Journal of Biomedical Informatics 90 (2019) 103096

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Unsupervised low-dimensional vector representations for words, phrases and text that are transparent, scalable, and produce similarity metrics that are not redundant with neural embeddings

Neil R. Smalheiser[a,*], Aaron M. Cohen[b], Gary Bonifield[a]

[a] Department of Psychiatry and Psychiatric Institute, University of Illinois College of Medicine, Chicago, IL 60612, USA
[b] Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239, USA

ARTICLE INFO

Keywords:
Word2vec
Pvtopic
Vector representation
Dimensional reduction
Text mining
Semantic similarity
Natural language processing
Implicit features

ABSTRACT

Neural embeddings are a popular set of methods for representing words, phrases or text as a low dimensional vector (typically 50–500 dimensions). However, it is difficult to interpret these dimensions in a meaningful manner, and creating neural embeddings requires extensive training and tuning of multiple parameters and hyperparameters. We present here a simple unsupervised method for representing words, phrases or text as a low dimensional vector, in which the meaning and relative importance of dimensions is transparent to inspection. We have created a near-comprehensive vector representation of words, and selected bigrams, trigrams and abbreviations, using the set of titles and abstracts in PubMed as a corpus. This vector is used to create several novel implicit word-word and text-text similarity metrics. The implicit word-word similarity metrics correlate well with human judgement of word pair similarity and relatedness, and outperform or equal all other reported methods on a variety of biomedical benchmarks, including several implementations of neural embeddings trained on PubMed corpora. Our implicit word-word metrics capture different aspects of word-word relatedness than word2vec-based metrics and are only partially correlated (rho = 0.5–0.8 depending on task and corpus).

The vector representations of words, bigrams, trigrams, abbreviations, and PubMed title + abstracts are all publicly available from http://arrowsmith.psych.uic.edu/arrowsmith_uic/word_similarity_metrics.html for release under CC-BY-NC license. Several public web query interfaces are also available at the same site, including one which allows the user to specify a given word and view its most closely related terms according to direct co-occurrence as well as different implicit similarity metrics.

1. Introduction

A recent lawsuit in the United Kingdom involved a patient suing a hospital over access to his health records. The patient felt that he had been subjected to unnecessary surgery, and sought access to all records related to his care. It developed that the hospital had supplied him with only a subset of the relevant information, relying on a technical distinction between the words *notes* and *records*. It is a good example of what the forensic linguist John Olsson has identified as a common tactic of scuzzy businesses: relying on technical distinctions between similar words to do something or other. But, what does it mean for two words to be "similar"? It seems obvious, but this turns out to be a vexing question–and one whose answer has implications for many fields of inquiry, ranging from psycholinguistics (where it is studied in the context of exploring hypotheses about how our brain recognizes words) to biomedical ontology (where it has implications for doing inference over the output of high-throughput assays) to natural language processing and text mining.

One reason that word similarity is difficult to study is that two words can be similar to each other in many different ways. We can divide these ways of being "similar" into a number of broad categories, ranging from relationships of abstract meaning (e.g. *egg* and *ovum*) based in large part on human intuitions, to purely statistical relationships derived from large sets of linguistic data (e.g. the word *gene* tends to be joined by the word *and* to the words *locus* and *region* more frequently than would be expected by chance, while the word *protein* is more likely to be joined with the word *enzyme*). Two words may be related, such as "pork" and "beans". Or they may co-occur in a phrase, such as "Cold Spring". Two words may also be related insofar as they can substitute for each other within a given utterance, as the word "cat" in "The cat is on the mat" can be substituted by a wide number of other entities that can sit or stand on a mat. Furthermore, two words may be

* Corresponding author.
E-mail address: neils@uic.edu (N.R. Smalheiser).

https://doi.org/10.1016/j.jbi.2019.103096
Received 8 January 2018; Received in revised form 27 November 2018; Accepted 31 December 2018
Available online 14 January 2019

implicitly related if they share some relation with a third entity; for example, “acupuncture” and “morphine” are both treatments for pain. In the present study, we describe and characterize several similarity metrics for relating words, terms, or passages of text to each other. We define direct similarity of two terms as measuring how often they co-occur within the same text, relative to the frequency expected by chance. Implicit term or text similarity metrics measure the degree to which two terms or texts share similar vector representations (see below). Both direct and implicit types of similarity have their place, depending on the intended purpose, and indeed, having a palette of text-related similarity metrics should be a boon for text mining models.

Many previous investigations have examined different metrics for estimating the similarity of two words or phrases, concepts, as well as sentences and documents. To give just a few examples, two words can be related according to their path distance on graphs based on network features, ontologies, or Wordnet [1]. Alternatively, they can related by direct measures of text similarity such as string matching [2], or by implicit measures such as the number of shared words in their dictionary definitions [3]. Two articles or documents can be judged for similarity of their text words, their metadata (e.g., number of shared MeSH terms, MeSH term pairs [4] or UMLS concepts [5]), or implicit information such as similarity in their cited or citing articles [6]. There are various unsupervised schemes for representing documents in terms of underlying word similarities (among them context based methods [7,8] such as latent semantic indexing [7], as well as PubMed Related articles [9] and neural vector spaces [10]), which can be used to create word-word or topic-based document-document similarity metrics that rank articles for relatedness (see also [11,12]).

