Improving Correlation with Human Judgments by Integrating Semantic Similarity with Second–Order Vectors

Bridget T. McInnes
Department of Computer Science
Virginia Commonwealth University
Richmond, VA 23284 USA
btmcinnes@vcu.edu

Ted Pedersen
Department of Computer Science
University of Minnesota
Duluth, MN 55812 USA
tpederse@d.umn.edu

Abstract

Vector space methods that measure semantic similarity and relatedness often rely on distributional information such as co–occurrence frequencies or statistical measures of association to weight the importance of particular co–occurrences. In this paper, we extend these methods by incorporating a measure of semantic similarity based on a human curated taxonomy into a second–order vector representation. This results in a measure of semantic relatedness that combines both the contextual information available in a corpus–based vector space representation with the semantic knowledge found in a biomedical ontology. Our results show that incorporating semantic similarity into a second order co–occurrence matrices improves correlation with human judgments for both similarity and relatedness, and that our method compares favorably to various different word embedding methods that have recently been evaluated on the same reference standards we have used.

1 Introduction

Measures of semantic similarity and relatedness quantify the degree to which two concepts are similar (e.g., *lung–heart*) or related (e.g., *lung–bronchitis*). Semantic similarity can be viewed as a special case of semantic relatedness – to be similar is one of many ways that a pair of concepts may be related. The automated discovery of groups of semantically similar or related terms is critical to improving 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).

There is a long history in using distributional methods to discover semantic similarity and relatedness (e.g., (Lin and Pantel, 2002; Reisinger and Mooney, 2010; Radinsky et al., 2011; Yih and Qazvinian, 2012)). These methods are all based on the distributional hypothesis, which holds that two terms that are distributionally similar (i.e., used in the same context) will also be semantically similar (Harris, 1954; Weeds et al., 2004). Recently word embedding techniques such as word2vec (Mikolov et al., 2013) have become very popular. Despite the prominent role that neural networks play in many of these approaches, at their core they remain distributional techniques that typically start with a word by word co–occurrence matrix, much like many of the more traditional approaches.

However, despite these successes distributional methods do not perform well when data is very sparse (which is common). One possible solution is to use second–order co–occurrence vectors (Schütze, 1992; Schütze, 1998). In this approach the similarity between two words is not strictly based on their co–occurrence frequencies, but rather on the frequencies of the other words which occur with both of them (i.e., second order co–occurrences). This approach has been shown to be successful in quantifying semantic relatedness (Islam and Inkpen, 2006; Pedersen et al., 2007). However, while more robust in the face of sparsity, second–order methods can result in significant amounts of noise, where contextual information that is overly general is included and does not contribute to quantifying the semantic relatedness between the two concepts.

Our goal then is to discover methods that automatically reduce the amount of noise in a second–order co–occurrence vector. We achieve this by incorporating pairwise semantic similarity scores

Proceedings of the BioNLP 2017 workshop, pages 107–116,
Vancouver, Canada, August 4, 2017. ©2017 Association for Computational Linguistics

derived from a taxonomy into our second–order vectors, and then using these scores to select only the most semantically similar co–occurrences (thereby reducing noise).

We evaluate our method on two datasets that have been annotated in multiple ways. One has been annotated for both similarity and relatedness, and the other has been annotated for relatedness by two different types of experts (medical doctors and medical coders). Our results show that integrating second order co–occurrences with measures of semantic similarity increases correlation with our human reference standards. We also compare our result to a number of other studies which have applied various word embedding methods to the same reference standards we have used. We find that our method often performs at a comparable or higher level than these approaches. These results suggest that our methods of integrating semantic similarity and relatedness values have the potential to improve performance of purely distributional methods.

2 Similarity and Relatedness Measures

This section describes the similarity and relatedness measures we integrate in our second–order co–occurrence vectors. We use two taxonomies in this study, SNOMED–CT and MeSH. SNOMED–CT (*Systematized Nomenclature of Medicine Clinical Terms*) is a comprehensive clinical terminology created for the electronic representation of clinical health information. MeSH (*Medical Subject Headings*) is a taxonomy of biomedical terms developed for indexing biomedical journal articles.

