Bioinformatics Advance Access published August 16, 2016

Subject Section

Corpus Domain Effects on Distributional Semantic Modeling of Medical Terms

Serguei V.S. Pakhomov[1,*], Greg Finley[2], Reed McEwan[2], Yan Wang[2]

Genevieve B. Melton[2]

[1]College of Pharmacy, University of Minnesota, 308 Harvard St. SE, Minneapolis, MN 55455., [2]Institute for Health Informatics, University of Minnesota, 505 Essex Street SE, Minneapolis, MN 55455.

*To whom correspondence should be addressed.

Associate Editor: Dr. Jonathan Wren

Abstract

Motivation: Automatically quantifying semantic similarity and relatedness between clinical terms is an important aspect of text mining from electronic health records, which are increasingly recognized as valuable sources of phenotypic information for clinical genomics and bioinformatics research. A key obstacle to development of semantic relatedness measures is the limited availability of large quantities of clinical text to researchers and developers outside of major medical centers. Text from general English and biomedical literature are freely available; however, their validity as a substitute for clinical domain to represent semantics of clinical terms remains to be demonstrated.

Results: We constructed neural network representations of clinical terms found in a publicly available benchmark dataset manually labeled for semantic similarity and relatedness. Similarity and relatedness measures computed from text corpora in three domains (Clinical Notes, PubMed Central articles, and Wikipedia) were compared using the benchmark as reference. We found that measures computed from full text of biomedical articles in PubMed Central repository (rho = 0.62 for similarity and 0.58 for relatedness) are on par with measures computed from clinical reports (rho = 0.60 for similarity and 0.57 for relatedness). We also evaluated the use of neural network based relatedness measures for query expansion in a clinical document retrieval task and a biomedical term word sense disambiguation task. We found that, with some limitations, biomedical articles may be used in lieu of clinical reports to represent the semantics of clinical terms and that distributional semantic methods are useful for clinical and biomedical natural language processing applications.

Contact: *pakh0002@umn.edu, gpfinley@umn.edu, rmcewan@umn.edu, wang2258@umn.edu, gmelton@umn.edu.*

1 Introduction

Automated approaches for representing the semantic content of terms and similarity and relatedness between them have been widely used in a number of Natural Language Processing (NLP) applications in both general English (Budanitsky & Hirst, 2006; Landauer, 2006; Resnik, 1999; Weeds & Weir, 2005) and specialized terminological domains such as bioinformatics (Ferreira, Hastings, & Couto, 2013; Lord, Stevens, Brass, & Goble, 2003; Mazandu, Chimusa, Mbiyavanga, & Mulder, 2016; Wang, Du, Payattakool, Yu, & Chen, 2007; Yang, Nepusz, & Paccanaro, 2012) and medicine (Garla & Brandt, 2012; Lee, Shah, Sundlass, & Musen, 2008; Liu, McInnes, Pedersen, Melton-Meaux, & Pakhomov, 2012; Pakhomov et al., 2010; Pedersen, Pakhomov, Patwardhan, & Chute, 2007; Sajadi, 2014). A subset of these methods, distributional semantics, relies on the co-occurrence information between words obtained from large corpora of text and makes the assumption that words with similar or related meanings tend to occur in similar contexts. This approach is foundational to a number of higher-level tasks including information retrieval, word sense ambiguity resolution, automatic synonym generation and recognition, literature-based knowledge discovery, among many other (see Cohen and Widdows (2009) for a comprehensive review).

In the general English domain, distributional semantic approaches to measuring semantic similarity and relatedness have been quite successful, achieving correlations in the 70's and 80's with human judgments (Faruqui & Dyer, 2014). In the biomedical domain, the problem of representing lexical semantics of medical terms in such a way as to match human judgments of semantic similarity and relatedness between them has proven to be more challenging. State of the art approaches developed so far for computing semantic similarity and relatedness achieve only modest agreement with human judgments. This applies to both distributional methods (i.e., knowledge-free) and those based on relations in manually constructed ontologies such as WordNet and the Unified Medical Language System (i.e., knowledge-based). For example, Garla and Brant (2012) reported on a large systematic investigation of a wide range of knowledge-based and knowledge-free approaches to computing semantic relatedness and similarity and found that knowledge-based approaches augmented with information content obtained from corpora of text (e.g., Leacock and Chodorow (1998)) outperformed distributional approaches based on first and second order semantic vectors (e.g., Lin (1998) and Patwardhan and Pedersen (2006)). The authors used a number of publicly available benchmarks including the University of Minnesota Semantic Relatedness Standard (Pakhomov et al., 2010) containing the largest number and diversity of medical term pairs to date. In terms of agreement with human ratings and the UMNSRS benchmark, the best correlations range between 0.30 and 0.46 (Spearman rank correlation). Liu et al. (Liu et al., 2012) reported similar correlations between the output of the second-order context vectors (Patwardhan & Pedersen, 2006) and a subset of the UMNSRS benchmark.

Another study by Sajadi et al. (2014) that also included the UMNSRS benchmark, among others, confirmed Garla and Brant's (2012) findings demonstrating that their knowledge-based method based on the Wikipedia network (HITS-sim) achieved significantly higher correlations with human ratings than distributional methods. The Spearman rank correlations for the UMNSRS benchmark were 0.51 and 0.58 for semantic relatedness and semantic similarity judgments, respectively. The distributional semantic methods evaluated in Sajadi et al. (2014) included *word2vec*, a neural network based mechanism for semantic representation based on word embeddings that was originally proposed by Mikolov et al. (2013). Sajadi et al. (2014) trained a skip-gram vector representation of medical terms using word2vec on the OHSUMED corpus (a collection of 348,566 biomedical research articles). This approach achieved a correlation of 0.39 on both relatedness and similarity human judgments with the UMNSRS benchmark. A similar study by Muneeb et al. (2015) also examined *word2vec* semantic representations derived from the PubMed Central Open Access (PMC) corpus. Similarly to the study by Sajadi et al. (2014), Muneeb et al. (2015) tested their approach on the UMNSRS benchmark and reported Spearman rank correlations of 0.45 and 0.52 for semantic relatedness and semantic similarity judgments, respectively. These previous studies targeted overall performance of distributional semantic and other approaches and did not examine the performance on subsets of the UMNSRS dataset consisting of pairs of different semantic types.