During the past five years, neural embedding methods such as word2vec [13] and GloVE [14] have been investigated widely to create low-dimensional vector representations of words and text passages. In such schemes, the implicit similarity of any two words or texts is given by the cosine similarity of their vectors [e.g., 10,15]. Despite the popularity of this approach, there are several issues that should be acknowledged. First, the most popular implementations use a sliding window that relates a central word to its nearby contextual words, and this tends to emphasize word substitutionality. Second, the vectors vary according to the type and size of the textual corpus used to train the models; typically, a very large corpus of text is used that may not be the same as the text to which the modeling is being applied. Third, there are many different variations and implementations of neural embeddings; not only are there many parameters and hyperparameters that need to be tuned to optimize any particular modeling task, but the optimal tuning will vary from one specific task to another. Fourth, typically a word is represented by a vector having 50–500 dimensions, whose values are assigned by a neural network operating as a black box. The interpretation and relative importance of each dimension are not transparent to interpretation, at least not without extensive further training and processing [16,17].

We therefore decided to create an alternative type of similarity metric that would deal with the limitations of neural embeddings. Specifically, we created a simple unsupervised method for representing words, phrases or text as a low dimensional vector, in which the meaning and relative importance of dimensions is transparent to inspection. The approach is general, although we have employed PubMed titles and abstracts as a corpus in order to apply the metrics to biomedical tasks. Our implicit term metrics give superior performance on a panel of biomedical term similarity and relatedness benchmarks. We also investigated how our metrics compare with several direct similarity metrics and word2vec-derived metrics for computing similarity of PubMed text passages (title + abstracts).

The present studies are part of an ongoing project to create a series of implicit similarity metrics to aid in text mining and modeling of the biomedical literature. For example, recently we created journal similarity metrics that relate any two journals A and B according to a) how similar they are topically, b) how similar they are in terms of sharing the same authors, and c) how likely it is that an author who publishes in A will publish later in B [18]. We also created MeSH (Medical Subject Headings) similarity metrics that relate any two MeSH terms A and B according to a) how often the two terms co-occur in a single article, and b) how often the two terms are found within the body of articles published by a single author [19]. Such implicit metrics are valuable in text mining models such as the Author-ity author name disambiguation project [20,21], where the goal is to consider two articles that share the same author [lastname, first initial] and estimate the probability that the two articles refer to the same author-individual. A match on (say) journal name is a powerful direct feature suggesting that the two articles may share the same author, but this feature can only be scored as nonmatch (=0) vs. match (=1). In contrast, two nonmatching journal names can be scored using implicit similarity metrics as a positive number between 0 and 1 in all cases. Thus, implicit features have a “smoothing” effect, improving robustness and sensitivity of sparse models.

2. Materials and methods

2.1. Terms

A database of all PubMed articles published 1966–2016, in English (or with titles and abstracts translated into English), was created. Titles and abstracts were tokenized as in [22] and stoplisted using the PubMed 365-word stoplist [23]. Words were included only if they appeared in at least 100 titles and in at least 25 abstracts.

The **basic word model** consists of 44,201 words that met these criteria and includes 41,918,680 word pairs that co-occurred in the title or abstract of the same PubMed article. For the **full model**, we included words as above, plus selected bigrams, trigrams, and abbreviations as follows: 1. Bigrams were formed from tokenized words that appear in the basic model. 2. The two words making up the bigram must have a direct odds ratio of 10 or greater. 3. Trigrams are formed from bigrams where trigram ABC is derived from bigrams AB and BC. Trigrams must appear in 1000 or more abstracts, and the trigram ABC must have a document frequency at least 30% of the minimum document frequencies of AB and BC.

Note that when computing multi-word phrases, instances of the individual words that make up these bigrams or trigrams were removed from further consideration, and were not included when computing direct and implicit word metrics of the individual words.

To select abbreviations for encoding by their own vectors: 1. We only considered instances of abbreviations that are written in all-capitals and listed in at least 25 articles within the ADAM abbreviation database [24,25]. 2. The frequency of the non-capitalized word must be greater than the frequency of the all-capitalized word. Again, when an abbreviation was selected, instances of the abbreviation were excluded when computing the same token considered as a single non-capitalized word. For example, ACT is an abbreviation whereas act is a non-capitalized word; instances of ACT were not included when computing the vector for act.)

The **full term model** is made up of 97,754 terms (43,494 words, 49,918 bigrams, 3,638 trigrams, and 704 abbreviations) and includes 92,757,119 co-occurring term pairs.

2.2. The direct odds ratio, a measure of direct term similarity

The Direct Odds Ratio measures how often two words are observed to co-occur in the same article, relative to the value that would be expected by chance (i.e., if each word occurred at random throughout the corpus of PubMed articles and independently of each other). The two terms may co-occur in either order, anywhere in the same article, and need not be within the same sentence or to within a proximity window of defined size. Two very common terms might be expected to co-occur often just by chance, whereas it will be highly significant if one

observes any co-occurrence of two very rarely occurring terms.