We obtain SNOMED–CT and MeSH via the Unified Medical Language System (UMLS) Metathesaurus (version 2016AA). The Metathesaurus contains approximately 2 million biomedical and clinical concepts from over 150 different terminologies that have been semi–automatically integrated into a single source. Concepts in the Metathesaurus are connected largely by two types of hierarchical relations: $parent/child$ (PAR/CHD) and $broader/narrower$ (RB/RN).

2.1 Similarity Measures

Measures of semantic similarity can be classified into three broad categories : path–based, feature–based and information content (IC). Path–based similarity measures use the structure of a taxonomy to measure similarity – concepts positioned close to each other are more similar than those further apart. Feature–based methods rely on set theoretic measures of overlap between features (union and intersection). The information content measures quantify the amount of information that a concept provides – more specific concepts have a higher amount of information content.

2.1.1 Path–based Measures

Rada et al. (1989) introduce the *Conceptual Distance* measure. This measure is simply the length of the shortest path between two concepts ($c1$ and $c2$) in the MeSH hierarchy. Paths are based on *broader than* (RB) and *narrower than* (RN) relations. Caviedes and Cimino (2004) extends this measure to use *parent* (PAR) and *child* (CHD) relations. Our $path$ measure is simply the reciprocal of this shortest path value (Equation 1), so that larger values (approaching 1) indicate a high degree of similarity.

$$path = \frac{1}{spath(c_1, c_2)} \tag{1}$$

While the simplicity of $path$ is appealing, it can be misleading when concepts are at different levels of specificity. Two very general concepts may have the same path length as two very specific concepts. Wu and Palmer (1994) introduce a correction to $path$ that incorporates the depth of the concepts, and the depth of their Least Common Subsumer (LCS). This is the most specific ancestor two concepts share. In this measure, similarity is twice the depth of the two concept's LCS divided by the product of the depths of the individual concepts (Equation 2). Note that if there are multiple LCSs for a pair of concepts, the deepest of them is used in this measure.

$$wup = \frac{2 * depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)} \tag{2}$$

Zhong et al. (2002) take a very similar approach and again scale the depth of the LCS by the sum of the depths of the two concepts (Equation 3), where $m(c) = k^{-depth(c)}$. The value of k was set to 2 based on their recommendations.

$$zhong = \frac{2 * m(lcs(c_1, c_2))}{m(c_1) + m(c_2)} \tag{3}$$

Pekar and Staab (2002) offer another variation on $path$, where the shortest path of the two concepts to the LCS is used, in addition to the shortest bath between the LCS and the root of the taxonomy (Equation 4).

$$pks = -\log \frac{spath(lcs(c_1, c_2), root)}{\sum_{x=c_1,c_2,root} spath(lcs(c_1, c_2), x)} \tag{4}$$

2.1.2 Feature–based Measures

Feature–based methods represent each concept as a set of features and then measure the overlap or sharing of features to measure similarity. In particular, each concept is represented as the set of their ancestors, and similarity is a ratio of the intersection and union of these features.

Maedche and Staab (2001) quantify the similarity between two concepts as the ratio of the intersection over their union as shown in Equation 5.

$$cmatch = \frac{|A(c_1) \bigcap A(c_2)|}{|A(c_1) \bigcup A(c_2)|} \tag{5}$$

Batet et al. (2011) extend this by excluding any shared features (in the numerator) as shown in Equation 6.

$$batet = -log_2(\frac{|A(c_1) \bigcup A(c_2)| - |A(c_1) \bigcap A(c_2)|}{|A(c_1) \bigcup A(c_2)|}) \tag{6}$$

2.1.3 Information Content Measures

Information content is formally defined as the negative log of the probability of a concept. The effect of this is to assign rare (low probability) concepts a high measure of information content, since the underlying assumption is that more specific concepts are less frequently used than more common ones.

