Retrofitting Word Vectors of MeSH Terms to Improve Semantic Similarity Measures

Zhiguo Yu, MS, Trevor Cohen, MBChB, PhD
Elmer V. Bernstam, MD, MSE, Todd R. Johnson, PhD
The University of Texas Health Science Center at Houston, Houston, TX
Zhiguo.yu@uth.tmc.edu

Byron C. Wallace, PhD
College of Computer and Information Science, Northeastern University, Boston, MA

Abstract

Estimation of the semantic relatedness between biomedical concepts has utility for many informatics applications. Automated methods fall into two broad categories: methods based on distributional statistics drawn from text corpora, and methods based on the structure of existing knowledge resources. In the former case, taxonomic structure is disregarded. In the latter, semantically relevant empirical information is not considered. In this paper, we present a method that retrofits the context vector representation of MeSH terms by using additional linkage information from UMLS/MeSH hierarchy such that linked concepts have similar vector representations. We evaluated the method relative to previously published physician and coder's ratings on sets of MeSH terms. Our experimental results demonstrate that the retrofitted word vector measures obtain a higher correlation with physician judgments. The results also demonstrate a clear improvement on the correlation with experts' ratings from the retrofitted vector representation in comparison to the vector representation without retrofitting.

1 Introduction

Groups of semantically similar concepts and terms are known to improve the retrieval (Rada et al., 1989) and clustering (Lin et al., 2007) of biomedical and clinical documents, and the development of biomedical terminologies and ontologies (Bodenreider and Burgun, 2004). However, automated estimation of semantic similarity remains a challenge. Most semantic similarity measures leverage the structure of an ontology or taxonomy (e.g. WordNet, Unified Medical Language System (UMLS)/Medical Subject Headings (MeSH)) to calculate, for example, the shortest path information between concept nodes (Pedersen et al., 2007; Caviedes and Cimino, 2004). Vector representations based on a co-occurrence matrix from a corpus has also been used to calculate the relatedness between concepts (Pedersen et al., 2007; Pedersen et al., 2004). Others use information content (IC) to estimate the semantic similarity and relatedness between two concepts, which incorporate the probability of the concept occurring in a corpus (Caviedes and Cimino, 2004; Ciaramita et al., 2008; Turney, 2005). Some topic modeling techniques (Blei et al., 2003; Yu et al., 2013) have also been applied to integrate the automatically generated themes (topics) from a specific corpus to the controlled vocabulary that indexed within this corpus to help improve the document retrieval and clustering performances (Yu et al., 2016).

In this paper, we introduce a new semantic similarity measure utilizing both vector space word representations and a biomedical taxonomy (UMLS/MeSH) to determine the degree of semantic similarity between pairs of concepts. For two concepts, we first learn their vector space word representations from distributional information of words in a large domain-relevant corpus. Although such vectors are semantically informative, they disregard the valuable information contained in semantic lexicons such as WordNet, FrameNet, and the Paraphrase Database. In 2014, Faruqui, et al. (Faruqui et al., 2014a) developed a "retrofitting"

Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 43–51,
Austin, TX, November 5, 2016.

method that addresses this limitation by incorporating information from such semantic lexicons into word vector representations, such that semantically linked words will have similar vector representations. We applied this technique to word vector representations of UMLS/MeSH concepts in an effort to improve their quality. We evaluated the method relative to previously published human expert similarity ratings of a Physician and Coder on sets of MeSH terms. Our experimental results demonstrate that the retrofitted word vector similarity measures have a higher correlation with Physician (but not Coder) judgments, compared with other existing techniques. The results also demonstrate a clear improvement on the correlation with experts' ratings from the retrofitted vector representation to the vector representation without retrofitting.