To compute the direct odds ratio for a given term pair A:B, one computes the observed co-occurrence divided by the expected co-occurrence if they appear independent of each other. The expected co-occurrence = (count(A) * count(B))/N where N is the total number of articles in MEDLINE. However, for most term pairs, the expected co-occurrence is quite low (< 1), and so the observed/expected ratio is greatly affected by small fluctuations in the observed co-occurrence count. To improve robustness, we used a bin method to compute the expected co-occurrence across all term pairs of roughly the same size. a) We first calculated, for each term pair, the geomean of their individual document occurrences count(A) and count(B):

$$\text{geomean}(A{:}\ B) = \sqrt{(\text{count}(A) * \text{count}(B))}$$

b) The set of all term pairs were ranked by geomean, and this set was divided into bins of 500 pairs (i.e., each bin consisting of pairs that have roughly the same geomean). For each bin, we calculated the co-occurrence count expected by chance, by taking the average of the co-occurrence counts co-occur(A:B) across all pairs in the same bin.

c) Finally, we calculated the direct odds ratio for each pair A:B present in that bin, by taking its observed A: B co-occurrence score divided by the average co-occurrence score for that bin:

$$\text{Direct odds ratio}(A{:}\ B) = \text{co-occur}(A{:}\ B)/(\text{mean co-occur}(\text{bin}))$$

This is similar to the manner in which odds ratios were computed for journal similarity metrics in D'Souza and Smalheiser [18] and for MeSH term similarity metrics in Smalheiser and Bonifield (2016) [1,2].

Note that the direct odds ratio is not only employed to create vector representations of words and terms (as discussed below), but also provides a **direct similarity metric**. That is, two words or terms show greater direct similarity if they tend to co-occur in the same articles more than would be expected by chance.

2.3. Computing vector representations for terms

For each term selected as above, we made a list of all other selected terms that co-occurred with it in titles and abstracts of MEDLINE articles, and ordered the list according to their direct odds ratios. The co-occurring terms with the highest direct odds ratios were used to form a vector. That is, for a given term, we took its co-occurring term having the highest direct odds ratio and placed it in dimension 1; took the term having the next highest direct odds ratio and placed it in dimension 2; and so on, until 300 dimensions were assigned or until the direct odds ratio fell below 1.25. **Thus, the vector representation of a word or term consists of a ranked list of its co-occurring context terms.**

We chose the cut-off criteria (300 dimensions or direct odds ratio of 1.25) because of previous research suggesting that 300 dimensions gives asymptotically optimal performance in both neural embedding models [e.g., 20] and latent semantic indexing [7,12]. We also felt that including terms having a direct odds ratio of less than 1.25 would be too close to random co-occurrence, hence might subject the vector to too much "noise". The vast majority of single words (99.9%) in the basic model, and 97% of terms in the full model, had at least 300 co-occurring terms whose direct odds ratio was 1.25 or greater. (The few exceptions were mainly multi-word terms of very low frequency.)

2.4. Computing vector representations of text passages

For each PubMed article, the title + abstract was concatenated to form a single text passage, and tokenized and processed to recognize words, bigrams, trigrams, abbreviations from our dataset. Each term was only counted once, i.e., there was no double-counting for the title or for multiple instances in the abstract. For each term found in the title + abstract, its 300-dimensional vector representation was listed. Then, we created a master list of all terms that occur across all these vectors. Note that any given term may occur in multiple vectors, and it may be associated with a different direct odds ratio in each of the vectors that it appears in. To create a single new vector out of the set of these terms, we assigned each term a value which is the sum of the $\log_e$ direct odds ratios of that term in each vector it appears in. Finally, we represented the PubMed title + abstract as a 300-dimensional vector as follows: The term having the highest value gets rank 1, next highest is rank 2, etc. down to 300 dimensions.

2.5. Creating implicit similarity metrics for terms

a) **Implicit unweighted shared term score**. For any two terms in our dataset, we examined their 300-dimensional vectors, and counted the number of words (or terms) shared in both vectors. The unweighted similarity is thus an integer ranging from 0 to 300.

b) **Implicit weighted score for the basic (single word) model**. For two words or terms A and B, we examined their 300-dimensional vectors, and list the words shared in both vectors. We then created a weighted sum as follows: For each shared vector word, choose the LESSER of the odds ratio in the two vectors. Then, the implicit weighted score = the sum of the $\log_e$ direct odds ratios across all shared words.

c) **Implicit weighted score for the full (term) model**. We give greater weight for shared trigrams and bigrams relative to shared words or abbreviations. Thus, we create a weighted sum as follows: For each shared vector term, choose the LESSER of the direct odds ratio in the two vectors. Then, the implicit weighted score = the sum of (w_i* $\log_e$ direct odds ratios) across all shared words, where w_i = 1 for words and abbreviations, 2 for bigrams, 3 for trigrams.

Note that our implicit term metrics use vectors whose entries consist of terms, not numbers, so cosine similarity is not an appropriate method to compare two of these vectors.