Resnik (1995) modified this notion of information content in order to use it as a similarity measure. He defines the similarity of two concepts to be the information content of their LCS (Equation 7).

$$res = IC(lcs(c_1, c_2) = -\log(P(lcs(c_1, c_2))) \tag{7}$$

Jiang and Conrath (1997), Lin (1998), and Pirró and Euzenat (2010) extend res by incorporating the information content of the individual concepts in various different ways. Lin (1998) defines the similarity between two concepts as the ratio of information content of the LCS with the sum of the individual concept's information content (Equation 8). Note that lin has the same form as wup and $zhong$, and is in effect using information content as a measure of specificity (rather than depth). If there is more than one possible LCS, the LCS with the greatest IC is chosen.

$$lin = \frac{2 * IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{8}$$

Jiang and Conrath (1997) define the distance between two concepts to be the sum of the information content of the two concepts minus twice the information content of the concepts' LCS. We modify this from a distance to a similarity measure by taking the reciprocal of the distance (Equation 9). Note that the denominator of jcn is very similar to the numerator of $batet$.

$$jcn = \frac{1}{IC(c_1) + IC(c_2) - 2 * IC(lcs(c_1, c_2))} \tag{9}$$

Pirró and Euzenat (2010) define the similarity between two concepts as the information content of the two concept's LCS divided by the sum of their individual information content values minus the information content of their LCS (Equation 10). Note that $batet$ can be viewed as a set–theoretic version of $faith$.

$$faith = \frac{IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2) - IC(lcs(c_1, c_2))} \tag{10}$$

2.2 Information Content

The information content of a concept may be derived from a corpus (corpus–based) or directly from a taxonomy (intrinsic–based). In this work we focus on corpus–based techniques.

For corpus–based information content, we estimate the probability of a concept c by taking the sum of the probability of the concept $P(c)$ and the probability its descendants $P(d)$ (Equation 11).

$$P(c*) = P(c) + \sum_{d \in descendant(c)} P(d) \tag{11}$$

The initial probabilities of a concept ($P(c)$) and its descendants ($P(d)$) are obtained by dividing the number of times each concept and descendant occurs in the corpus, and dividing that by the total numbers of concepts (N).

Ideally the corpus from which we are estimating the probabilities of concepts will be sense–tagged. However, sense–tagging is a challenging problem in its own right, and it is not always possible to carry out reliably on larger amounts of text. In fact in this paper we did not use any sense–tagging of the corpus we derived information content from.

Instead, we estimated the probability of a concept by using the *UMLSonMedline* dataset. This was created by the National Library of Medicine and consists of concepts from the 2009AB UMLS and the counts of the number of times they occurred in a snapshot of Medline taken on 12 January, 2009. These counts were obtained by using the Essie Search Engine (Ide et al., 2007) which queried Medline with normalized strings from the 2009AB MRCONSO table in the UMLS. The frequency of a CUI was obtained by aggregating the frequency counts of the terms associated with the CUI to provide a rough estimate of its frequency. The information content measures then use this information to calculate the probability of a concept.

Another alternative is the use of *Intrinsic Information Content*. It assess the informativeness of concept based on its placement within a taxonomy by considering the number of incoming (ancestors) relative to outgoing (descendant) links (Sánchez et al., 2011) (Equation 12).

$$IC(c) = -log(\frac{\frac{|leaves(c)|}{|subsumers(c)|} + 1}{max\_leaves + 1}) \tag{12}$$

where $leaves$ are the number of descendants of concept c that are leaf nodes, $subsumers$ are the number of concept c's ancestors and max_leaves are the total number of leaf nodes in the taxonomy.

2.3 Relatedness Measures

Lesk (1986) observed that concepts that are related should share more words in their respective definitions than concepts that are less connected. He was able to perform word sense disambiguation by identifying the senses of words in a sentence with the largest number of overlaps between their definitions. An overlap is the longest sequence of one or more consecutive words that occur in both definitions. Banerjee and Pedersen (2003) extended this idea to WordNet, but observed that WordNet glosses are often very short, and did not contain enough information to distinguish between multiple concepts. Therefore, they created a *super–gloss* for each concept by adding the glosses of related concepts to the gloss of the concept itself (and then finding overlaps).

Patwardhan and Pedersen (2006) adapted this measure to second–order co–occurrence vectors. In this approach, a vector is created for each word in a concept's definition that shows which words co–occur with it in a corpus. These word vectors are averaged to create a single co-occurrence vector for the concept. The similarity between the concepts is calculated by taking the cosine between the concepts second–order vectors. Liu et al. (2012) modified and extended this measure to be used to quantify the relatedness between biomedical and clinical terms in the UMLS. The work in this paper can be seen as a further extension of Patwardhan and Pedersen (2006) and Liu et al. (2012).

