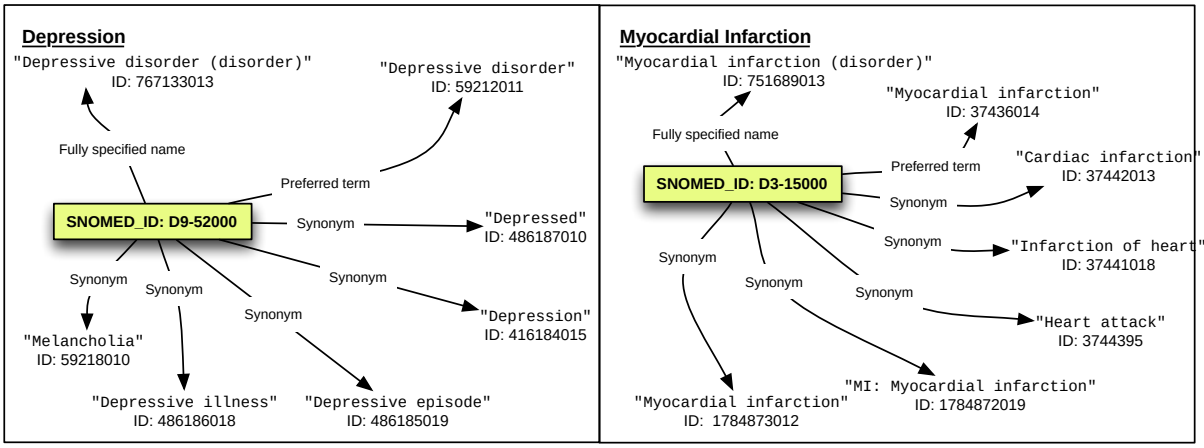

Figure 1: Example SNOMED CT concepts: *depression* and *myocardial infarction*.

Figure 1 shows two example SNOMED CT concepts: *depression* and *myocardial infarction*. Note that in the *depression* example, the preferred term "depressive disorder" maps to single-word terms like "depressed" and "depression". Furthermore, it can be noted that the synonym "melancholia" does not contain the term "depression" or one of its morphological variants.

Synonymy

Previous research on synonym extraction in the biomedical informatics literature has utilized diverse methodologies. In the context of information retrieval from clinical documents, Zeng et al.[5] used three query expansion methods – reading synonyms and lexical variants directly from the UMLS[6], generating topic models from clinical documents, and mining the SemRep[7] predication database – and found that an entirely corpus-based statistical method (i.e. topic

modeling) generated the best synonyms. Conway & Chapman[8] used a rule-based approach to generate potential synonyms from the BioPortal ontology web service, verifying the acceptability of candidate synonyms by checking for their presence in a very large corpus of clinical text, with the goal of populating a lexical-oriented knowledge organization system. In the Natural Language Processing community, there is a rich tradition of using lexico-syntactic patterns to extract synonyms (and other) relations[9].

Distributional Semantics

Models of distributional semantics exploit large corpora to capture the meaning of terms based on their distribution in different contexts. The theoretical foundation underlying such models is the *distributional hypothesis*[10], which states that words with similar meanings tend to appear in similar contexts. Distributional methods have become popular with the increasing availability of large corpora and are attractive due to their ability, in some sense, to render semantics computable: an estimate of the semantic relatedness between two terms can be quantified. These methods have been applied successfully to a range of natural language processing tasks, including document retrieval, synonym tests and word sense disambiguation[11]. An obvious use case of distributional methods is for the extraction of semantic relations, such as synonyms, hypernyms and co-hyponyms (terms with a common hypernym)[12]. Ideally, one would want to differentiate between such semantic relations; however, with these methods, the semantic relation between two distributionally similar terms is unlabeled. As synonyms are interchangeable in most contexts – meaning that they will have similar distributional profiles – synonymy is certainly a semantic relation that will be captured. However, since hypernyms and hyponyms – in fact, even antonyms – are also likely to occur in similar contexts, such semantic relations will likewise be extracted.