2.6. Computing implicit similarity metrics for text passages

For two articles (title + abstract) or other text passages, we examine their 300-dimensional vector representations, and listed the words (or terms) shared in both vectors. The **implicit shared terms** metric is simply the number of shared terms, i.e., an integer between 0 and 300. The **implicit weighted score** is computed as follows: For each shared vector term, choose the LESSER of the direct odds ratio in the two vectors. Then, the implicit weighted score = the sum of (w_i* $\log_e$ direct odds ratios) across all shared terms, where w_i = 1 for words and abbreviations, 2 for bigrams, 3 for trigrams.

2.7. Word2vec based similarity metrics

For computing word2vec term similarity, we used the University of Turku word2vec 200-dimensional vectors computed for biomedical words that was trained on PubMed titles and abstracts downloaded from http://evexdb.org/pmresources/vec-space-models/. To represent phrases, some bigrams were already represented as single entities in the Finnish word2vec corpus. However, we verified that the performance of word2vec on the 30 term relatedness benchmark (Supplementary File 1) was improved by summing individual word vectors for multi-word phrases [17,26], thus summing was performed on the other benchmarks as well. This vector representation is referred to here as "**word2vec**". To obtain the word2vec similarity of two words or phrases, the cosine similarity of their word2vec vectors was computed, giving a real number between −1 and +1.

To represent PubMed articles, word2vec 300-dimensional vectors were computed using the paragraph2vec code https://github.com/hassyGo/paragraph-vector using the following parameters: -wvdim 300 -pvdim 300 -itr 10 [26,27]. Training was performed across the entire PubMed corpus of article titles and abstracts (1966–2016). This training procedure created a separate word2vec representation of biomedical words which will be referred to as "**pvtopic**" since it follows the word2vec vector representation described in Hashimoto et al. [27].

Although Word2vec and pvtopic are both implementations of neural embeddings based on a PubMed corpus, word2vec gave better performance on the 30 term relatedness benchmark (see Supplementary File 1), so word2vec was used for term similarity and relatedness comparisons, whereas pvtopic was used for representing PubMed articles in order to follow the method of Hashimoto et al. [27]. The paragraph2vec code computed 300-dimensional vector representations for all PubMed articles; each article was represented by concatenating its title with its abstract to form one text. To obtain the similarity of two text passages, the cosine similarity of their vectors were computed. To ensure robustness, similarity scores were only computed for articles that contained abstracts and whose title + abstract contained at least 25 words.

3. Results

3.1. Comparison of metrics for term similarity of selected biomedical words

The **direct similarity** of two words or terms measures how often the two co-occur in the same article, relative to the co-occurrence that would be expected simply from the document frequencies of the two considered independently. For example, “peanut” and “butter” co-occur often (because of the phrase “peanut butter”) and thus show high direct similarity, i.e., a high direct odds ratio. In contrast, the **implicit similarity** of two words or terms measures how well they share the same context words. For example, “randomization” and “randomisation” have very low direct similarity since they rarely co-occur in the same article (one is American spelling and one is British spelling), but they share many of the same surrounding context words (e.g., “clinical trial” and “RCT”) and so have very high implicit similarity. Finally, the word2vec metrics measure the substitutability of one word or term for another within passages of text. For example, in the sentence “I took the train from Rome to Florence”, the word “I” could be readily substituted by another personal name (e.g., “he”, “she”, “Sheldon”, etc.), the word “train” could be substituted by other modes of transportation (e.g., “bus” or “bike”), and “Florence” could be substituted by the name of another city that can be reached by train (but less likely by a distant city such as Chicago).

To give an example of how the different metrics rank related words most highly in the PubMed corpus, Table 1 shows the top 10 most related words to “tennis” by direct odds ratio, by implicit weighted score, and by word2vec. The direct odds ratio highlights words such as “player”, “elite”, “elbow”, and “epicondylitis” which are described in articles that study tennis players. In contrast, 9 of the top 10 words by the word2vec metric are names of other sports, such as basketball and baseball. The implicit weighted score falls in between the other two metrics, sharing 3 words with the direct odds ratio and 5 words with word2vec. Only a single word, “soccer”, is shared in the top ten by all three metrics.

To give another illustrative example, Table 2 shows the top 10 most related words to “Italy”. The top 10 words by direct odds ratio consists entirely of the names of cities and regions within Italy. In contrast, the top 10 words according to the word2vec metric consist entirely of the names of other European countries. Again, the implicit weighted score falls in between the other two, as it highlights words that are used often in the same context as the word Italy, including words such as “Italian”, “Europe”, “European” and “Mediterranean”, as well as the names of five other European countries. These examples confirm our expectation that the direct, implicit and word2vec metrics are capturing different types of similarity, at least in part.

Table 1
Top 10 words most related to “Tennis” according to three metrics.

Rank	Direct odds ratio	Implicit weighted score	Word2vec
1	Player	Soccer	Basketball
2	Ball	Athletic	Volleyball
3	Elbow	Basketball	Soccer
4	Epicondylitis	Athlete	Badminton
5	Sport	Sport	Handball
6	Athlete	Throwing	Baseball
7	Elite	Volleyball	Football
8	Tournament	Baseball	Rugby
9	Soccer	Football	Softball
10	Golf	Collegiate	Amateur

Table 2
Top 10 words most related to “Italy” according to three metrics.