2 Related Work

There are two major classes of semantic similarity measurement methods. The most common class uses an ontology or taxonomy to calculate the shortest path between two concepts. Rada, et al. (Rada et al., 1989) introduces the measure of conceptual distance to quantify the similarity between concepts in the UMLS. Wu and Palmer (Wu and Palmer, 1994) extend this measure by calculating the length of shortest path between two concepts that connects the concepts through their least common subsumer (LCS). The LCS is the most specific ancestor shared by two concepts. In 2005, Nguyen and Al-Mubaid (Nguyen and Al-Mubaid, 2006) proposed a new path-based measure using $is - a$ relation in MeSH. They incorporate both the depth and LCS in their measure. In their results, they compared with the measures introduced by Leacock & Chodorow (Leacock and Chodorow, 1998), Wu & Palmer (Wu and Palmer, 1994), and the Path measure. Batet, et al. (Batet et al., 2011) introduce a measure that incorporates the common concepts shared between the two concepts and their LCS. Recently, McInnes, et al. (McInnes et al., 2014) introduced U-path measure using undirected path to determine the degree of semantic similarity between two concepts in a dense taxonomy with multiple inheritance. In 2009, McInnes, et al. (McInnes et al., 2009) presented a UMLS-Similarity tool which contains five semantic similarity measures proposed by Rada, et al. (Rada et al., 1989) , Wu & Palmer (Wu and Palmer, 1994), Leacock & Chodorow (Leacock and Chodorow, 1998), and Nguyen & Al-Mubaid (Nguyen and Al-Mubaid, 2006), and the Path measure.

The second class of techniques uses training corpora and information content (IC) to estimate the semantic similarity between two concepts. IC measures the specificity of a concept in a hierarchy. The IC-based measures account for the probability of the concept occurring in a corpus. A concept with a high IC value is more specific to a topic than one with a low IC value. Resnik (Resnik, 1995), Jiang & Conrath (Jiang and Conrath, 1997) and Lin (Lin, 1998), all have published works on the IC-based similarity measures. Resnik (Resnik, 1995) measures the similarity between two concepts by finding the IC of the LCS of the two concepts. Jiang & Conrath (Jiang and Conrath, 1997) and Lin (Lin, 1998) extended Resnik's IC-based measure by incorporating the IC of the individual concepts. Jiang & Conrath measure similarity by finding the IC of each individual concept and of the LCS of them. However, Lin's measure is similar to that of Wu & Palmer (Wu and Palmer, 1994), where depth is replaced by information content.

Context vector metrics based on distributional statistics have also been used to calculate semantic similarity (Patwardhan, 2006; Patwardhan, 2003). By building co-occurrence vectors that represent the contextual profile of concepts, the relatedness between concepts can then be calculated using cosine similarity between vectors corresponding to two given concepts (Pedersen et al., 2007).

Though IC-based measures do draw upon distributional information, this is used in a very restricted way to determine the specificity of a concept. Context vector metric-based distributional statistics do not have such limitations on the use of distributional information. However, the taxonomic structure is not taken into account in distributional methods. "Correlation with human pairwise judgment" evaluation is widely used in computational linguistics. There are a number of evaluation sets exist in the biomedical domain. 'MayoSRS', developed by

Pakhomov, et al. (Pakhomov et al., 2011), consists of 101 clinical term pairs whose relatedness was determined by nine medical coders and three physicians from the Mayo Clinic. In this paper, we used 'MiniMayoSRS,' a subset of 'MayoSRS.' The average correlation between physicians is 0.68. The average correlation between medical coders is 0.78 (Pedersen et al., 2007). 'UMNSRS', developed by Pakhomov, et al. (Pakhomov et al., 2010), consists of 725 clinical term pairs whose semantic similarity and relatedness was determined independently by four medical residents from the University of Minnesota Medical School.

3 Method

In this section, we provide a brief description of the method used for retrofitting word vector to semantic lexicons, present the design of our work flow, describe the test data and the semantic lexicon we created , and also present the evaluation measures we used.

3.1 Retrofitting Word Vector to Semantic Lexicons

Vector space word representations are a critical component of many natural language processing systems. It is common to represent words as discrete indices in a vocabulary, but this fails to capture the rich relational structure of the human semantic lexicon (Maas et al., 2011). Retrofitting is a simple and effective method to improve word vectors using word relation knowledge found in semantic lexicons. It is used as a post-processing step to improve vector quality (Faruqui et al., 2014a).