Distributional methods can be usefully divided into *spatial models* and *probabilistic models*. Spatial models represent terms as vectors in a high-dimensional space, based on the frequency with which they appear in different contexts, and where proximity between vectors is assumed to indicate semantic relatedness. Probabilistic models view documents as a mixture of topics and represent terms according to the probability of their occurrence during the discussion of each topic: two terms that share similar topic distributions are assumed to be semantically related. There are pros and cons of each approach; however, scalable versions of spatial models have proved to work well for very large corpora.[11]

Spatial models differ mainly in the way context vectors, representing term meaning, are constructed. In many methods, they are derived from an initial term-context matrix that contains the (weighted, normalized) frequency with which the terms occur in different contexts. The main problem with using these term-by-context vectors is their dimensionality, equal to the number of contexts (e.g. # of documents / vocabulary size). The solution is to project the high-dimensional data into a lower-dimensional space, while approximately preserving the relative distances between data points. In latent semantic analysis (LSA)[13], the term-context matrix is reduced by an expensive matrix factorization technique known as singular value decomposition. Random indexing (RI)[14] is a scalable and computationally efficient alternative to LSA, in which explicit dimensionality reduction is circumvented: a lower dimensionality d is instead chosen *a priori* as a model parameter and the d-dimensional context vectors are then constructed incrementally. RI can be viewed as a two-step operation:

1. Each context (e.g. each document or unique term) is assigned a sparse, ternary and randomly generated *index vector*: a small number (1-2%) of 1s and -1s are randomly distributed; the rest of the elements are set to zero. By generating sparse vectors of a sufficiently high dimensionality in this way, the context representations will, with a high probability, be *nearly* orthogonal.
2. Each unique term is also assigned an initially empty *context vector* of the same dimensionality. The context vectors are then incrementally populated with context information by adding the index vectors of the contexts in which the target term appears.

There are a number of model parameters that need to be configured according to the task that the induced *term space* will be used for. For instance, the types of semantic relations captured by an RI-based model depends on the context definition[15]. By employing a document-level context definition, relying on direct co-occurrences, one models *syntagmatic* relations. That is, two terms that frequently co-occur in the same documents are likely to be about the same topic, e.g. <*car, motor, race*>. By employing a sliding window context definition, where the index vectors of

the surrounding terms within a, usually small, window are added to the context vector of the target term, one models *paradigmatic* relations. That is, two terms that frequently occur with the same set of words – i.e. share neighbors – but do not necessarily co-occur themselves, are semantically similar, e.g. *<car, automobile, vehicle>*. Synonymy is an instance of a paradigmatic relation.

RI, in its original conception, does not take into full account term order information, except by giving increasingly less weight to index vectors of terms as the distance from the target term increases. Random permutation (RP)[16] is an elegant modification of RI that attempts to remedy this 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. RI has performed well on tasks such as taking the TOEFL (Test of English as a Foreign Language) test. However, by incorporating term order information, RP was shown to outperform RI on this particular task[16]. Combining RI and RP models has been demonstrated to yield improved results on the synonym extraction task[17].

The predefined dimensionality is yet another model parameter that has been shown to be potentially very important, especially when the number of contexts (the size of the vocabulary) is large, as it often is in the clinical domain. Since the traditional way of using distributional semantics is to model only unigram-unigram relations – a limitation when wishing to model the semantics of phrases and longer textual sequences – a possible solution is to identify and model multiword terms as single tokens. This will, however, lead to an explosion in the size of the vocabulary, necessitating a larger dimensionality. In short, dimensionality and other model parameters need to be tuned for the dataset and task at hand.[18]

Materials and Methods

The method and experimental setup can be summarized in the following steps: (1) data preparation, (2) term extraction and identification, (3) model building and parameter tuning, and (4) evaluation (Figure 2). Term spaces are constructed with various parameter settings on two dataset variants: one with unigram terms and one with multiword terms. The models – and, in effect, the method – are evaluated for their ability to identify synonyms of SNOMED CT preferred terms. After optimizing the parameter settings for each group of models on a development set, the best models are evaluated on unseen data.