We hypothesize that one of the reasons the UMNSRS benchmark has been difficult to approximate with automated approaches (particularly with knowledge-free distributional approaches) is because it consists of a large number of clinical concepts from a variety of semantic types (disorders, symptoms, and drugs). Thus, in order to model human judgments of semantic relatedness and similarity of this benchmark using distributional methods one may need to use very large amounts of textual data from a corpus that closely matches the domain of clinical language represented by this benchmark. Previous studies tended to rely on relatively small corpora that are from a closely related but not perfectly matched domains (e.g., biomedical articles). One exception to this was a study by Pedersen et al. (2007) that used Mayo Clinic clinical notes as a source of training data; however, while that study had a good domain match, the size of the corpus was relatively modest by current standards (~232M tokens).

The objective of the current study is to examine the effect of corpus size and sublanguage domain match on the agreement between relatedness and similarity measures produced by the popular and highly efficient distributional semantic method (*word2vec*) and human judgments in the UMNSRS benchmark dataset.

2 Methods

2.1 Reference standard

In this study, we used a previously developed public reference standard – University of Minnesota Semantic Relatedness Standard (UMNSRS). This reference standard consists of over 550 pairs of medical terms mapped to concept unique identifiers in the Unified Medical Language System (UMLS). By design, only single word terms had been selected for this reference standard to represent SYMPTOMS and DISORDERS; however, the selection of DRUGS was not restricted to single word terms and thus the reference contains several multi-word terms (e.g, "Vitamin K", "folic acid", "fish oil", etc.).

Each pair of terms was assessed by 4 medical residents and scored with respect to the degree to which the terms were similar or related to each other using a continuous scale. Thus UMNSRS consists of two subsets - one contains averaged semantic similarity ratings by 4 medical residents (UMNSRS-Similarity) and the other contains semantic relatedness ratings by a different set of 4 medical residents (UMNSRS-Relatedness). UMNSRS-Similarity contains a total of 566 pairs of terms and UMNSRS-Relatedness contains 587 pairs. Each of these datasets was designed to test for similarity and relatedness between medical concepts of several semantic types. Thus, UMNSRS contains 6 semantic categories: DISORDER-DISORDER, SYMPTOM-SYMPTOM, DRUG-DRUG, DISORDER-DRUG, DISORDER-SYMPTOM, and SYMPTOM-DRUG pairs. Further details on this reference standard are reported elsewhere (Pakhomov et al., 2010).

2.2 Text corpora

We focused on three sublanguage domains – clinical, biomedical, and general English. The clinical domain was represented by a corpus of clinical notes (CLINICAL) from the Fairview Health System documenting of 5 years worth of clinical encounters between 2010 and 2014, inclusively. Since Fairview Health System uses Epic Electronic Health Records (EHR) system, we used its secondary data use tool called Clarity to extract the text of clinical reports from the Epic database. We constructed several sets of smaller corpora from this dataset in order to test the effect of corpus size on computation of semantic relatedness and similarity. To this end, we split the clinical notes repository by year and created four additional corpora that spanned the following years: 2014, 2013-4, 2012-4, and 2011-4. The biomedical domain was represented by the PMC dataset consisting of full-length biomedical articles (available as of September 2015). The general English domain was represented by all WIKIPEDIA articles (available as of September 2015). All text data were pre-processed minimally by removing non-alphanumeric characters and lowercasing the rest. No further alterations including morphological normalization or stop word removal were done.

2.3 Modifications to the reference standard

For the present study, we modified the UMNSRS dataset in order to account for the differences in vocabulary in the corpora that were used to obtain the contexts for medical terms. Generally, the number of terms from UMNSRS that we could not find in the three corpora was low. For example, the list of terms from UMNSRS that were not found in the CLINICAL-ALL corpus is provided in the Appendix A.

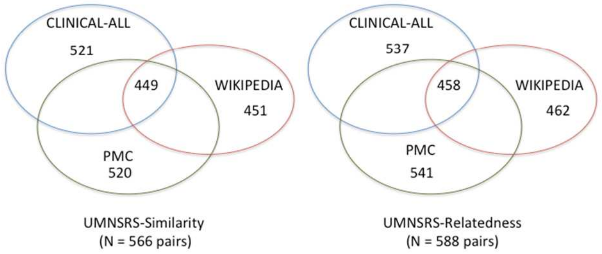

Figure 1. Pairs of terms remaining in UMNSRS-Similarity and UMNSRS-Relatedness datasets after reduction to match the vocabulary in the three domain corpora.

We had to discard 40 pairs of words from the UMN-Similarity and 43 pairs from the UMNSRS-Relatedness subsets based on the CLINICAL corpus. These numbers were similar for PMC – 39 pairs discarded based on UMNSRS-Similarity and 40 based on UMNSRS-Relatedness. Substantially more pairs had to be discarded with Wikipedia – 108 based on UMNSRS-Similarity and 119 based on UMNSRS-Relatedness. We also had to modify UMNSRS in order to account for some of the term pairs having been presented to human raters more than once. These repeated pairs were included in the reference standard for estimating internal rater consistency; however, we wanted to exclude them from calculating correlations with automated measures as they are not independent and can potentially artificially affect correlations. As illustrated in Figure 1, combining exclusions of word pairs from the UMNSRS-Similarity and UMNSRS-Relatedness datasets based on all three corpora resulted in the overall reduction from 566 to 449 pairs in the UMNSRS-Similarity and from 588 to 458 pairs in the UMNSRS-Relatedness datasets[1].

2.4 Computation of semantic relatedness and testing

We used *word2vec* toolkit developed by Mikolov et al. (2013) to generate word embeddings for each word in the three corpora. Default parameters were used. We trained a bag-of-words (CBOW) model for word embeddings with a window size set to 8 words. The dimensionality of vectors representing word embeddings was set to 200. Prior work comparing CBOW to the skip-gram representations is inconclusive and suggests that superiority of the representations may be task-dependent. While Mikolov et al. (2013) and Levy et al. (2015) show that skip-gram representations are superior to CBOW, Baroni et al. (2014) show the opposite, and Muneeb et al. (2015) show that the skip-gram representation outperforms CBOW on semantic similarity tasks but the two are equivalent on semantic relatedness tasks. Due to lack of clear evidence that one representation is superior to the other and since the objective of the current study is not to optimize word2vec performance, we arbitrarily decided to use the CBOW representation as the basis for all experiments.

During testing, we extracted word vectors for each term pair in the reference standard and computed the cosine between them in a standard fashion using their dot product and magnitude. The cosines computed this way for each term pair in the reference standard were then compared to manual similarity and relatedness judgments using Spearman rank correlation in each dataset.