Rank	Direct odds ratio	Implicit weighted score	Word2vec
1	Emilia	Spain	France
2	Tuscany	Europe	Spain
3	Veneto	Italian	Greece
4	Sardinia	France	Portugal
5	Campania	Greece	Romania
6	Lombardy	Portugal	Belgium
7	Romagna	European	Germany
8	Sicily	Poland	Turkey
9	Turin	Inhabitant	Switzerland
10	Milan	Mediterranean	Sweden

3.2. Evaluation on term relatedness and similarity benchmarks

Several prominent biomedical term relatedness and term similarity benchmarks were downloaded from http://rxinformatics.umn.edu/SemanticRelatednessResources.htm which are shown and described in Supplementary files 1–5.

3.2.1. Performance of metrics on a 30 term biomedical term relatedness benchmark

First, we compared how well various word similarity metrics align with the mean judgments of three physician and nine medical coders of term relatedness (rated on a scale of 1 to 4) in the Pedersen 30 word pair benchmark [5]. These term pairs are a subset of the larger Pakhomov 101 term pair dataset (see below), chosen because they showed high inter-rater reliability.

The 30 word benchmark, physician and coder relatedness scores, and similarity scores computed by the various metrics are shown in Supplementary File 1. As shown in Table 3, the implicit shared terms (both weighted and unweighted, computed from the full term model) gave high alignment with human judgments of term similarity, as assessed by Spearman rank correlation. The correlations were extremely high (0.86 for physicians, 0.81 for coders), which to our knowledge exceeds that previously reported for any term similarity metrics tested where the pair of terms is the sole input to the metric (i.e., excluding approaches which employ supplementary outside information from knowledgebases) [e.g., 5,10,28,29]. The direct odds ratio also gave high Spearman correlations (0.82–0.84). Of several different ways examined to compute the word2vec and pvtopic similarities (see Supplementary File 1), the best performance obtained for word2vec was 0.79 for physicians, 0.80 for coders, and the best Spearman rank correlation of pvtopic was 0.68 for physicians, 0.69 for coders. These values are comparable to or better than the performance of other word2vec-based metrics reported by others on this benchmark.

3.2.2. Benchmark of 101 biomedical term pairs rated for relatedness

Because the 30 word benchmark is relatively small, we further evaluated our metrics on the larger set of 101 term pairs that were manually rated for relatedness on a scale of 1 to 10 by 13 medical coders, consisting mostly of disease names and signs and symptoms of disease (Pakhomov et al. [30]). See Supplementary File 2 for the dataset (note that one term pair failed to map to our dataset, giving 100 term pairs).

Table 3
Spearman rank correlations among human relatedness scores and similarity metrics for the 30 biomedical term relatedness benchmark.

	Physician	Coder	Direct odds ratio	Imp shared	Imp weighted	Best word2vec	Best pvtopic
Physician	1.0000	0.8886	0.8425	0.8611	0.8597	0.7926	0.6759
Coder		1.0000	0.8247	0.8103	0.8127	0.8008	0.6924
Direct odds ratio			1.0000	0.9133	0.9047	0.7516	0.8282
Implicit shared terms				1.0000	0.9781	0.7644	0.8650
Implicit weighted score					1.0000	0.7749	0.8630
Best word2vec						1.0000	0.6605
Best pvtopic							1.0000

Implicit shared terms and implicit weighted score are computed from the full term model (that includes words as well as selected bigrams, trigrams, and abbreviations). Shown are the versions of word2vec and pvtopic that gave the best performance among those tested, namely, summing vectors to represent phrases, and computing similarity on exact words used in the benchmarks, rather than on the terms that mapped to our dataset (Supplementary File 1).

Table 4
Spearman rank correlations among human relatedness scores and similarity metrics for the 101 biomedical term relatedness benchmark.

	Mean	Direct odds	Implicit shared	Implicit weighted	Word2vec
Mean relatedness judgment	1.0000	0.7582	0.7426	0.7354	0.4968
Direct odds ratio		1.0000	0.8673	0.8703	0.5539
Implicit shared terms			1.0000	0.9879	0.6823
Implicit weighted score				1.0000	0.6837
Word2vec					1.0000

The performance of our metrics is shown in Table 4. The Spearman rank correlations with coder relatedness judgment were similar for the direct odds ratio (0.76) and the two implicit metrics (implicit shared terms (0.74) and implicit weighted score (0.735). In contrast, word2vec showed a much lower correlation with human judgment (0.496) and this difference was statistically significant compared to the implicit weighted score ($p = 0.027$, unpaired, 2-tailed *t*-test using values derived from 5-fold cross-validation. To perform cross-validation, the dataset was randomly divided into 5 subsets; the performance was computed for each method for each subset; then the mean and variance across the 5 subsets was calculated for each method and used to assess statistical significance between methods using unpaired, 2-tailed *t*-test).

Examining the Spearman correlations among the different metrics themselves is a way of assessing whether they are redundant, or measure different aspects of similarity and relatedness in the context of a given task or situation. In this dataset, the direct odds ratio was highly correlated with the implicit metrics (> 0.85) but showed only a modest correlation with word2vec (0.55) and the implicit metrics also showed modest correlations with word2vec (0.68) (Table 4). These cross-metric correlations are even lower than was observed for the 30 term pair benchmark (Table 3), and confirm again that our implicit metrics are not redundant with word2vec.