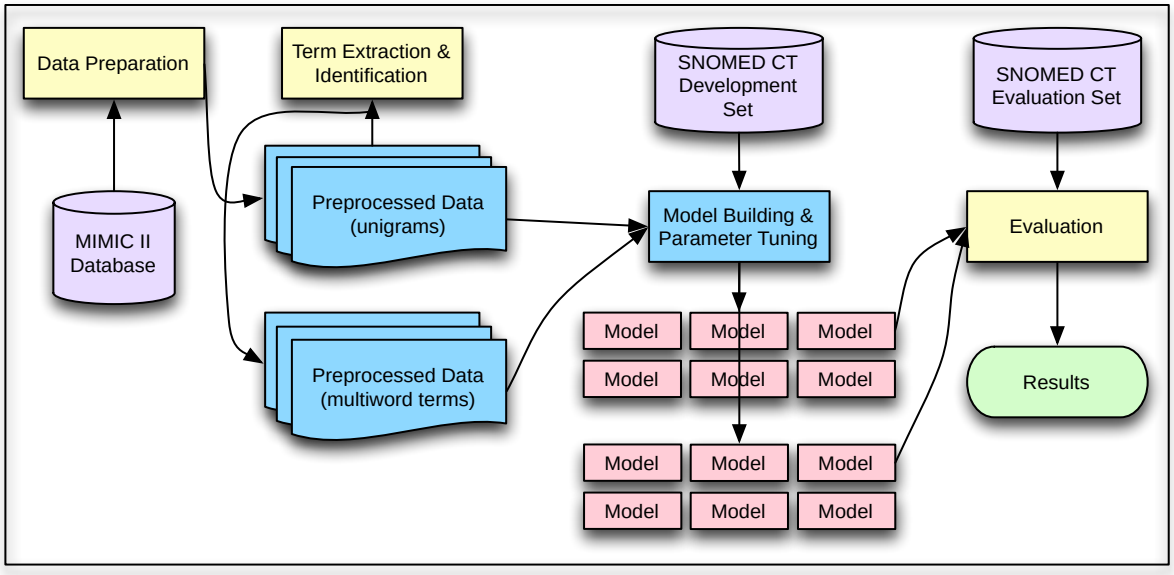

Figure 2: An overview of the process and the experimental setup.

In Step 1, the clinical data from which the term spaces will be induced is extracted from the MIMIC-II database and preprocessed. MIMIC-II[3] is a publicly available database encompassing clinical data for over 40,000 hospital

stays of more than 32,000 patients, collected over a seven-year period (2001-2007) from intensive care units (medical, surgical, coronary and cardiac surgery recovery) at Boston's Beth Israel Deaconess Medical Center. In addition to various structured data, such as laboratory results and ICD-9 diagnosis codes, the database contains text-based records, including nursing progress notes, discharge summaries and radiology interpretations. We create a corpus comprising all text-based records (∼250 million tokens) from the MIMIC-II database. The documents in the corpus are then preprocessed to remove metadata, such as headings (e.g. *FINAL REPORT*), incomplete sentence fragments, such as enumerations (of e.g. medications), as well as digits and punctuation marks.

In Step 2, we extract and identify multiword terms in the corpus. This will allow us to extract synonymous relations between terms of varying length. This is done by first extracting all unigrams, bigrams and trigrams from the corpus with TEXT-NSP[19] and treating them as candidate terms for which the C-value is then calculated. The C-value statistic[20 21] has been used successfully for term recognition in the biomedical domain, largely due to its ability to handle nested terms[22]. It is based on term frequency and term length (number of words); if a candidate term is part of a longer candidate term, it also takes into account how many other terms it is part of and how frequent those longer terms are (Figure 3). By extracting n-grams and then ranking the n-grams according to their C-value, we are incorporating the notions of both *unithood* – indicating collocation strength – and *termhood* – indicating the association strength of a term to domain concepts[22]. In this still rather simple approach to term extraction, however, we do not take any other linguistic knowledge into account. As a simple remedy for this, we create a number of filtering rules that remove terms beginning and/or ending with certain words, e.g. prepositions (*in*, *from*, *for*) and articles (*a*, *the*). Another alteration to the term list – now ranked according to C-value – is to give precedence to SNOMED CT preferred terms by adding/moving them to the top of the list, regardless of their C-value (or failed identification). The reason for this is that we are aiming to identify SNOMED CT synonyms of preferred terms and, by giving precedence to preferred terms – but not synonyms, as that would constitute cheating – we are effectively strengthening the statistical foundation on which the distributional method bases its semantic representation. 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 thus treat multiword terms as separate tokens with their own particular distributions in the data, to a greater or lesser extent different from those of their constituents.