Figure 1 shows a small word graph example with edges connecting semantically related words. The words, $cancer$, $tumor$, $neoplasm$, $sarcoma$, and $swelling$, are similar words to each other in a lexical knowledge resource. Grey nodes are observed word vectors built from the corpus, which are independent of each other. White nodes are inferred word vectors, waiting to be retrofitted. The edge between each pair of white nodes means they are similar words to each other. The inferred word vector (e.g., q_tumor) is expected to be close to its corresponding observed word vector (e.g., $q\hat{}\_tumor$) and close to its synonym neighbors (e.g., q_cancer and $q\_neoplasm$). The objective is to minimize the following:

$$\Psi(Q) = \sum_{i=1}^{n}[\alpha_i \left\|q_i - \hat{q_i}\right\|^2 + \sum_{(i,j)\in E} \beta_{ij} \left\|q_i - q_j\right\|^2] \quad (1)$$

where α and β values control the relative strengths of associations, Q is the retrofitted vectors, and $(i, j) \in E$ means there is an edge between node q_i and q_j. Ψ is convex in Q. An efficient iterative updating method is used to find this convex. First, retrofitted vectors in Q are initialized to be equal to the observed vectors. The next step is to take the first derivative of Ψ with respect to q_i vector and use the following to update it online.

$$q_i = \frac{\sum_{j:(i,j)\in E} \beta_{ij} q_j + \alpha_i \hat{q_i}}{\sum_{j:(i,j)\in E} \beta_{ij} + \alpha_i} \quad (2)$$

It takes approximately 10 iterations to converge to the difference in Euclidean distance of adjacent nodes of less than 0.01 in practice. An implementation of this algorithm has been published online by the authors (Faruqui et al., 2014b). We used this implementation in the current work.

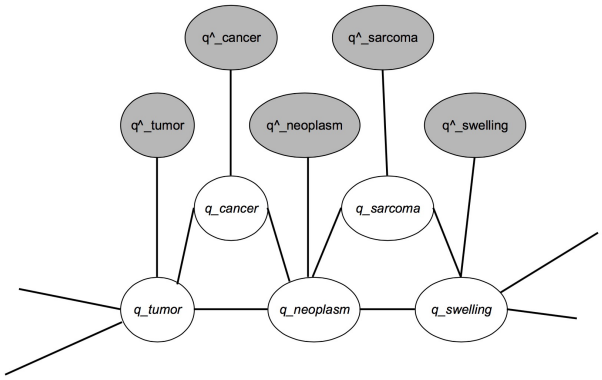

Figure 1: Word graph with edges between related words, observed (grey node), inferred (white node).

3.2 Work Flow

Our work flow is presented in Figure 2. The input is a pair of concepts. The output is a similarity score. The next step after *input data* is to *fetch relevant*

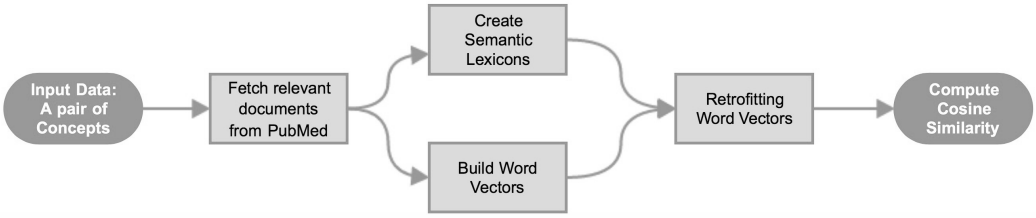

Figure 2: Work flow.