3.2.3. Performance on medical resident judgments of biomedical term similarity and relatedness

The UMNSRS-Similarity benchmark of 566 term pairs involves diseases or conditions and includes many drug names (Pakhomov et al. [31]). The dataset, which consists mostly of single words, was rated for similarity by medical residents and is shown in Supplementary File 3 (note: we only evaluated the 501 term pairs that mapped to our dataset). The performance of the various metrics is shown in Table 5. Again, the correlations with human judgments of word similarity for direct odds ratio (0.70) and implicit weighted score (0.69) were high and not significantly different from each other ($p = 0.4676$), whereas the performance of word2vec (0.58) was significantly worse than the other metrics (implicit weighted score was significantly greater than word2vec at $p = 0.0169$, unpaired, 2-tailed *t*-test using values derived from 5-fold cross-validation).

Finally, two versions of the UMNSRS-Relatedness benchmark [31] were examined; the full version containing 588 term pairs (of which 511 term pairs mapped to our dataset), and a modified version consisting of 458 term pairs that occur across different domains (Pakhomov et al. [32]), of which 420 term pairs mapped to our dataset. As shown in Supplementary Files 4 and 5, the benchmarks consist mostly of single words. In the 511 word benchmark (Table 6), the Spearman rank correlations with human judgments of word relatedness were 0.63 for direct odds ratio and 0.60 for implicit weighted score; these correlations were not significantly different from each other (direct odds ratio vs. implicit weighted score $p = 0.3964$).

In contrast, word2vec gave significantly lower performance (0.48) than the other metrics (implicit weighted score vs. word2vec $p = 0.0224$), and similar to word2vec performance reported by Muneeb et al. [33] and Chiu et al. [34]. In the 420 word version (Table 7), all metrics gave slightly better performance, but again, word2vec was significantly lower than the other metrics (implicit weighted score vs. word2vec $p = 0.0015$). The comparison of our implicit metrics against word2vec is fair insofar as neither metric was deliberately tuned for the tasks at hand. However, it should be acknowledged that custom task-specific tuning of the word2vec parameters can improve performance to the point where both metrics are comparable: Henry et al. [29] and Zhu

Table 5
Spearman rank correlations among human similarity scores and similarity metrics for the UMNSRS-Similarity benchmark (evaluating the 501 term pairs that mapped to our dataset).

	Mean	Direct odds ratio	Imp shared	Imp weighted	Word2vec
Mean similarity judgement	1.0000	0.7013	0.6649	0.6931	0.5797
Direct odds ratio		1.0000	0.8322	0.8566	0.7387
Implicit shared terms			1.0000	0.9826	0.8194
Implicit weighted score				1.0000	0.8236
Word2vec					1.0000

Table 6
Spearman rank correlations among medical student relatedness scores and similarity metrics for the UMNSRS-Relatedness benchmark (evaluating the 511 term pairs that mapped to our dataset).

	Mean	Direct odds ratio	Imp shared	Imp weighted	Word2vec
Mean similarity judgement	1.0000	0.6338	0.5645	0.5973	0.4832
Direct odds ratio		1.0000	0.7998	0.8272	0.7197
Implicit shared terms			1.0000	0.9819	0.8229
Implicit weighted score				1.0000	0.8261
Word2vec					1.0000

et al. [35] have reported Spearman correlations of 0.64–0.68 across various word2vec models against UMNSRS-Similarity and 0.59–0.72 against UMNSRS-Relatedness.

Across all benchmarks, the direct odds ratio exhibits a relatively high Spearman rank correlation with the implicit metrics (0.79–0.91), but a consistently lower correlation with word2vec (0.55–0.75). The implicit metrics show a substantial but partial correlation with word2vec (0.68–0.83).

3.3. Evaluation on PubMed article record pairs that match (or do not match) on author

Besides evaluating the similarity metrics on biomedical term similarity and relatedness tasks, we also evaluated them for their ability to compare two text passages. Texts can be compared for similarity (or relatedness) in different ways. For example, in information retrieval, one text may be given as a query and other texts may be ranked in terms of their topical relevance. Here, we compared two title + abstract fields of sole-authored PubMed articles and evaluated how well the similarity metrics could distinguish article pairs written by the same individual, vs. written by different individuals. This task has a clear, objective endpoint and is of practical value, since the similarity metrics are potential features to be used in modeling author name disambiguation [3,4] in cases where articles are sole-authored and therefore presumably written by the individual listed as author.

Using the 2009 Author-ity name disambiguation dataset [3,4,36], we randomly chose 1,000 pairs of sole-author articles published in 1987–2009 in English that had abstracts (and the total number of words in title + abstract was ≥ 25) that were predicted to be written by the same individual, vs. 1,000 pairs of sole-author articles that were predicted to be written by different individuals (i.e., sharing the same last name but having a mismatch on the first initial of the author's name). Pairs were excluded if they had identical titles or abstracts. For each pair of articles, we computed the following six similarity measures on the title + abstract. Direct similarity measures included a) number of shared title words (after stoplisting, tokenization and stemming), b) longest common character string (of at least 3 characters, including spaces), and c) number of shared rare terms (i.e., terms that occurred in fewer than 25 abstracts in MEDLINE, weighted by counting shared bigrams as 2 and shared trigrams as 3). We also examined three implicit similarity measures: d) implicit shared terms, e) implicit weighted score, and f) pvtopic word2vec-based similarity score. The article pairs and computed scores are shown in Supplementary File 6.