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

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

Figure 3: The formula for calculating C-value of candidate terms.

In Step 3, term spaces are induced from the dataset variants: one containing only unigram terms (*UNIGRAM TERMS*) and one containing also longer terms: unigram, bigram and trigram terms (*MULTIWORD TERMS*). The following model parameters are experimented with:

- **Model Type**: random indexing (RI), random permutation (RP). *Does the method for synonym identification benefit from incorporating word order information?*
- **Sliding Window**: 1+1, 2+2, 3+3, 4+4, 5+5, 6+6 surrounding terms. *Which paradigmatic context definition is most beneficial for synonym identification?*
- **Dimensionality**: 500, 1000, 1500 dimensions. *How does the dimensionality affect the method's ability to identify synonyms, and is the impact greater when the vocabulary size grows exponentially as it does when treating multiword terms as single tokens?*

Evaluation takes place in both Step 3 and Step 4. The term spaces are evaluated for their ability to identify synonyms of SNOMED CT preferred terms that each appears at least fifty times in the corpus (Table 1). A vast number of

SNOMED CT terms do not appear in the corpus; requiring that they appear a certain number of times arguably makes the evaluation more realistic. Although fifty is a somewhat arbitrarily chosen number, it is likely to ensure that the statistical foundation is solid. A preferred term is provided as input to a term space and the twenty most semantically similar terms are output, provided that they also appear at least fifty times in the data. For each preferred term, recall is calculated using the twenty most semantically similar terms generated by the model (i.e for each SNOMED CT concept, recall is the proportion of SNOMED CT synonyms returned for that concept when the model is queried using that concept's preferred term). The SNOMED CT data is divided into a *development set* and an *evaluation set* in a 50/50 split. The development set is used in Step 3 to find the optimal parameter settings for the respective datasets and the task at hand. The best parameter configuration for each type of model (*UNIGRAM TERMS* and *MULTIWORD TERMS*) is then used in the final evaluation in Step 4. Note that the requirement of each synonym pair appearing at least fifty times in the data means that the development and evaluation sets for the two types of models are not identical, e.g. the test sets for *UNIGRAM TERMS* will not contain any multiword terms. This, in turn, means that the results of the two types of models are not directly comparable.

	UNIGRAM TERMS		**MULTIWORD TERMS**	
Semantic Type	**Preferred Terms**	**Synonyms**	**Preferred Terms**	**Synonyms**
attribute	4	6	6	8
body structure	2	2	12	12
cell	0	0	1	1
cell structure	1	1	1	1
disorder	26	32	68	81
environment	3	3	3	3
event	1	1	1	1
finding	39	54	69	86
morphologic abnormality	35	45	42	54
observable entity	17	22	22	26
organism	2	2	3	3
person	2	2	2	2
physical force	1	1	1	1
physical object	9	10	12	15
procedure	23	28	49	64
product	6	6	3	3
qualifier value	133	173	153	190
regime/therapy	0	0	3	3
situation	0	0	1	1
specimen	0	0	2	0
substance	24	24	24	25
Total	328	412	478	580

Table 1: The frequency of SNOMED CT preferred terms and synonyms that are identified at least fifty times in the MIMIC II Corpus. The *UNIGRAM TERMS* set contains only unigram terms, while the MULTIWORD TERMS set contains unigram, bigram and trigram terms.