documents from PubMed. In our test data, each concept is mapped to MeSH term(s) (Please see details in the paragraph *Test Data* of this section.) We then randomly fetch 1000 citations indexed with those MeSH term(s) from PubMed. In the *Build Word Vectors* step, we build each MeSH term a word vector using the approach described in (Yu et al., 2016). We use titles and abstracts of returned citations and only select those MeSH terms indexed in more than 100 citations as our candidate semantic lexicon. The main MeSH term mapped from the input concept is indexed in all 1000 citations. The retrofitted vector quality suffers if we take into account MeSH terms that appear in a small number of citations. For each selected MeSH term, we collect all the words from the citations indexed with that MeSH term. After removing the stop words, We use *tf-idf* (Equation 3) to weight the remaining words and then normalize the weights so that they sum to one. In *Create Semantic Lexicons*, we use both the UMLS-similarity tool developed by McInnes, et al. (McInnes et al., 2009) and the MeSH tree structure as the source from which it estimates semantic relatedness. For details see the paragraph *Semantic Lexicons* in this section. *Retrofitting Word Vectors* retrofits the word vectors using the created semantic lexicons to generate new word vectors. We then calculate cosine similarity (Equation 4) based on the concepts pair's new word vectors. On account of the stochastic nature of the literature sampling, we test each pair of concepts five times and average performance over these five times as its final similarity score.

$$\textit{tf-idf}_{w,d} = \textit{tf}_{w,d} * log\frac{N}{df_{w,D}} \quad (3)$$

where $tf_{w,d}$ is the term frequency of word w in document d, $df_{w,D}$ is the document frequency that word w appears in all documents D, and N is the total number of documents.

$$cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2} \quad (4)$$

where A_i and B_i are components of vector A and B respectively.

3.3 Test Data

We used the set of 30 concept pairs from Pedersen, Pakhomov, and Patwardhan (2005) (Pedersen et al., 2007), which was annotated by 3 physicians and 9 medical index coders. Each pair was annotated on a 4 point scale: "*practically synonymous, related, marginally, and unrelated*". Table 1 displays the details of these concepts pairs along with both ratings.

Neguyen and Al-Mubaid use 25 out of the 30 pairs of terms in the dataset. 5 pairs of terms (highlighted in both table 1 and table 2) were excluded because they did not exist in MeSH version 2006. To make it comparable with their results, we also use these 25 pairs of terms. The mappings of the terms to MeSH terms were obtained firstly by using the online MetaMap tool (Aronson and Lang, 2010). Then we used the MeSH browser 2016 (MeSH, 2016) to get the most updated MeSH terms.

3.4 Semantic Lexicons

We tested two semantic lexicons in our experiments. The first is from the results of McInnes, et al.'s UMLS-Similarity tool (McInnes et al., 2009). UMLS-Similarity contains five semantic similarity measures proposed by Rada, et al. (Rada et al., 1989), Wu & Palmer (wup) (Wu and Palmer, 1994), Leacock & Chodorow (lch) (Leacock and Chodorow, 1998), and Nguyen & Al-Mubaid (nam) (Nguyen and Al-Mubaid, 2006), and the Path measure. Leacock & Chodorow's measure achieved best performance among these five semantic similarity measures. In our experiment, we used this mea-

Term 1	Term 2	Physicians	Coders
Renal failure	Kidney failure	4.0000	4.0000
Heart	Myocardium	3.3333	3.0000
Stroke	Infarct	3.0000	2.7778
Abortion	miscarriage	3.0000	3.3333
Delusion	Schizophrenia	3.0000	2.2222
Congestive heart failure	Pulmonary edema	3.0000	1.4444
Metastasis	Adenocarcinoma	2.6667	1.7778
Calcification	Stenosis	2.6667	2.0000
Diarrhea	**Stomach cramps**	2.3333	1.3333
Mitral stenosis	Atrial fibrillation	2.3333	1.3333
Chronic obstructive pulmonary disease	**Lung infiltrates**	2.3333	1.8889
Rheumatoid arthritis	Lupus	2.0000	1.1111
Brain tumor	Intracranial hemorrhage	2.0000	1.3333
Carpal tunnel syndrome	Osteoarthritis	2.0000	1.1111
Diabetes mellitus	Hypertension	2.0000	1.0000
Acne	Syringe	2.0000	1.0000
Antibiotic	Allergy	1.6667	1.2222
Cortisone	Total knee replacement	1.6667	1.0000
Pulmonary embolus	**Myocardial infarction**	1.6667	1.2222
Pulmonary Fibrosis	Lung Cancer	1.6667	1.4444
Cholangiocarcinoma	Colonoscopy	1.3333	1.0000
Lymphoid hyperplasia	Laryngeal Cancer	1.3333	1.0000
Multiple Sclerosis	Psychosis	1.0000	1.0000
Appendicitis	Osteoporosis	1.0000	1.0000
Rectal polyp	**Aorta**	1.0000	1.0000
Xerostomia	Alcoholic cirrhosis	1.0000	1.0000
Peptic ulcer disease	Myopia	1.0000	1.0000
Depression	Cellulitis	1.0000	1.0000
Varicose vein	**Entire knee meniscus**	1.0000	1.0000
Hyperlipidemia	Metastasis	1.0000	1.0000