The metrics were first assessed in terms of their coverage, that is, the number of article pairs that received nonzero similarity scores, and in terms of their discrimination ratio, that is, the mean similarity value seen across the positive sample divided by the mean value of the negative sample. As shown in Table 8, all of the examined metrics were significantly different (positive vs. negative sample) at $p < 0.0001$ (2-tailed *t*-test, unpaired), indicating that all of the metrics tested were strongly discriminative of authorship. The shared rare terms metric had by far the best discrimination ratio but the worst coverage, indicating that two title + abstracts are unlikely to share any rare terms at all even when they are written by the same individual, but this is extremely unlikely when they are written by different individuals. In contrast, the longest common character string metric had complete coverage (every article pair shared at least 3 common characters) but had the lowest discrimination ratio; even article pairs written by different individuals could share long multi-word strings such as "in the context of global climate change". The number of shared title words metric fell in between the other two direct metrics; only about half of the positive article pairs shared one or more title terms, but less than 1/20th of the negative pairs shared any title terms at all. The three implicit metrics exhibited a nice balance of high coverage (97.4% − 100%) and high discrimination ratios (16.08–21.27).

Table 8
Article similarity metrics computed for pairs of sole-author articles written by the same individual (positive sample) vs. written by different individuals (negative sample).

	Shared title words	Longest common string	Rare terms	Implicit shared terms	Implicit weighted score	Pvtopic
Positive sample:						
Mean:	1.26	20.64	3.77	150.91	2328.80	0.18
Std. Dev.:	1.83	14.50	10.47	102.41	2312.15	0.11
# nonzero values	508	1000	292	974	974	1000
Negative sample:						
Mean:	0.05	9.06	0.00	9.05	109.48	0.01
Std. Dev.:	0.24	2.55	0.03	21.61	315.56	0.07
# nonzero values	48	1000	1	426	426	1000
Discrim ratio	24.65	2.28	3774	16.67	21.27	16.08

All metrics were significantly different (positive vs. negative sample) at $p < 0.0001$, using 2-tailed *t*-test, unpaired.

In order to learn whether the different metrics were capturing distinct features or were largely redundant on this task, we compared the Spearman rank correlations of the similarity scores of each metric against each other. For each metric, the different article pairs across the

Table 7
Spearman rank correlations among medical student relatedness scores and similarity metrics for the modified UMNSRS-Relatedness benchmark (evaluating the 420 term pairs that mapped to our dataset).

	Mean	Direct odds ratio	Imp shared	Imp weighted	Word2vec
Mean similarity judgement	1.0000	0.6658	0.5892	0.6226	0.5062
Direct odds ratio		1.0000	0.7867	0.8129	0.6992
Implicit shared terms			1.0000	0.9828	0.8137
Implicit weighted score				1.0000	0.8125
Word2vec					1.0000

Table 9
Spearman rank correlations among similarity metrics in pairs of sole-author articles written by the same individual (positive sample) vs. written by different individuals (negative sample) and for the two sets combined.

	Implicit shared terms	Implicit weighted score	Shared title words	Longest common string	Rare terms	Pvtopic
Positive sample:						
Implicit shared terms	1.0000	0.9548	0.5148	0.6146	0.4720	0.6424
Implicit weighted score		1.0000	0.4778	0.5916	0.4325	0.6219
Shared title words			1.0000	0.5909	0.5594	0.5393
Longest common string				1.0000	0.7200	0.6326
Rare terms					1.0000	0.5858
pvtopic						1.0000
Negative sample:						
Implicit shared terms	1.0000	0.9979	0.1190	0.2129	0.0307	0.3270
Implicit weighted score		1.0000	0.1189	0.2232	0.0254	0.3255
Shared title words			1.0000	0.1506	−0.0071	0.0912
Longest common string				1.0000	−0.0446	0.1566
Rare terms					1.0000	0.0464
pvtopic						1.0000
Combined:						
Implicit shared terms	1.0000	0.9923	0.6130	0.7081	0.5093	0.7866
Implicit weighted score		1.0000	0.5994	0.7054	0.4943	0.7795
Shared title words			1.0000	0.5959	0.5806	0.5906
Longest common string				1.0000	0.5840	0.6644
Rare terms					1.0000	0.5341
Pvtopic						1.0000

positive sample will naturally vary considerably in terms of their similarity. This is true too to some extent for pairs contained in the negative sample. However, since the overall similarities are so low in the negative sample, the range of variation and the shape of the distribution of scores is likely to be quite different in the positive vs. the negative samples. Thus, this analysis was studied separately for the positive sample, for the negative sample, and for the two samples combined together (which will exhibit the largest overall range of similarity scores).