Results

One interesting result to report, although not the focus of this paper, concerns the coverage of SNOMED CT in a large clinical corpus like MIMIC-II. First of all, it is interesting to note that only 9,267 out of the 105,437 preferred terms with one or more synonyms are unigram terms (24,866 bigram terms, 21,045 trigram terms and 50,259 terms that consist of more than three words/tokens). Out of the 158,919 synonyms, 12,407 are unigram terms (43,513 bigram terms, 32,367 trigram terms and 70,632 terms that consist of more than three words/tokens). 7,265 SNOMED CT terms (preferred terms and synonyms) are identified in the MIMIC-II corpus (2,632 unigram terms, 3,217 bigram

terms and 1,416 trigram terms); the occurrence of longer terms in the corpus has not been verified in the current work. For the number of preferred terms and synonyms that appear more than fifty times in the corpus, consult Table 1.

When tuning the parameters of the *UNIGRAM TERMS* models and the *MULTIWORD TERMS* models, the pattern is fairly clear: for both dataset variants, the best model parameter settings are based on random permutation and a dimensionality of 1,500. For *UNIGRAM TERMS*, a sliding window of 5+5 yields the best results (Table 2). The general tendency is that results improve as the dimensionality and the size of the sliding window increase. However, increasing the size of the context window beyond 5+5 surrounding terms does not boosts results further.

	RANDOM INDEXING						**RANDOM PERMUTATION**					
Sliding Window →	**1+1**	**2+2**	**3+3**	**4+4**	**5+5**	**6+6**	**1+1**	**2+2**	**3+3**	**4+4**	**5+5**	**6+6**
500 *Dimensions*	0.17	0.18	0.20	0.20	0.20	0.21	0.17	0.20	0.21	0.21	0.22	0.22
1,000 *Dimensions*	0.16	0.19	0.20	0.21	0.21	0.21	0.19	0.22	0.21	0.21	0.22	0.23
1,500 *Dimensions*	0.17	0.21	0.22	0.22	0.22	0.22	0.18	0.22	0.22	0.23	**0.24**	0.23

Table 2: Model parameter tuning: results, recall top 20, for *UNIGRAM TERMS* on the development set.

For *MULTIWORD TERMS*, a sliding window of 4+4 yields the best results (Table 3). The tendency is similar to that of the *UNIGRAM TERMS* models: incorporating term order (RP) and employing a larger dimensionality leads to the best performance. In contrast to the *UNIGRAM TERMS* models, the most optimal context window size is in this case slightly smaller.

	RANDOM INDEXING						**RANDOM PERMUTATION**					
Sliding Window →	**1+1**	**2+2**	**3+3**	**4+4**	**5+5**	**6+6**	**1+1**	**2+2**	**3+3**	**4+4**	**5+5**	**6+6**
500 *Dimensions*	0.08	0.12	0.13	0.13	0.13	0.12	0.08	0.13	0.14	0.14	0.13	0.12
1,000 *Dimensions*	0.09	0.12	0.13	0.13	0.13	0.13	0.10	0.13	0.14	0.13	0.12	0.11
1,500 *Dimensions*	0.09	0.13	0.13	0.14	0.14	0.14	0.10	0.14	0.14	**0.16**	0.14	0.14

Table 3: Model parameter tuning: results, recall top 20, for MULTIWORD TERMS on the development set.

Once the optimal parameter settings for each dataset variant had been configured, they were evaluated for their ability to identify synonyms on the unseen evaluation set. Overall, the best *UNIGRAM TERMS* model (RP, 5+5 context window, 1,500 dimensions) achieved a recall top 20 of 0.24 (Table 4), i.e. 24% of all unigram synonym pairs that occur at least fifty times in the corpus were successfully identified in a list of twenty suggestions per preferred term. For certain semantic types, such as *morphologic abnormality* and *procedure*, the results are slightly higher: almost 50%. For *finding*, the results are slightly lower: 15%.