3 Method

In this section, we describe our *second–order similarity vector* measure. This incorporates both contextual information using the term pair's definition and their pairwise semantic similarity scores derived from a taxonomy. There are two stages to our approach. First, a co–occurrence matrix must be constructed. Second, this matrix is used to construct a second–order co–occurrence vector for each concept in a pair of concepts to be measured for relatedness.

3.1 Co–occurrence Matrix Construction

We build an $m \times n$ similarity matrix using an external corpus where the rows and columns represent words within the corpus and the element contains the similarity score between the row word and column word using the similarity measures discussed above. If a word maps to more than one possible sense, we use the sense that returns the highest similarity score.

For this paper our external corpus was the NLM 2015 Medline baseline. Medline is a bibliographic database containing over 23 million citations to journal articles in the biomedical domain and is maintained by National Library of Medicine. The 2015 Medline Baseline encompasses approximately 5,600 journals starting from 1948 and contains 23,343,329 citations, of which 2,579,239 contain abstracts. In this work, we use Medline titles and abstracts from 1975 to present day. Prior to 1975, only 2% of the citations contained an abstract. We then calculate the similarity for each bigram in this dataset and include those that have a similarity score greater than a specified threshold on these experiments.

3.2 Measure Term Pairs for Relatedness

We obtain definitions for each of the two terms we wish to measure. Due to the sparsity and inconsistencies of the definitions in the UMLS, we not only use the definition of the term (CUI) but also include the definition of its related concepts. This follows the method proposed by Patwardhan and Pedersen (2006) for general English and WordNet, and which was adapted for the UMLS and the medical domain by Liu et al. (2012). In particular we add the definitions of any concepts connected via a parent (PAR), child (CHD), RB (broader than), RN (narrower than) or TERM (terms associated with CUI) relation. All of the definitions for a term are combined into a single *super–gloss*. At the end of this process we should have two super–glosses, one for each term to be measured for relatedness.

Next, we process each super–gloss as follows:

1. We extract a first–order co–occurrence vector for each term in the super–gloss from the co–occurrence matrix created previously.

2. We take the average of the first order co–occurrence vectors associated with the terms in a super–gloss and use that to represent the meaning of the term. This is a second–order co–occurrence vector.

3. After a second–order co–occurrence vector has been constructed for each term, then we calculate the cosine between these two vectors to measure the relatedness of the terms.

4 Data

We use two reference standards to evaluate the semantic similarity and relatedness measures [1]. UMNSRS was annotated for both similarity and relatedness by medical residents. MiniMayoSRS was annotated for relatedness by medical doctors (MD) and medical coders (coder). In this section, we describe these data sets and describe a few of their differences.

MiniMayoSRS: The MayoSRS, developed by 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. The relatedness of each term pair was assessed based on a four point scale: (4.0) practically synonymous, (3.0) related, (2.0) marginally related and (1.0) unrelated. MiniMayoSRS is a subset of the MayoSRS and consists of 30 term pairs on which a higher inter–annotator agreement was achieved. The average correlation between physicians is 0.68. The average correlation between medical coders is 0.78. We evaluate our method on the mean of the physician scores, and the mean of the coders scores in this subset in the same manner as reported by Pedersen et al. (2007).

UMNSRS: The University of Minnesota Semantic Relatedness Set (UMNSRS) was developed by Pakhomov et al. (2010), and 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. The similarity and relatedness of each term pair was annotated based on a continuous scale by having the resident touch a bar on a touch sensitive computer screen to indicate the degree of similarity or relatedness. The Intraclass Correlation Coefficient (ICC) for the reference standard tagged for similarity was 0.47, and 0.50 for relatedness. Therefore, as suggested by Pakhomov and colleagues,we use a subset of the ratings consisting of 401 pairs for the similarity set and 430 pairs for the relatedness set which each have an ICC of 0.73.

5 Experimental Framework

We conducted our experiments using the freely available open source software package UMLS::Similarity (McInnes et al., 2009) version 1.47[2]. This package takes as input two terms (or UMLS concepts) and returns their similarity or relatedness using the measures discussed in Section 2.

Correlation between the similarity measures and human judgments were estimated using Spearman's Rank Correlation (ρ). Spearman's measures the statistical dependence between two variables to assess how well the relationship between the rankings of the variables can be described using a monotonic function. We used Fisher's r-to-z transformation (Fisher, 1915) to calculate the significance between the correlation results.