2.5 Evaluation on a document retrieval task

In order to evaluate the *word2vec* based semantic relatedness measures on a task of direct relevance to clinical NLP and clinical research, we defined a document retrieval task in which we searched a collection of EHR clinical notes for all clinical encounters between January 1, 2015 and October 1, 2015 in the Fairview Health Services for patients diagnosed with heart failure. This particular task was motivated by prior research showing advantages of using unstructured text of EHR to identify candidate patients for participation in a cohort research study of heart failure in the community (Bursi et al., 2006). The advantages included better concurrency and completeness of case ascertainment for potential inclusion into the study as compared to using diagnostic codes.

As part of evaluation for the current study, we defined a similar document retrieval task in which we indexed all EHR records from the Fairview Health Services using Elasticsearch technology with the default Snowball analyzer. The index was then queried using the Patient Information Extraction for Research (PIER) system (McEwan et al., 2016) with a disjunction of the following terms: "heart failure" OR "HF" OR "CHF." The results of this query were parsed to extract the set of unique patient IDs that formed the baseline for subsequent comparisons.

The search query was then augmented with additional search terms obtained from a *word2vec* model that was constructed from two data sources: CLINICAL-ALL and PMC corpora. The WIKIPEDIA corpus was excluded from further evaluations based on the results of cross-domain comparisons on the UMNSRS reference standard. Since this subsequent validation task was not constrained by the UMNSRS reference standard having only single word terms, we constructed these models by using multi-word expressions (up to 4 word sequences). This was achieved by taking two pre-processing passes over the corpora with the *word2phrase* tool to construct a bigrams-of-bigrams corpus (the bigram threshold was set to 200 on the first pass and 100 on the second pass),

[1] Available at http://rxinformatics.umn.edu/SemanticRelatednessResources.html

followed by training a model with *word2vec* using the same default parameters as described in the previous section.

We took the top 5, 10, 20 and 40 most semantically closely related words to the term "heart failure" as defined by the cosine distance between their semantic vectors computed with *word2vec*. An expanded query was then defined that included the search terms used in the baseline query and the top 5, 10, 20 or 40 semantically related terms, thus forming four points of comparison with the baseline.

In order to control for both the baseline and semantically expanded queries returning results not relevant to heart failure diagnosis due to language use phenomena such as negation, uncertainty and family history contexts, we also ran a query on the structured part of the EHR for ICD-9 billing code 428.x (heart failure). The choice of the time period between January 1, 2015 and October 1, 2015 was dictated by the fact that Fairview Health Services transitioned to using ICD-10 in October of 2015. In order to avoid potential issues with cross-classification and translation between ICD-9 and ICD-10 in this transitional period, we limited all queries of both structured and unstructured data to before October 1, 2015.

The use of ICD-9 code 428.x as a primary or secondary diagnosis has been shown in multiple studies to have high precision (reported positive predictive value ~ 87%) but relatively low recall (reported sensitivity ~ 60%) for identification of patients with heart failure (see (McCormick, Lacaille, Bhole, & Avina-Zubieta, 2014) for review). Thus, we focused on measuring the value added by the semantically expanded queries in terms of improvements in recall - additional patients that this query found as compared to baseline. However, while querying the text of EHR to identify patients with heart failure has been shown to have high sensitivity (82% recall), it also had low precision (49% positive predictive value) (Pakhomov et al., 2007). Therefore, for the current validation study, we focused on comparing the recall of both the baseline and semantically expanded queries of patients that also had an ICD-9 code of 428.x in their EHR. The disadvantage of restricting the evaluation of recall to patients with code 428 is that we will not be able to estimate how many more valid patients *both* the baseline and semantically expanded queries can identify. However, the advantage of this approach is the certainty afforded by the high precision of the reference standard, which is more important for comparisons of recall between the expanded and baseline queries *relative* to each other.

2.6 Evaluation on a word sense disambiguation task

We also investigated the use of word embeddings as semantic representations for word sense disambiguation (WSD) in a biomedical NLP context. WSD methods that rely on distributional semantics involve automatically choosing the most context-appropriate sense of a word by maximizing the similarity of the sense and the context. With these approaches, a basic representation of the semantic content of the senses and contexts takes the form of semantic vectors containing co-occurrence counts. For example, McInnes et al. (2011) introduce a modification of the Lesk algorithm for performing WSD using machine-readable dictionaries. They assign each sense an expanded definition by concatenating the definitions of the sense and of related concepts in the UMLS. They represent senses as second-order co-occurrence vectors by calculating the centroid of word co-occurrence vectors for each word in the extended definition. For each test example, a context vector is calculated similarly, by averaging co-occurrence vectors of the words within a fixed window size (six words) around the ambiguous term, and the sense vector with highest cosine similarity to the context vector is chosen.

In the current study we compared this second-order co-occurrence vector based approach to an alternative approach in which co-occurrence vectors are replaced with word embeddings. Co-occurrence vectors were derived separately from the CLINICAL-ALL and PMC corpora. Word embeddings were also generated for each corpus as described in section 2.4.

We implemented these two approaches to word sense representation in a knowledge-based WSD system similar to that employed by McInnes et al. (2011). The main differences from the latter consisted of smoothing in the form of taking a square root of all co-occurrence counts in order to emphasize rare words, and reducing the weight of the words occurring in the extended definitions by 50% to give more weight to the words in the primary definitions of concepts. The two approaches to word sense representation implemented in our WSD system were evaluated on 203 biomedical terms and acronyms contained in the publicly available National Library of Medicine MSH-WSD dataset (see Jimeno-Yepes, McInnes, & Aronson, 2011 for a detailed description of this dataset). For the current study, we split this dataset 50/50 into a development and validation subsets using stratified random sampling. The parameters of both approaches to word sense representation were tuned on the development set and evaluated on the validation set.

3 Results

3.1 Training corpora and reference standards

The general characteristics of corpora used to train neural representations of individual single-word terms are shown in Table 1.