Table 1: Test set of 30 medical term pairs sorted in the order of the averaged physician‘ scores.

sure in UMLS-Similarity to calculate the similarity score between each selected MeSH term and the main MeSH term. We calculated the average of all these scores as the threshold. We then chose those MeSH terms whose scores are over this threshold as the main MeSH term's semantic lexicon terms. The second semantic lexicon is constructed using MeSH tree structure information. For each main MeSH term, we chose its parents and child terms from the MeSH tree as its lexicon terms.

3.5 Evaluation

In our experiment, we used three types of vector representations to calculate the semantic similarity: MeSH term word vectors without retrofitting; MeSH term word vectors retrofitted with UMLS-Similarity results; and MeSH term word vectors retrofitted using the MeSH tree structure. We rank the 25 pairs of terms based on similarity scores and calculate the correlation between our rankings and the Physician and Coder judgments using the Spearman rank correlation coefficient. We compare our correlation results with those reported by Nguyen, et al. (Nguyen and Al-Mubaid, 2006) and

Term 1	Term 2	Word Vector	Retrofitted with UMLS-Similarity Results	Retrofitted with MeSH Tree Structure
Renal failure	Kidney failure	1.00	1.00	1.00
Heart	Myocardium	0.86	0.85	0.86
Stroke	Infarct	0.70	0.71	0.70
Abortion	miscarriage	0.79	0.74	0.76
Delusion	Schizophrenia	0.81	0.83	0.81
Congestive heart failure	Pulmonary edema	0.73	0.72	0.73
Metastasis	Adenocarcinoma	0.88	0.84	0.83
Calcification	Stenosis	0.47	0.46	0.47
Diarrhea	**Stomach cramps**	N/A	N/A	N/A
Mitral stenosis	Atrial fibrillation	0.71	0.71	0.71
Chronic obstructive pulmonary disease	**Lung infiltrates**	N/A	N/A	N/A
Rheumatoid arthritis	Lupus	0.70	0.71	0.70
Brain tumor	Intracranial hemorrhage	0.69	0.68	0.69
Carpal tunnel syndrome	Osteoarthritis	0.66	0.66	0.66
Diabetes mellitus	Hypertension	0.82	0.81	0.81
Acne	Syringe	0.54	0.54	0.54
Antibiotic	Allergy	0.67	0.67	0.67
Cortisone	Total knee replacement	0.47	0.44	0.47
Pulmonary embolus	**Myocardial infarction**	N/A	N/A	N/A
Pulmonary Fibrosis	Lung Cancer	0.72	0.70	0.72
Cholangiocarcinoma	Colonoscopy	0.63	0.62	0.61
Lymphoid hyperplasia	Laryngeal Cancer	0.70	0.70	0.70
Multiple Sclerosis	Psychosis	0.69	0.67	0.67
Appendicitis	Osteoporosis	0.55	0.55	0.54
Rectal polyp	**Aorta**	N/A	N/A	N/A
Xerostomia	Alcoholic cirrhosis	0.67	0.67	0.66
Peptic ulcer disease	Myopia	0.47	0.47	0.48
Depression	Cellulitis	0.55	0.54	0.54
Varicose vein	**Entire knee meniscus**	N/A	N/A	N/A
Hyperlipidemia	Metastasis	0.56	0.55	0.55

Table 2: Results of Word vector representations.

generated by UMLS-Similarity tool (McInnes et al., 2009).