[1] http://www.people.vcu.edu/ btmcinnes/downloads.html

[2] http://search.cpan.org/edist/UMLS-Similarity/

6 Results and Discussion

Table 1 shows the Spearman's Rank Correlation between the human scores from the four reference standards and the scores from the various measures of similarity introduced in Section 2. Each class of measure is followed by the scores obtained when integrating our second order vector approach with these measures of semantic similarity.

6.1 Results Comparison

The results for UMNSRS tagged for similarity (*sim*) and MiniMayoSRS tagged by coders show that all of the second-order similarity vector measures (*Integrated*) except for *vector-jcn* obtain a higher correlation than the original measures. We found that *vector-res* and *vector-faith* obtain the highest correlations of all these results with human judgments.

For the UMNSRS dataset tagged for relatedness and MiniMayoSRS tagged by physicians (MD), the original *vector* measure obtains a higher correlation than our measure (*Integrated*) although the difference is not statistically significant ($p \leq 0.2$).

In order to analyze and better understand these results, we filtered the bigram pairs used to create the initial similarity matrix based on the strength of their similarity using the *faith* and the *res* measures. Note that the *faith* measure holds to a 0 to 1 scale, while *res* ranges from 0 to an unspecified upper bound that is dependent on the size of the corpus from which information content is estimated. As such we use a different range of threshold values for each measure. We discuss the results of this filtering below.

6.2 Thresholding Experiments

Table 2 shows the results of applying the threshold parameter on each of the reference standards using the *res* measure. For example, a threshold of 0 indicates that all of the bigrams were included in the similarity matrix; and a threshold of 1 indicates that only the bigram pairs with a similarity score greater than one were included.

These results show that using a threshold cutoff of 2 obtains the highest correlation for the UMNSRS dataset, and that a threshold cutoff of 4 obtains the highest correlation for the MiniMayoSRS dataset. All of the results show an increase in correlation with human judgments when incorporating a threshold cutoff over all of the original

Table 1: Spearman's Correlation Results

	UMNSRS		MiniMayoSRS	
	Resident		MD	Coder
	sim	rel	relatedness	
Path				
path	0.52	0.28	0.35	0.45
wup	0.50	0.24	0.39	0.51
pks	0.49	0.25	0.38	0.50
zhong	0.50	0.25	0.42	0.50
Integrated				
vector-path	0.60	0.43	0.54	0.54
vector-wup	0.60	0.42	0.55	0.55
vector-pks	0.60	0.42	0.53	0.53
vector-zhong	0.58	0.41	0.54	0.53
Feature				
batet	0.16	0.33	0.16	0.15
cmatch	0.33	0.17	0.35	0.35
Integrated				
vector-batet	0.59	0.43	0.53	0.51
vector-cmatch	0.60	0.43	0.54	0.55
IC				
res	0.49	0.26	0.36	0.47
lin	0.51	0.29	0.44	0.54
jcn	0.52	0.33	0.42	0.52
faith	0.51	0.29	0.43	0.54
Integrated				
vector-res	0.57	0.41	0.58	**0.65**
vector-lin	0.57	0.41	0.59	0.64
vector-jcn	0.42	0.15	0.26	0.41
vector-faith	**0.59**	0.42	0.58	0.63
Intrinsic IC				
ires	0.49	0.26	0.40	0.50
ilin	0.50	0.28	0.41	0.50
ijcn	0.51	0.29	0.39	0.50
ifaith	0.50	0.28	0.41	0.50
Integrated				
vector-ires	0.57	0.41	0.50	0.52
vector-ilin	0.57	0.41	0.55	0.59
vector-ijcn	0.50	0.41	0.54	0.54
vector-ifaith	0.58	0.42	0.58	0.64
Relatedness				
lesk	0.49	0.33	0.52	0.56
o1vector	0.47	0.36	0.43	0.54
o2vector	0.54	**0.45**	**0.63**	0.59

Table 2: Threshold Correlation with *vector-res*

		UMNSRS		MiniMayoSRS	
T	# bigrams	sim	rel	MD	coder
0	850,959	0.58	0.41	0.58	0.65
1	166,003	0.56	0.39	0.60	0.67
2	65,502	**0.64**	**0.47**	0.56	0.62
3	27,744	0.60	0.46	0.62	0.71
4	10,991	0.56	0.43	**0.75**	**0.76**
5	3,305	0.26	0.16	0.36	0.36