Table 1. Descriptive statistics for text corpora

Corpus	Vocab. size (N word types)	Corpus size (N word tokens)	Dates
CLINICAL-ALL	401,087	4,169,696,714	2010-4
CLINICAL-4B	385,540	3,882,516,140	2011-4
CLINICAL-3B	342,074	3,137,442,621	2012-4
CLINICAL-2B	262,250	1,853,601,667	2013-4
CLINICAL-500M	132,739	452,314,196	2014 (Jan-Aug)
CLINICAL-100M	73,147	123,969,678	Jan 2014
CLINICAL-10M	26,618	9,931,684	Jan 2014 first 10 M wrds
CLINICAL-1M	9,616	973,143	Jan 2014 first 1 M wrds
PMC	1,596,146	4,533,613,517	Up to Sept 2015
WIKIPEDIA	1,350,678	1,700,369,832	Up to Sept 2015

Since we are comparing three heterogeneous sources of training data with varying vocabulary coverage on a common reference standard, we had to select pairs of terms from the reference standard where both terms were in the vocabulary of each of the three corpora. This resulted in reducing the size of the reference standard. Introducing systematic bias is one potential consequence of reducing the size of the reference standard in this manner. To test for bias, we compared the agreement of the human raters (measured as the standard deviation in response included in the UMNSRS dataset) on the pairs that happened to be excluded from the overlap between the three sources of data, and thus from the comparisons, to the agreement on pairs that were included in the overlap. For the UMNSRS-Similarity dataset, the average standard deviations in human raters response were 274.2 and 275.4 for the excluded (N=109) and included (N=449) pairs, respectively. This difference was not statistically significant (p=0.795). Likewise, for the UMNSRS-Relatedness dataset, the average standard deviations in human raters response were 284.4 and 287.7 for the excluded (N=170) and included (N=458) pairs, respectively. This difference was not statistically significant (p=0.795).

3.2 Comparisons across domains

The results of comparing the correlations between human and machine-generated similarity and relatedness judgments based on the three different domain corpora are shown in Tables 2 and 3. As evident from these tables, the overall results (obtained from all available pairs regardless of the semantic type of the terms in the pair) between the CLINICAL and PMC domains are almost identical and both perform considerably closer to human ratings than those obtained from the WIKIPEDIA domain. The differences between the CLINICAL and PMC domains begin to emerge when comparisons are made on subsets of the UNMSRS pairs selected based on the semantic types of the terms in the pair.

When examining semantic similarity between terms (Table 2), it is evident that vector representations of term semantics trained on the PMC domain reflects human judgments of similarity on disorder-disorder pairs (rho = 0.74) and disorder-symptom pairs (rho = 0.49) much closer than representations trained on the CLINICAL domain (rho = 0.61 and 0.42, respectively). However, for disorder-drug and symptom-drug pairs the opposite is true – representations trained on the CLINICAL domain are closer to human judgments than those trained on the PMC domain (rho 0.69 vs. 0.62 for disorder-drug and 0.51 vs. 0.40 for symptom-drug pairs).

Table 2. Comparison of correlations with human raters of automated cosine distance-based approaches trained on corpora from several domains and tested on UMNSRS-Similarity (Di – disorder, S – symptom, Dr – drug)

Corpus	Spearman rank correlation coefficient (rho)						
	All pairs (n=449)	Di-Di (n=84)	S-S (n=82)	Dr-Dr (n=57)	Di-S (n=78)	Di-Dr (n=77)	S-Dr (n=71)
CLINICAL-ALL	0.60	0.61	0.56	0.76	0.42	0.69	0.51
PMC	0.62	0.74	0.55	0.77	0.49	0.66	0.40
WIKIPEDIA	0.48	0.59	0.43	0.67	0.22	0.53	0.37

Table 3. Comparison of correlations with human raters of automated cosine distance-based approaches trained on corpora from several domains and tested on UMNSRS-Relatedness (Di – disorder, S – symptom, Dr – drug)

Corpus	Spearman rank correlation coefficient (rho)						
	All pairs (n=458)	Di-Di[#] (n=88)	S-S (n=83)	Dr-Dr (n=65)	Di-S (n=85)	Di-Dr (n=70)	S-Dr (n=67)
CLINICAL-ALL	0.57	0.59	0.57	0.68	0.41	0.63	0.59
PMC	0.58	0.59	0.64	0.73	0.42	0.52	0.54
WIKIPEDIA	0.45	0.59	0.36	0.62	0.35	0.47	0.25*

* p-value less than 0.05; all other p-values are less than 0.01

Table 4. Correlations with human judgments as a function of corpus size in the CLINICAL domain.

	Corpus Size (word tokens) (x 10^9)	N pairs in UMNSRS-Sim in common with CLINCAL, PMC, and WIKIPEDIA	N pairs in UMNSRS-Rel in common with CLINCAL, PMC, and WIKIPEDIA	Correlation with UMNSRS-Sim ratings	Correlation with UMNSRS-Rel ratings
CLINICAL-1M	~ 0.001	188	187	0.46	0.43
CLINICAL-10M	~ 0.01	326	319	0.56	0.54
CLINICAL-100M	~ 0.12	403	405	0.60	0.57
CLINICAL-500M	~ 0.45	419	425	0.60	0.58
CLINICAL-2B	~ 1.85	441	452	0.63	0.59
CLINICAL-3B	~ 3.14	445	455	0.61	0.58
CLINICAL-4B	~ 3.88	449	458	0.60	0.56
CLINICAL-ALL	~ 4.16	449	458	0.61	0.57

The examination of semantic relatedness (Table 3) demonstrates that representations trained on the PMC domain are closer to human judgments in measuring the degree of relatedness between symptoms and drugs (rho 0.64 vs. 0.57 for symptom-symptom pairs and 0.73 vs. 0.68 for drug-drug pairs). As with similarity judgments, term representations trained on the CLINICAL domain are better at capturing relatedness between disorders and drugs and symptoms and drugs (rho 0.63 vs. 0.52

and 0.59 vs. 0.54, respectively). While overall semantic representations of terms trained on WIKIPEDIA are not as close to human judgments as the other two domains, these representations appear to be on par with PMC and CLINICAL in measuring relatedness between disorders (rho = 0.59 for all three domains). Representations based on WIKIPEDIA are weakest at measuring similarity and relatedness of disorder-symptom and symptom-drug pairs

3.3 Effects of corpus size

As shown in Table 1, the WIKIPEDIA corpus is substantially smaller in terms of the number of word tokens than either the CLINCAL-ALL or the PMC corpora. This potentially raises a question whether some of the lower correlations between the manual ratings and those produced automatically on the basis of neural representations of word semantics obtained with WIKIPEDIA are due to the size of the corpus rather than the mismatch between domains. Since we could not increase the size of the WIKIPEDIA corpus (all available text was already used), we decreased the size of one of the other two corpora. The CLINICAL-ALL corpus lends itself particularly well to this task as the data in this corpus are organized by year and month in which a clinical note was created. This temporal structure of the corpus also provides the opportunity to test for any recency effects that may bias models trained on more recent data to perform better than older data or vice versa. The results of comparisons across models trained on different size corpora are provided in Table 4 and show that any appreciable decreases in correlations with human ratings of similarity and relatedness begin to appear only when the corpus size drops significantly below 100M tokens. At a size of ~ 2B tokens (roughly equivalent to WIKIPEDIA's ~1.7B), the correlations with human judgments of similarity and relatedness are still much higher for the term representations trained on the CLINICAL domain than on WIKIPEDIA.