4 Results and Discussion

Table 2 shows the term pairs in the dataset and the similarity of the terms determined by our measure using three different vector representations. Table

Measures		Physician	Coder
path	Nguyen and Al-Mubaid	0.627	0.852
path	UMLS-Similarity	0.486	0.581
lch	Nguyen and Al-Mubaid	0.672	0.856
lch	UMLS-Similarity	0.486	0.581
wup	Nguyen and Al-Mubaid	0.652	0.794
wup	UMLS-Similarity	0.453	0.535
nam	Nguyen and Al-Mubaid	0.666	**0.862**
nam	UMLS-Similarity	0.448	0.551
Vector representation	Without retrofitting	0.646	0.632
Vector representation	Retrofitted with UMLS-Similarity	**0.696**	0.665
Vector representation	Retrofitted with MeSH tree structure	0.675	0.655

Table 3: Spearman's Rank Correlation Results. Our results are compared with the results of these different measures reported by Nguyen and Al-Mubaid and also generated by UMLS-Similarity tool. path:path based similarity measure; lch: the similarity measure proposed by Leacock & Chodorow in 1998; wup: the similarity measure proposed by Wu & Palmer in 1994; nam:the similarity measure proposed by Nguyen & Al-Mubaid in 2006

3 shows the correlation results between our methods and the judgments made by physicians and coders, as well as the results reported by Nguyen, et al. (Nguyen and Al-Mubaid, 2006) and McInnes, et al. (McInnes et al., 2009), using the UMLS-Similarity tool.

From Table 3, we can see our retrofitted vector representation with UMLS-Similarity obtains a highest correlation with the Physician judgments. Though our retrofitted vector representation achieved a lower correlation with the Coder judgments than the results reported by Nguyen and Al-Mubaid (Nguyen and Al-Mubaid, 2006), we still see an improvement from the retrofitted vector representations as compared with the original vector representation without retrofitting. Since UMLS-Similarity's results are lower than our vector representations, it is understandable that our retrofitted vector representations still can not surpass the results achieved by Nguyen and Al-Mubaid's method. From Table 3, we can also see that our vector representations obtain lower correlations with the coder judgments than with the physician judgments. This contrasts with both the UMLS-Similarity results and those reported by Nguyen and Al-Mubaid. We believe that the reason for this phenomenon is that the coder group were more familiar with the ontology or taxonomy than the physician group. When reviewing these pairs of concepts, coders may interpret the terms in relation to the ontology or taxonomy, whereas physicians may be more likely to understand them at a broader contextual level. Because our vector representation methods all originated as context vectors, this may explain why our methods achieved higher correlation with physician judgments.

Among the three types of vector representations, the retrofitted vector representation with UMLS-Similarity had a higher correlation with both physician and coder judgments than the vectors retrofitted using the MeSH tree structure. We believe this occurred because the way we created the semantic lexicon from the MeSH tree structure had a limited effect on the original vector representations. From Table 2, we can see that the semantic lexicon based on the MeSH tree structure only affected 10 of 25 pairs of terms. The semantic lexicon based on UMLS-Similarity results affected 16 of 25 pairs of terms. We used the MeSH term's parents and children as the lexicon terms, and it is unlikely for a PubMed article to be indexed with both parent and child terms. The UMLS-Similarity approach is more permissive. Two MeSH terms are accepted as a lexicon term only when they have above-threshold similarity as estimated by path-based measures.

5 Conclusions and Future Work

In this paper, we introduced a semantic similarity measure that utilizes vector word representation and the linkage information in an ontology or taxonomy. By retrofitting vector representations with additional ontology or taxonomy information, we can generate vector representations in which lexically-linked concepts are more likely to have similar vector representations. This leads to better approximation of human judgments on the task of estimating semantic relatedness. We show that our method obtains a higher correlation with physician judgments than UMLS-Similarity, and previously reported results. We also demonstrate a clear improvement from the retrofitted vector representation as compared to the vector representation without retrofitting. In the future we plan to expand this technique to other knowledge sources and datasets. We also plan to use more sophisticated and better established approaches to generate concept vectors, e.g. methods of distributional semantic (Cohen et al., 2010), word embedding (Mikolov et al., 2013), and compare with more recently evaluations using neural network based similarity and relatedness measures (Pakhomov et al., 2016).