Table 3: Threshold Correlation with *vector-faith*

	#	UMNSRS		MiniMayoSRS	
T	bigrams	sim	rel	MD	coder
0	838,353	0.59	0.42	**0.58**	**0.63**
0.1	197,189	0.58	0.41	0.57	**0.63**
0.2	121,839	0.58	0.41	**0.58**	**0.63**
0.3	71,353	0.63	0.46	0.54	0.55
0.4	45,335	0.64	0.48	0.50	0.51
0.5	29,734	**0.66**	**0.49**	0.49	0.53
0.6	19,347	0.65	0.49	0.52	0.56
0.7	11,946	0.64	0.48	0.53	0.55
0.8	7,349	0.64	0.49	0.53	0.56
0.9	4,731	0.62	0.49	0.53	0.57

measures. The increase in the correlation for the UMNSRS tagged for similarity is statistically significant ($p \leq 0.05$), however this is not the case for the UMNSRS tagged for relatedness nor for the MiniMayoSRS data.

Similarly, Table 3 shows the results of applying the threshold parameter (T) on each of the reference standards using the $faith$ measure. Although, unlike res whose scores are greater than or equal to 0 without an upper limit, the $faith$ measure returns scores between 0 and 1 (inclusive). Therefore, here a threshold of 0 indicates that all of the bigrams were included in the similarity matrix; and a threshold of 0.1 indicates that only the bigram pairs with a similarity score greater than 0.1 were included. The results show an increase in accuracy for all of the datasets except for the MiniMayoSRS tagged for physicians. The increase in the results for the UMNSRS tagged for similarity and the MayoSRS is statistically significant ($p \leq 0.05$). This is not the case for the UMNSRS tagged for relatedness nor the MiniMayoSRS.

Overall, these results indicate that including only those bigrams that have a sufficiently high similarity score increases the correlation results with human judgments, but what quantifies as sufficiently high varies depending on the dataset and measure.

6.3 Comparison with Previous Work

Recently, word embeddings (Mikolov et al., 2013) have become a popular method for measuring semantic relatedness in the biomedical domain. This is a neural network based approach that learns a representation of a word by word co–occurrence matrix. The basic idea is that the neural network learns a series of weights (the hidden layer within the neural network) that either maximizes the probability of a word given its context, referred to as the continuous bag of words (CBOW) approach, or that maximizes the probability of the context given a word, referred to as the Skip–gram approach. These approaches have been used in numerous recent papers.

Muneeb et al. (2015) trained both the Skip–gram and CBOW models over the PubMed Central Open Access (PMC) corpus of approximately 1.25 million articles. They evaluated the models on a subset of the UMNSRS data, removing word pairs that did not occur in their training corpus more than ten times. Chiu et al. (2016) evaluated both the the Skip–gram and CBOW models over the PMC corpus and PubMed. They also evaluated the models on a subset of the UMNSRS ignoring those words that did not appear in their training corpus. Pakhomov et al. (2016) trained CBOW model over three different types of corpora: clinical (clinical notes from the Fairview Health System), biomedical (PMC corpus), and general English (Wikipedia). They evaluated their method using a subset of the UMNSRS restricting to single word term pairs and removing those not found within their training corpus. Sajadi et al. (2015) trained the Skip–gram model over CUIs identified by MetaMap on the OHSUMED corpus, a collection of 348,566 biomedical research articles. They evaluated the method on the complete UMNSRS, MiniMayoSRS and the MayoSRS datasets; any subset information about the dataset was not explicitly stated therefore we believe a direct comparison may be possible.

In addition, a previous work very closely related to ours is a retrofitting vector method proposed by Yu et al. (2016) that incorporates ontological information into a vector representation by includ-