3.4 Document retrieval task

The sets of top 100 most semantically related terms to the term "heart failure" derived from CLINCIAL_ALL and PMC corpora are shown in Appendix B in the Supplement.

As evident from Appendix B, the sets of closely related terms derived from both corpora include not only synonymous or nearly synonymous more specific terms (e.g., "biventricular heart failure", "diastolic heart failure", "systolic heart failure"), but also non-synonymous and functionally related terms for conditions closely associated with heart failure (e.g., "dilated cardiomyopathy", "pulmonary hypertension", "volume overload", "chronic cor pulmonale", "reduced EF", "elevated BNP"). The sets of top 100 most closely semantically related terms to "heart failure" derived from CLINCIAL-ALL and PMC corpora contain 19 phrases that occur in both sets - highlighted in bold in Appendix B.

The results of the comparisons between the baseline query and semantically expanded queries with 5, 10, 20 and 40 most closely related terms for heart failure are summarized in Table 5. Only phrases from the semantically related sets that did not contain any of the terms used in the baseline query (i.e., "heart failure", "HF, "CHF") were used to expand the baseline query. The actual phrases that were added (via OR boolean operator) to the baseline query are marked in Appendix B in the columns labeled "Top, 5, 10, 20, 40."

The results in Table 5 show that queries expanded with semantically related phrases identify more patients with code 428.x than the baseline query. Both queries expanded with top 5 phrases from CLINICAL-ALL corpus and PMC corpus perform very similarly. Queries expanded with top 10, 20 and 40 PMC phrases perform slightly better with respect to recall; however, the CLINICAL-ALL queries yield higher precision and F1 scores.

Table 5. Comparison of results on the document retrieval task between two sources of phrases for query expansion.

	N patients identified (% of total study sample)	N patients with 428.x (increase over baseline)	Recall (95% CI)	Precision (95% CI)	F1
Study sample (Jan-Oct, 2015)	610282 (100%)	--	--	--	--
ICD-9 (428.x) query	7497 (1.2%)	7497	--	--	--
Baseline ("HF" or "CHF" or "heart failure") query					
	45728 (7.5%)	6716	0.896 (.889-.903)	0.147 (.144-.150)	0.253
Baseline + related phrases derived from CLINICAL-ALL corpus query					
Top 5	54992 (9.0%)	6817 (+101)	0.909 (.903-.916)	0.124 (.121-.126)	0.218
Top 10	56601 (9.3%)	6835 (+119)	0.912 (.905-.918)	0.121 (.118-.123)	0.214
Top 20	57516 (9.4%)	6842 (+126)	0.913 (.906-.919)	0.120 (.116-.122)	0.212
Top 40	213243 (35%)	7337 (+621)	0.979 (.975-.982)	0.034 (.034-.035)	0.066
Baseline + related phrases derived from PMC corpus query					
Top 5	53035 (8.7%)	6759 (+43)	0.902 (.895-.908)	0.127 (.125-.130)	0.223
Top 10	138827 (22.7%)	6928 (+212)	0.924 (.918-.930)	0.050 (.049-.051)	0.095
Top 20	196617 (32.2%)	7027 (+311)	0.937 (.932-.943)	0.036 (.035-.037)	0.069
Top 40	334524 (54.8%)	7129 (+413)	0.951 (.946-.956)	0.021 (.021-.022)	0.041

3.5 WSD task

The results of cross-domain comparisons of the approaches to semantic representation based on co-occurrence vectors and word embeddings are summarized in Table 6. The approach based on word embeddings outperforms co-occurrence vectors regardless of whether the training data came from the clinical (McNemar's $\chi^2 = 36.6$, $p < 0.001$) or biomedical ($\chi^2 = 128.2$, $p < 0.001$) domains. Both co-occurrence and word embeddings approaches perform better on this task when trained on the PMC corpus that matches the domain of the NLM WSD dataset (co-occurrence: $\chi^2 = 130.8$; embeddings: $\chi^2 = 281.2$).

Embedding-based sense representations also exhibited an interesting property: a strong negative correlation between the cosine similarity of the sense vectors for a single term and disambiguation accuracy that term (Pearson's r = -0.43, p < 0.001 for CLINICAL-ALL; r = -0.56, p < 0.001 for PMC). It would follow intuitively that more similar senses should be more difficult to disambiguate. However, this correlation was not as reliably observed for the second-order co-occurrence vectors (r = -0.06, p = 0.40 for CLINICAL-ALL; r = -0.15, p = 0.03 for PMC). Regardless of their suitability for WSD, co-occurrence representations appear not to directly model the closeness of easily confusable senses as well as trained embeddings.

4 Discussion

In the current study, we performed a large-scale evaluation of a popular neural network learning approach (*word2vec*) to representing the meaning of words in the medical domain and measuring the strength of association between them. Our results are overall slightly better than the best results reported so far on the UMNSRS benchmark by Sajadi et al. (2014) that used a graph-based approach (HITS-sim) to represent word semantics that leveraged Wikipedia as a network (rho 0.58 vs. 0.51 for relatedness and 0.62 vs. 0.58 for similarity). However, these differences are probably not significant from a practical standpoint due to reasons having to do with inter-rater agreement explained later in the discussion.

Table 6. WSD accuracy achieved on the NLM WSD corpus with different representations of word senses.