With the best *MULTIWORD TERMS* model (RP, 4+4 context window, 1,500 dimensions), the average recall top 20 is 0.16. Again, the results vary depending on the semantic type: higher results are achieved for entities such as *disorder* (0.22) *morphologic abnormality* (0.29) and *physical object* (0.42), while lower results are obtained for *finding* (0.12), *observable entity* (0.09), *qualifier value* (0.08) and *substance* (0.08). For 22 of the correctly identified synonym pairs, at least one of the terms in the synonymous relation was a multiword term.

Discussion

When modeling unigram terms, as is the traditional approach when employing models of distributional semantics, the results are fairly good: almost 25% of all synonym pairs that appear with some regularity in the corpus are successfully identified. However, the problem with the traditional unigram-based approach is that the vast majority of SNOMED CT terms – and other biomedical terms for that matter – are multiword terms: fewer than 10% are unigram terms. This highlights the importance of developing methods and techniques that are able to model the meaning of multiword expressions. In this paper, we attempted to do this in a fairly straightforward manner: identify multiword terms and

Semantic Type	# Synonym Pairs	Recall (Top 20)
attribute	2	0.50
body structure	1	0.00
cell structure	1	1.00
disorder	17	0.23
environment	2	0.00
event	1	0.00
finding	27	0.15
morphologic abnormality	26	0.43
observable entity	11	0.22
organism	1	0.00
person	1	1.00
physical force	1	0.00
physical object	6	0.30
procedure	15	0.46
product	2	0.50
qualifier value	89	0.15
substance	12	0.33
All	215	0.24

Table 4: Final evaluation: results, recall top 20, for *UNIGRAM TERMS* on the evaluation set.

treat each one as a distinct semantic unit. This approach allowed us to identify more synonym pairs in the corpus (292 vs. 215). The results were slightly lower compared to the *UNIGRAM TERMS* model, although the results are not directly comparable since they were evaluated on different datasets. The *UNIGRAM TERMS* model was unable to identify any synonymous relations involving a multiword term, whereas the *MULTIWORD TERMS* model successfully identified 22 such relations. This demonstrates that multiword terms can be handled with some amount of success in distributional semantic models. However, our approach relies to a large degree on the ability to identify high quality multiword terms, which was not the focus of this paper. The term extraction could be improved substantially by using a linguistic filter that produces better candidate terms than n-grams. Using a shallow parser to extract phrases is one such obvious improvement.

Another issue concerns the evaluation of the method. Relying heavily on a purely quantitative evaluation, as we have done, can provide only a limited view of the usefulness of the models. Only counting the number of synonyms that are currently in SNOMED CT – and treating this as our gold standard – does not say anything about the quality of the "incorrect" suggestions. There may be valid synonyms that are currently not in SNOMED CT. One example of this is the preferred term *itching*, which, in SNOMED CT, has two synonyms: *itch* and *itchy*. The model was able to identify the former but not the latter; however, it also identified *itchiness*. Another phenomenon which is perhaps of less interest in the case of SNOMED CT, but of huge significance for developing terminologies that are to be used for information extraction purposes: the identification of misspellings. When looking up *anxiety*, for instance, the synonym *anxiousness* was successfully identified; other related terms were *agitation* and *aggitation* [sic]. Many of the suggested terms are variants of a limited number of concepts. Future work should thus involve review of candidate synonyms by human evaluators.

Semantic Type	# Synonym Pairs	Recall (Top 20)
attribute	3	0.33
body structure	6	0.17
cell structure	1	1.00
cell	1	0.00
disorder	40	0.22
environment	2	0.00
event	1	0.00
finding	43	0.12
morphologic abnormality	29	0.29
observable entity	15	0.09
organism	2	0.00
person	1	1.00
physical force	1	0.00
physical object	7	0.42
procedure	34	0.17
product	2	0.50
qualifier value	88	0.08
regime/therapy	2	0.50
situation	1	0.00
specimen	1	0.00
substance	12	0.08
All	292	0.16

Table 5: Final evaluation: results, recall top 20, for MULTIWORD TERMS on the evaluation set.