Table 4: Comparison with Previous Work

Method	UMNSRS				MayoSRS	MiniMayoSRS		
	Subsets		Full		(N=101)	(N=29)		
	sim	rel	sim (N=566)	rel (N=587)	rel	MD	coder	avg
vector–res (ours)	0.64 (N=401)	0.49 (N=430)	0.59	0.48	0.51	0.75	0.76	0.76
vector–faith (ours)	0.66 (N=401)	0.49 (N=430)	0.61	0.49	0.46	0.58	0.63	0.63
(Yu et al., 2016)						0.70	0.67	
(Sajadi et al., 2015)			0.39	0.39	0.63			0.8
(Pakhomov et al., 2016)	0.62 (N=449)	0.58 (N=458)						
(Muneeb et al., 2015)	0.52 (N=462)	0.45 (N=465)						
(Chiu et al., 2016)	0.65 (N=UK)	0.60 (N=UK)						

ing semantically related words. In their measure, they first map a biomedical term to MeSH terms, and second build a word vector based on the documents assigned to the respective MeSH term. They then retrofit the vector by including semantically related words found in the Unified Medical Language System. They evaluate their method on the MiniMayoSRS dataset.

Table 4 shows a comparison to the top correlation scores reported by each of these works on the respective datasets (or subsets) they evaluated their methods on. N refers to the number of term pairs in the dataset the authors report they evaluated their method. The table also includes our top scoring results: the *integrated vector-res* and *vector-faith*. The results show that integrating semantic similarity measures into second–order co–occurrence vectors obtains a higher or on–par correlation with human judgments as the previous works reported results with the exception of the UMNSRS rel dataset. The results reported by Pakhomov et al. (2016) and Chiu et al. (2016) obtain a higher correlation although the results can not be directly compared because both works used different subsets of the term pairs from the UMNSRS dataset.

7 Conclusion and Future Work

We have presented a method for quantifying the similarity and relatedness between two terms that integrates pair–wise similarity scores into second–order vectors. The goal of this approach is two–fold. First, we restrict the context used by the vector measure to words that exist in the biomedical domain, and second, we apply larger weights to those word pairs that are more similar to each other. Our hypothesis was that this combination would reduce the amount of noise in the vectors and therefore increase their correlation with human judgments. We evaluated our method on datasets that have been manually annotated for relatedness and similarity and found evidence to support this hypothesis. In particular we discovered that guiding the creation of a second–order context vector by selecting term pairs from biomedical text based on their semantic similarity led to improved levels of correlation with human judgment.

We also explored using a threshold cutoff to include only those term pairs that obtained a sufficiently large level of similarity. We found that eliminating less similar pairs improved the overall results (to a point). In the future, we plan to explore metrics to automatically determine the threshold cutoff appropriate for a given dataset and measure. We also plan to explore additional features that can be integrated with a second–order vector measure that will reduce the noise but still provide sufficient information to quantify relatedness. We are particularly interested in approaches that learn word, phrase, and sentence embeddings from structured corpora such as literature (Hill et al., 2016a) and dictionary entries (Hill et al., 2016b). Such embeddings could be integrated into a second–order vector or be used on their own.

Finally, we compared our proposed method to other distributional approaches, focusing on those that used word embeddings. Our results showed that integrating semantic similarity measures into second–order co–occurrence vectors obtains the same or higher correlation with human judgments as do various different word embedding approaches. However, a direct comparison was not possible due to variations in the subsets of the UMNSRS evaluation dataset used. In the future, we would not only like to conduct a direct comparison but also explore integrating semantic similarity into various kinds of word embeddings by training on pair–wise values of semantic similarity as well as co–occurrence statistics.

References

S. Banerjee and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. Acapulco, Mexico, pages 805–810.

M. Batet, D. Sánchez, and A. Valls. 2011. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of biomedical informatics* 44(1):118–125.

O. Bodenreider and A. Burgun. 2004. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. In *Proceedings of the 11th World Congress on Medical Informatics (MEDINFO)*. San Fransico, CA, pages 327–331.

J.E. Caviedes and J.J. Cimino. 2004. Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics* 37(2):77–85.

B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo. 2016. How to Train Good Word Embeddings for Biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. pages 166–174.

R.A. Fisher. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* pages 507–521.

Z. Harris. 1954. Distributional structure. *Word* 10(23):146–162.

F. Hill, K. Cho, and A. Korhonen. 2016a. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 1367–1377.

F. Hill, K. Cho, A. Korhonen, and Y. Bengio. 2016b. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics* 4:17–30.

N.C. Ide, R.F. Loane, and D. Demner-Fushman. 2007. Essie: a concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association* 14(3):253–263.