	CLINICAL-ALL	PMC
Co-occurrence vectors	0.700	0.740*
Word embeddings	0.722	0.777*
Majority sense	0.524	
Random chance	0.492	

* indicates significant difference from CLINICAL-ALL value at p < 0.001

For comparison, Sajadi et al. (2014) also used the *word2vec* approach (with skip-gram representation of word contexts) trained on OHSUMED corpus of 348,566 MEDLINE citations and the text of the Unified Medical Language System terms but obtained much lower correlations with human ratings (rho = 0.39). The citations included in the OHSUMED corpus are truncated at 250 words[2]; therefore, the total size of the OHSUMED collection is at most 87M tokens. The main differences of our study's use of word2vec from the one reported by Sajadi et al. (2014) are that we used the bag-of-words representation of contexts instead of the skip-gram representation, and a much larger corpus of text (over 4B tokens) for training consisting of entire articles rather than just the MEDLINE citations containing only the abstract and title of the articles. We believe that that the improved performance of *word2vec* approach observed in our study was mostly due to the latter two differences – larger corpus and inclusion of entire articles; however, we did not directly test these assumptions in the current study. At the outset of the current study, we intuitively expected to find domain and corpus size effects. We expected that, since the UMNSRS reference standard was created based on ratings by clinicians, semantic representations derived from the CLINICAL domain would agree better with clinicians' ratings than representations derived from PMC or WIKIPEDIA. The findings of this study confirm that, as expected, semantic representations derived from the general English domain (WIKIPEDIA) overall agree to a lesser extent with clinician's ratings than the other two domains. However, we did not expect to find that representations derived from biomedical literature would perform similarly or better to representations derived from clinical notes. This is an important finding because it suggests that distributed semantic representations derived from full texts of biomedical articles are equivalent (at least for the purpose to finding semantically similar or related terms) to semantic representations derived from clinical corpora. This is important because access to clinical corpora is highly restricted and difficult due to patient confidentiality and security of protected health information concerns. The open-access biomedical articles from PMC have no such restrictions and are freely and easily accessible by anyone. The only exceptions are disorder-drug and symptom-drug pairs for which representations derived from the CLINICAL domain were closer to clinician's ratings of similarity and relatedness than those derived from PMC. These exceptions as well as the overall findings, however, need to be examined in light of the inter-rater agreement on the pairs of terms include in UMNSRS benchmark.

As reported in our previous work (Pakhomov et al, 2010), the inter-rater agreement on all pairs of terms in the UMNSRS benchmark is in the moderate range (ICC 0.5 for relatedness and 0.47 for similarity) due to individual variability in conceptualizing the referents of medical terms and their similarity and relatedness to each other. The highest inter-rater agreement on the UMNSRS dataset was achieved on disorder-disorder pairs (ICC 0.56 for both similarity and relatedness) and similarity between symptoms and drugs (ICC 0.58 and 0.63, respectively). The lowest inter-rater agreement was observed on similarity for disorder-drug and symptom-drug pairs (ICC 0.33 and 0.24, respectively) – poor agreement. Thus, while it makes sense intuitively that in the current study we observe lower agreement on disorder-drug and symptom-drug between relatedness and similarity measures derived from the PMC domain than from the CLINCIAL domain, the observed Spearman rank correlations are substantially higher than the inter-rater agreement intraclass correlations for these categories of term pairs. Thus, we must exercise more caution in the interpretation of the results of the current study pertaining to disorder-drug and symptom-drug pairs. In the same vein, it should also be noted that the correlations in the 0.50 – 0.60 range reported in the current study and previous studies involving UMNSRS benchmark are in the same range as the intra-class correlation coefficients used to measure agreement and, therefore, may constitute ceiling performance that can be measured with the full UMNSRS benchmark.

In order to test whether our findings will hold at higher levels of inter-rater agreement, we used the median of the standard deviations in human rater responses to select pairs with higher agreement. By doing so, we selected subsets of 224 pairs from the UMNSRS-Similarity and 228 pairs from the UMNSRS-Relatedness benchmarks. For both of these selected subsets, the intra-class correlation coefficient among the 4 human raters per benchmark was 0.86 (excellent range). We then ran correlations between these selected manual ratings and automated measures of similarity and relatedness computed from the three domain corpora. As a result, we observed a substantial improvement in correlations (more than 5 points) with the UMNSRS-Similarity subset, but not UMNSRS-Relatedness. All correlations with the UMNSRS-Similarity subset improved approximately by the same amount across all three domains. On the UMNSRS-Relatedness subset, however, the correlations either stayed the same or changed very slightly (less than 5 points) in both

[2] http://trec.nist.gov/data/filtering/README.t9.filtering

negative and positive directions. The detailed comparisons are shown in Table 7.

Table 7. Changes in correlations with human ratings of similarity and relatedness after selecting pairs with high inter-rater agreement (ICC).

	UMNSRS-Similarity		UMNSRS-Relatedness	
	All pairs	**High agreement pairs**	**All pairs**	**High agreement pairs**
	n=449	n=224	n=458	n=228
ICC	0.47	0.86	0.50	0.86
CLINICAL-ALL	0.60	0.69 ↑	0.57	0.57 •
PMC	0.62	0.68 ↑	0.58	0.55 ↓
WIKIPEDIA	0.48	0.62 ↑	0.45	0.49 ↑

Another interesting finding of the current study is that increasing the size of the corpus beyond a certain size does not seem to provide an additional advantage. In the current study, the performance on both similarity and relatedness benchmarks plateaued between 10M and 100M tokens, which is consistent with the findings of another previous study by Pedersen et al. (2007) in which we found that the performance of another distributional semantics corpus-based approach to computing semantic relatedness plateaued after the training corpus reached 300,000 clinical notes (~ 66M tokens). These findings may be interpreted as providing additional evidence to show that the size of the corpus used for distributional semantic representations of medical terms does not matter beyond a certain point (e.g. 100M tokens); however, an alternative explanation is that the corpus size does matter but the plateauing of the correlations with human ratings is a function of the test data rather than the training data and has more to do with the inter-rater agreement. One possible way to test this hypothesis in future work is to apply corpus-based semantic relatedness measures in a secondary evaluation paradigm in which measures derived from corpora of different size would be used for another task such as word sense ambiguity resolution, spelling correction, or query expansion for information retrieval.

In addition to cross-domain comparisons of the *word2vec* based semantic relatedness measures on the UMNSRS reference standard, we also compared the use of semantically related phrases derived from clinical and biomedical domains on two tasks directly relevant to medical NLP. Automatically derived phrases for expanding text queries to identify patients with heart failure presented in the current study are consistent with the terms defined by experts in cardiovascular research that were used in prior studies examining the utility of NLP for identification of patients with heart failure from the unstructured text of EHR (Pakhomov, Buntrock, & Chute, 2005). These manually defined terms consisted of “cardiomyopathy,” “heart failure,” “congestive heart failure,” “pulmonary edema,” “decompensated heart failure,” “volume/fluid overload.” The sets of top 100 automatically derived terms related to “heart failure” from both the clinical (CLINCIAL-ALL) and biomedical (PMC) corpora in the current study contain all but one (“pulmonary edema”) of these terms manually determined to be good search terms for heart failure cases.