Acknowledgments

This work was supported in part by the UTHealth Innovation for Cancer Prevention Research Training Program Predoctoral Fellowship (Cancer Prevention and Research Institute of Texas (CPRIT) grant # RP140103), NIH NCATS grant UL1 TR000371, NIH NCI grant U01 CA180964 and the Brown Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the CPRIT.

References

Alan R. Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *JAMIA*, 17(3):229–236.

Montserrat Batet, David Sánchez, and Aida Valls. 2011. An ontology-based measure to compute semantic similarity in biomedicine. *J. of Biomedical Informatics*, 44(1):118–125, February.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Olivier Bodenreider and Anita Burgun. 2004. Aligning knowledge sources in the umls: methods, quantitative results, and applications. *Studies in health technology and informatics*, 107(0 1):327.

Jorge E Caviedes and James J Cimino. 2004. Towards the development of a conceptual distance metric for the umls. *Journal of biomedical informatics*, 37(2):77–85.

Massimiliano Ciaramita, Aldo Gangemi, and Esther Ratsch. 2008. Unsupervised learning of semantic relations for molecular biology ontologies. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 91–107.

Trevor Cohen, Roger Schvaneveldt, and Dominic Widdows. 2010. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of biomedical informatics*, 43(2):240–256.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014a. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014b. Retrofitting word vectors to semantic lexicons code.

Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

Yongjing Lin, Wenyuan Li, Keke Chen, and Ying Liu. 2007. A document clustering and ranking system for exploring medline citations. *Journal of the American Medical Informatics Association*, 14(5):651–661.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 142–150, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bridget T McInnes, Ted Pedersen, and Serguei VS Pakhomov. 2009. Umls-interface and umls-similarity:

open source software for measuring paths and semantic similarity. In *AMIA Annual Symposium Proceedings*, volume 2009, page 431. American Medical Informatics Association.

Bridget T McInnes, Ted Pedersen, Ying Liu, Genevieve B Melton, and Serguei V Pakhomov. 2014. U-path: An undirected path-based measure of semantic similarity. In *AMIA Annual Symposium Proceedings*, volume 2014, page 882. American Medical Informatics Association.

MeSH. 2016. National library of medicine (nlm), mesh browser.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Hoa A Nguyen and Hisham Al-Mubaid. 2006. New ontology-based semantic similarity measure for the biomedical domain. In *Granular Computing, 2006 IEEE International Conference on*, pages 623–628. IEEE.

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. In *AMIA annual symposium proceedings*, volume 2010, page 572. American Medical Informatics Association.

Serguei VS Pakhomov, Ted Pedersen, Bridget McInnes, Genevieve B Melton, Alexander Ruggieri, and Christopher G Chute. 2011. Towards a framework for developing semantic relatedness reference standards. *Journal of biomedical informatics*, 44(2):251–265.

Serguei VS Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B Melton. 2016. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, page btw529.

Siddharth Patwardhan. 2003. *Incorporating dictionary and corpus information into a context vector measure of semantic relatedness*. Ph.D. thesis, University of Minnesota, Duluth.

Siddharth Patwardhan. 2006. Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *In: Proceedings of the EACL*, pages 1–8.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

Ted Pedersen, Serguei V. S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *J. of Biomedical Informatics*, 40(3):288–299, June.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1):17–30.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Peter D. Turney. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 1136–1141, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhiguo Yu, Todd R Johnson, and Ramakanth Kavuluru. 2013. Phrase based topic modeling for semantic information processing in biomedicine. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 440–445. IEEE.

Zhiguo Yu, Elmer Bernstam, Trevor Cohen, Byron C Wallace, and Todd R Johnson. 2016. Improving the utility of mesh® terms using the topicalmesh representation. *Journal of biomedical informatics*, 61:77–86.