A. Islam and D. Inkpen. 2006. Second order co-occurrence pmi for determining the semantic similarity of words. In *Proceedings of the International Conference on Language Resources and Evaluation, Genoa, Italy*. pages 1033–1038.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*. Tapei, Taiwan, pages 19–33.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*. Toronto, Canada, pages 24–26.

D. Lin. 1998. An information-theoretic definition of similarity. In *Intl Conf ML Proc.*. San Francisco, CA, pages 296–304.

D. Lin and P. Pantel. 2002. Concept discovery from text. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-02)*. Taipei, Taiwan, pages 577–583.

Y. Lin, W. Li, K. Chen, and Y. Liu. 2007. A document clustering and ranking system for exploring MEDLINE citations. *Journal of the American Medical Informatics Association* 14(5):651–661.

Y. Liu, B.T. McInnes, T. Pedersen, G. Melton-Meaux, and S. Pakhomov. 2012. Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. In *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics*. ACM, pages 363–372.

A. Maedche and S. Staab. 2001. *Comparing ontologies-similarity measures and a comparison study*. AIFB.

B.T. McInnes, T. Pedersen, and S.V. Pakhomov. 2009. UMLS-Interface and UMLS-Similarity: Open source software for measuring paths and semantic similarity. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*. San Fransico, CA, pages 431–435.

T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

T.H. Muneeb, Sunil Sahu, and Ashish Anand. 2015. Evaluating distributed word representations for capturing semantics of biomedical concepts. In *Proceedings of BioNLP 15*. Association for Computational Linguistics, Beijing, China, pages 158–163.

S. Pakhomov, B.T. McInnes, T. Adam, Y. Liu, T. Pedersen, and G.B. Melton. 2010. Semantic similarity and relatedness between clinical terms: An experimental study. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*. Washington, DC, pages 572–576.

S.V. Pakhomov, G. Finley, R. McEwan, Y. Wang, and G.B. Melton. 2016. Corpus domain effects on distributional semantic modeling of medical terms — Bioinformatics — Oxford Academic. *Bioinformatics* 32:3635–3644.

S.V.S. Pakhomov, T. Pedersen, B. McInnes, G.B. Melton, A. Ruggieri, and C.G. Chute. 2011. Towards a framework for developing semantic relatedness reference standards. *Journal of Biomedical Informatics* 44(2):251–265.

S. Patwardhan and T. Pedersen. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*. Trento, Italy, pages 1–8.

T. Pedersen, S.V.S. Pakhomov, S. Patwardhan, and C.G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics* 40(3):288–299.

V. Pekar and S. Staab. 2002. Taxonomy learning: Factoring the structure of a taxonomy into a semantic classification decision. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '02, pages 1–7.

G. Pirró and J. Euzenat. 2010. A feature and information theoretic framework for semantic similarity and relatedness. In *The Semantic Web–ISWC 2010*, Springer, pages 615–630.

R. Rada, H. Mili, E. Bicknell, and M. Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics* 19(1):17–30.

K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*. ACM, pages 337–346.

J. Reisinger and R.J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 109–117.

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Montreal, Canada, pages 448–453.

A. Sajadi, E.E. Milios, V. Kešelj, and J.C.M. Janssen. 2015. Domain-specific semantic relatedness from Wikipedia structure: A case study in biomedical text. In *Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2015))*. Cairo, Egypt, pages 347–360.

D. Sánchez, M. Batet, and D. Isern. 2011. Ontology-based information content computation. *Knowledge-Based Systems* 24(2):297–303.

H. Schütze. 1992. Dimensions of meaning. In *Proceedings of the ACM/IEEE Conference on Supercomputing*. Minneapolis, MN, pages 787–796.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1):97–123.

J. Weeds, D. Weir, and D. McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, page 1015.

Z. Wu and M. Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Meeting of Association of Computational Linguistics*. Las Cruces, NM, pages 133–138.

W. Yih and V. Qazvinian. 2012. Measuring word relatedness using heterogeneous vector space models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 616–620.

Z. Yu, T. Cohen, B. Wallace, E. Bernstam, and T. Johnson. 2016. Retrofitting word vectors of mesh terms to improve semantic similarity measures. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics, Auxtin, TX, pages 43–51.

J. Zhong, H. Zhu, J. Li, and Y. Yu. 2002. Conceptual graph matching for semantic search. In *Proceedings of the 10th International Conference on Conceptual Structures*. pages 92–106.