Our results also show that expanding text search queries with semantically related terms based on word embeddings can significantly improve recall of the queries. This is an important finding in the context of using NLP to identify potential candidates for clinical research studies such as clinical trials and cohort studies. Improved recall is particularly important for cohort studies in which the completeness of ascertaining cases with a condition of interest is critical to minimizing potential bias. Using structured billing codes (a.k.a. claims data) for case ascertainment has been shown to have limitations in terms of accuracy and completeness, particularly in community-based samples (Bazarian, Veazie, Mookerjee, & Lerner, 2006; Fan et al., 2013; Pakhomov et al., 2007). Furthermore, a number of conditions (e.g., symptoms and physical examination findings) may not have a diagnostic code entered as part of the medical record. Using NLP to search unstructured text of EHR can complement the use of billing codes for improved recall of potential candidates for prospective and retrospective studies. Improved recall is also relevant for clinical trials from a more practical standpoint. Being able to identify any number of additional potential candidates for a clinical trial can improve recruitment rates and consequently shorten the duration of the trial resulting in faster delivery of new therapeutic interventions to the bedside.

In the current study, we used the ICD-9 code 428.x to define the reference standard (with the understanding of the limitations inherent in this approach) for comparing query expansions to each other and to the baseline query. In order to minimize uncontrolled variability in these comparisons, we did not use advanced NLP tools (e.g., negation and family history detection) to make the text queries more precise. Therefore, the precision of the queries reported in this study does not reflect the actual precision that can be expected from text queries enhanced with NLP. In prior work, we showed that using negation detection can improve precision of text queries to approximately 50% (Pakhomov et al., 2007), and this number is likely to be higher for more sophisticated and customized NLP systems. In the current study, we focused on changes in precision between semantically expanded queries and the baseline. Our findings indicate that queries expanded with top 5 semantically related phrases from the CLINICAL-ALL and PMC corpora perform better than the baseline query and are comparable to each other. The performance of these queries diverges on larger sets of expansions (10, 20, and 40) with a large decrease in precision on queries derived from the PMC corpus. This drop in precision can be attributed to the terms “cardiac” and “hypertension” that are relatively high on the list of PMC derived phrases. Queries with these two terms would capture the majority of patients with heart disease (with or without heart failure) and thus negatively affect precision. However, despite the lower precision, these results indicate that the public PMC corpus can potentially be used as an alternative source of semantic relatedness information for expanding queries when searching clinical texts. This is particularly important in settings where the query expansion is not fully automated but is curated by the end user that can manually select more specific terms from the set of semantically related suggestions. While these findings are encouraging, clearly, further investigation is necessary to test their generalizability.

Similarly to the evaluation on the clinical document retrieval task, the evaluation of using word embeddings on a biomedical term WSD task showed that this type of representation improves WSD accuracy over basic co-occurrence based approaches. The accuracy of ~78% that we achieved on this dataset with word embeddings derived from the PMC corpus is comparable to other previously reported machine readable dictionary based results on the same dataset (Garla & Brandt, 2013; Jimeno-Yepes et al., 2011; B. T. McInnes & Pedersen, 2013). The improvement in WSD accuracy with word embeddings is likely due to the fact that even extended sense definitions are bound to be sparse, compared to occurrences in a text corpus of any reasonable size. Word embeddings may offer the advantage of smoothing over the gaps in representations based on sparse sense definitions. The fact that better WSD accuracy is achieved with methods trained on PMC data is not surprising because the NLM WSD dataset is derived from the same domain. What is interesting is that the accuracy achieved with word embeddings derived from the clinical domain is not that much lower than the accuracy achieved with in-domain training data, which constitutes additional evidence that the clinical and biomedical sources of text are similar with respect to semantic representations that can be derived from them.

5 Conclusions

Based on the results of the current study we can draw several conclusions with varying degrees of confidence. We are confident that easy-to-use and fast-to-compute *word2vec* neural representations of word embeddings are highly effective in capturing semantic relatedness and similarity relations between medical terms. We are moderately confident that equivalent, if not superior, results can be obtained with semantic representations derived from the freely available PMC biomedical corpus as from highly restricted clinical data. We are less confident at this time in concluding that there exists a threshold (e.g. 100M tokens) beyond which one can expect to stop seeing improvements in semantic representations of medical terms. Evaluating the use of semantic relatedness measures based on word embeddings in clinical and biomedical NLP applications showed that deriving word vectors from in-domain data offers a slight advantage over using text from a related but not the same domain.

This study has important implications for the development of NLP and information retrieval tools for clinical and translational research. Availability of usable semantic representations of clinical terms is important for improving concept-based information and document retrieval from unstructured text of electronic health records that, in turn, is critical for defining cohorts of patients for participation in experimental and observational clinical studies. Being able to construct distributional semantic representations of clinical terms from open-access biomedical literature that are equivalent to those obtained from difficult to access clinical corpora will enable wider adoption and more rapid further development and evaluation of these methods.

6 Limitations

The results of this study should be interpreted in light of several limitations. First, the UMNSRS benchmark, by design, contains mostly single word terms and has moderate inter-rater agreement. Thus, the results presented in this manuscript may not readily generalize to multi-word terms in other benchmarks and would require further investigation. Since the benchmark contains single word terms, we used the single word version of *word2vec*, as well; however, *word2vec* enables computation of embeddings at the phrase level that can also be investigated further in future work on clinical reference standards for semantic relatedenss. The moderate inter-rater agreement may introduce a ceiling for comparing various methods for computing semantic relatedness and similarity of around 0.5 - 0.6 correlation. The correlations reported in this study for the CLINICAL and PMC based measures are in that moderate range and thus it is possible that using a reference standard with better inter-rater agreement may show differences between using PMC and CLINICAL corpora. In the current study, we compared measures computed from these corpora on subsets of the UMNSRS benchmark with high agreement and did not find that measures computed from these corpora performed differently relative to each other than on the full UMNSRS benchmark. This finding provides additional evidence that similarity and relatedness measures derived from full text of biomedical articles are likely to be equivalent to those derived from clinical reports.