Conclusions

We have demonstrated how distributional analysis of a large corpus of clinical narratives can be used to identify synonymy between SNOMED CT terms. In addition to capturing synonymous relations between pairs of unigram terms, we have shown that we are also able to extract such relations between terms of varying length. This language independent method can be used to port SNOMED CT – and other terminologies and ontologies – to other languages.

Acknowledgements

This work was partially supported by NIH grant R01LM009427 (authors WC & MC), the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection, ref. no. IIS11-0053 (author AH) and the Stockholm University Academic Initiative through the Interlock project.

References

1. International Health Terminology Standards Development Organisation: SNOMED CT;. Available from: `http://www.ihtsdo.org/snomed-ct/` [cited 10th March 2013].

2. International Health Terminology Standards Development Organisation: Supporting Different Languages;. Available from: `http://www.ihtsdo.org/snomed-ct/snomed-ct0/different-languages/` [cited 10th March 2013].

3. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, et al. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database. Crit Care Med. 2011;39(5):952–960.

4. SNOMED CT User Guide 2013 International Release:;. Available from: http://www.webcitation.org/6IlJeb5Uj [cited 10th March 2013].

5. Zeng QT, Redd D, Rindflesch T, Nebeker J. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. AMIA Annu Symp Proc. 2012;2012:1050–9.

6. Unified Medical Language System:;. Available from: http://www.nlm.nih.gov/research/umls/ [cited 10th March 2013].

7. Cohen T, Widdows D, Schvaneveldt RW, Davies P, Rindflesch TC. Discovering discovery patterns with predication-based Semantic Indexing. J Biomed Inform. 2012 Dec;45(6):1049–65.

8. Conway M, Chapman W. Discovering lexical instantiations of clinical concepts using web services, WordNet and corpus resources. In: AMIA Fall Symposium; 2012. p. 1604.

9. Hearst M. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of COLING 1992; 1992. p. 539–545.

10. Firth JR, Palmer FR. Selected papers of J. R. Firth, 1952-59. Indiana University studies in the history and theory of linguistics. Bloomington: Indiana University Press; 1968.

11. Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. J Biomed Inform. 2009 April;42(2):390–405.

12. Panchenko A. Similarity measures for semantic relation extraction. PhD thesis, Université catholique de Louvain & Bauman Moscow State Technical University; 2013.

13. Deerwester S, Dumais S, Furnas G. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990;41(6):391–407.

14. Kanerva P, Kristofersson J, Holst A. Random indexing of text samples for latent semantic analysis. In: Proceedings of 22nd Annual Conference of the Cognitive Science Society; 2000. p. 1036.

15. Sahlgren M. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. PhD thesis, Stockholm University; 2006.

16. Sahlgren M, Holst A, Kanerva P. Permutations as a Means to Encode Order in Word Space. In: Proceedings of the 30th Annual Meeting of the Cognitive Science Society; 2008. p. 1300–1305.

17. Henriksson A, Moen H, Skeppstedt M, Eklund AM, Daudaravicius V. Synonym extraction of medical terms from clinical text using combinations of word space models. In: Proceedings of Semantic Mining in Biomedicine (SMBM); 2012. p. 10–17.

18. Henriksson A, Hassel M. Optimizing the dimensionality of clinical term spaces for improved diagnosis coding support. In: Proceedings of the LOUHI Workshop on Health Document Text Mining and Information Analysis; 2013. p. 1–6.

19. Banerjee S, Pedersen T. The design, implementation, and use of the Ngram Statistic Package. Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing). 2003;p. 370–381.

20. Frantzi K, Ananiadou S. Extracting nested collocations. In: Proceedings of the Conference on Computational Linguistics (COLING); 1996. p. 41–46.

21. Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries. 2000;3(2):115–130.

22. Zhang Z, Iria J, Brewster C, Ciravegna F. A Comparative Evaluation of Term Recognition Algorithms. In: Proceedings of Language Resources and Evaluation (LREC); 2008. .