The second main limitation is that we did not experiment with various settings on *word2vec* such as context size and number of dimensions. Due to the number of experiments we ran in this study (combinations of corpora and benchmarks) and the size of the corpora, experimenting with parameter optimization would have been prohibitively time consuming. Thus, we relied on the default parameters. It is possible that larger context window sizes and/or larger number of dimensions to represent medical terms may yield further improvements; however, these improvements are likely to be incremental and would require a reference standard with greater inter-rater agreement to be reliably measured and compared across domains.

Another limitation that stems from the nature of the UMNSRS reference standard is that we only examined single word terms. Most medical terms consist of more that one word. The results obtained on single word terms may not generalize to multi-word terms and thus the latter need to be examined in future studies. However, in the current study, we were able to use neural representations of multi-word phrase semantics but only for evaluating a text query expansion application. Future studies will need to focus on constructing large representative collections of multi-word medical terms manually rated for semantic relatedness and similarity to enable direct evaluations of semantic representation methods.

Apart from the study's described limitations, it also has unique strengths. We used very large text corpora to derive semantic representation of terms. Most prior studies of distributional methods for computation of semantic relatedness in the medical domain relied on smaller corpora. The large corpus size also allowed us to experiment with various corpus sizes beyond those reported in prior literature. Another strength is that we used a benchmark that was used in a number of other previous studies, which enables direct comparisons between the results.

7 Acknowledgements

This work was supported by the National Library of Medicine (R01 LM009623-01).

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 238–247). Baltimore, Maryland.

Bazarian, J. J., Veazie, P., Mookerjee, S., & Lerner, E. B. (2006). Accuracy of Mild Traumatic Brain Injury Case Ascertainment Using ICD-9 Codes. *Academic Emergency Medicine*, *13*(1), 31–38.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Semantic Distance. *Computational Linguistics*, *32*, 13–47.

Bursi, F., Weston, S. A., Redfield, M. M., Jacobsen, S. J., Pakhomov, S., Nkomo, V. T., … Roger, V. L. (2006). Systolic and diastolic heart failure in the community. *JAMA*, *296*(18), 2209–2216.

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

Fan, J., Arruda-Olson, A. M., Leibson, C. L., Smith, C., Liu, G., Bailey, K. R., & Kullo, I. J. (2013). Billing code algorithms to identify cases of peripheral artery disease from administrative data. *Journal of the American Medical Informatics Association*, *20*(e2), e349–e354.

Faruqui, M., & Dyer, C. (2014). Community Evaluation and Exchange of Word Vectors at wordvectors.org. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (System Demonstration)* (p. 156).

Ferreira, J. D., Hastings, J., & Couto, F. M. (2013). Exploiting disjointness axioms to improve semantic similarity measures. *Bioinformatics*, *29*(21), 2781–2787.

Garla, V. N., & Brandt, C. (2012). Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC Bioinformatics*, *13*(1), 261.

Garla, V. N., & Brandt, C. (2013). Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. *Journal of the American Medical Informatics Association: JAMIA*, *20*(5), 882–886.

Jimeno-Yepes, A. J., McInnes, B. T., & Aronson, A. R. (2011). Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC Bioinformatics*, *12*(1), 223.

Landauer, T. K. (2006). *Handbook of latent semantic analysis*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Leacock, C., & Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 265–283). Cambridge, MA: MIT Press.

Lee, W.-N., Shah, N., Sundlass, K., & Musen, M. (2008). Comparison of ontology-based semantic-similarity measures. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, 384–388.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics, 3*, 211–225.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 296–304). Madison, WI.

Liu, Y., McInnes, B. T., Pedersen, T., Melton-Meaux, G., & Pakhomov, S. (2012). Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet (p. 363). ACM Press.

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, *19*(10), 1275–1283.

Mazandu, G. K., Chimusa, E. R., Mbiyavanga, M., & Mulder, N. J. (2016). A-DaGO-Fun: an adaptable Gene Ontology semantic similarity-based functional analysis tool. *Bioinformatics*, *32*(3), 477–479.

McCormick, N., Lacaille, D., Bhole, V., & Avina-Zubieta, J. A. (2014). Validity of Heart Failure Diagnoses in Administrative Databases: A Systematic Review and Meta-Analysis. *PLoS ONE*, *9*(8), e104519.

McEwan, R., Melton, G., Knoll, B., Wang, Y., Hultman, G., Dale, J., … Pakhomov, S. (2016). NLP-PIER: A Scalable Natural Language Processing, Indexing, and Searching Architecture for Clinical Notes. In *Proceedings of the 2016 Joint Summits of the American medical Informatics Association* (in press). San Francisco, CA.

McInnes, B., Pedersen, T., Liu, Y., Pakhomov, S. V., & Melton, G. (2011). Using second-order vectors in a knowledge-based method for acronym disambiguation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (pp. 145–153). Portland, Oregon, USA.

McInnes, B. T., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of Biomedical Informatics*, *46*(6), 1116–1124.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119).

Muneeb, T., Sunil, K., & Anand, A. (2015). Evaluating distributed word representations for capturing semantics of biomedical concepts. In *Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015)* (pp. 158–163). Bajing, China.

Pakhomov, S., Buntrock, J., & Chute, C. G. (2005). Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. *Journal of Biomedical Informatics*, *38*(2), 145–153.

Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., & Melton, G. B. (2010). Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2010*, 572–576.

Pakhomov, S., Weston, S. A., Jacobsen, S. J., Chute, C. G., Meverden, R., & Roger, V. L. (2007). Electronic medical records for clinical research: application to the identification of heart failure. *The American Journal of Managed Care*, *13*(6 Part 1), 281–288.

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. *In Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, 1–8.

Pedersen, T., Pakhomov, S. V., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform*, *40*, 288–99.

Resnik, P. (1999). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J of Artif Intell Res*, *11*, 95–130.

Sajadi, A. (2014). Graph-Based Domain-Specific Semantic Relatedness from Wikipedia. In M. Sokolova & P. van Beek (Eds.), *Advances in Artificial Intelligence* (Vol. 8436, pp. 381–386).

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, *23*(10), 1274–1281.

Weeds, J., & Weir, D. (2005). Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. *Computational Linguistics*, *31*(4), 439–475.

Yang, H., Nepusz, T., & Paccanaro, A. (2012). Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics*, *28*(10), 1383–1389.