As shown in Table 9, the implicit shared terms and implicit weighted scores were highly correlated (> 0.95) in all cases, but were much less correlated with the word2vec-based pvtopic score (rank correlations ranging from 0.33 in the negative sample to 0.79 in the combined sample). This indicates that the implicit term metrics introduced in this paper are identifying different aspects of textual similarity than pvtopic, even though both types of metrics showed similar performance at discriminating authorship. All three of the implicit metrics showed limited correlations with the direct metrics (~ 0.5 – ~ 0.7 in the combined sample), and the direct metrics showed similar limited correlations amongst themselves as well.

4. Discussion

In the present paper, we have utilized PubMed titles and abstracts as a corpus to compute several term and text passage similarity metrics, which have been characterized and compared in detail, and are presented as datasets that can be publicly queried or downloaded for further use. Although our interest is focused on biomedical applications, analogous metrics can be computed for any large textual corpus and should have general applicability in text mining across domains.

Three types of term similarity metrics were studied in this paper: a) the **direct odds ratio**, which measures how often two words are observed to co-occur in the same article, relative to the co-occurrence that would be expected by chance; b) **word2vec** similarity, which represents each word as a neural embedding vector and measures similarity as the cosine similarity of any two vectors; and c) our novel unweighted and weighted implicit similarity scores: **implicit shared terms** and **implicit weighted score**, which represent each term as a vector consisting of their 300 most similar context terms (ranked according to direct odds ratio). We also studied text passage similarity metrics consisting of: a) the **pvtopic** formulation of text similarity computed using paragraph2vec [27], and b) our novel unweighted and weighted **implicit score** measures, which represent each title + abstract as a vector consisting of their 300 most similar context terms (according to direct odds ratio), and compute the similarity of any two text passages by counting how many terms are shared in their vector representations.

The major methodological innovation introduced here is the way we have formulated the term and text vector representations and computed their similarities. To our knowledge, no other term or text similarity metrics combine these desirable features: 1. They are unsupervised methods, which have a minimum of tunable parameters, and are readily computed from direct co-occurrence data. 2. They represent words, phrases, or text as a low dimensional vector, in which the meaning and relative importance of dimensions is transparent to inspection. 3. They measure implicit similarity, i.e., the sharing of similar context words and phrases. The most similar system to the present one is Liu et al. [37] who also represented documents as an implicit similarity vector. In their scheme, keyphrases (rather than text words and phrases) are extracted from text, and each keyphrase is represented as a weighted vector ('silouette') consisting of co-occurring keyphrases. After pruning, the optimal number of dimensions in this vector is similar to that studied here, i.e., 200–400. Like our system, their method is transparent and interpretable, and uses implicit similarity measures. However, our representation of a document emerges naturally from its words and phrases, which theirs does not. Also, unlike our scheme, they a) compute a single ranking of latent keywords across the entire corpus of documents, b) they use a different weighting scheme that requires extensive training to assign the weights, and c) they compute the similarity of two documents as the cosine similarity of their vectors.

When choosing a similarity metric for a given task on a given corpus, there exist several considerations. Certainly, one would like a metric that exhibits high performance and is easy to compute. As well, two other factors may be less obvious. One is generalizability of the metric [29,30]: Does the metric give "reasonably" high performance on a wide range of tasks and corpora without the need for custom tuning? Another is the extent to which one metric to be employed in modeling text (e.g., classification or clustering tasks) is redundant with other features already incorporated in the model. The direct odds ratio, implicit shared terms, and implicit weighted score all gave similar high performance on a variety of biomedical term similarity benchmarks, in some cases higher than any other metrics that have been examined to date where the pair of terms is the sole input to the metric (i.e., excluding approaches which employ supplementary outside information

from knowledgebases) [e.g., 5,10,28,29]. These all showed relatively good results without any need for custom tuning. Both the direct odds ratio and implicit shared terms are easy to compute. Yet they clearly capture somewhat different aspects of similarity (Tables 1 and 2) and are not truly redundant in either term similarity / relatedness tasks (Tables 3–7) or article similarity rankings (Tables 8 and 9). The implicit term metrics can be thought of as a generalization or smoothing of the direct odds ratio, providing better sampling (and thus more robustness) when direct co-occurrences are low. Also, the implicit text metrics are more sensitive than direct similarity metrics, since they can provide a measure of similarity even when there are no shared terms at all. Thus, we suggest that the novel weighted and unweighted implicit similarity metrics presented here may be valuable for biomedical text modeling.

We also compared the direct odds ratio, implicit shared terms, and implicit weighted score metrics against word2vec and pvtopic (the word2vec-based implementation employed for text). They exhibited limited cross-correlation of similarity scores across a variety of corpora and tasks, indicating clearly that neural embeddings capture somewhat different aspects of similarity than the implicit metrics introduced here.

Compared to the direct and implicit term metrics, the neural embeddings were observed to have inferior performance on the biomedical term similarity / relatedness benchmarks, as assessed both by our tests (against several different implementations of word2vec that have been developed by other groups for similar tasks) and those reported by others (see Results). However, we acknowledge that that word2vec performance can be optimized for specific tasks by adjusting different choices of parameters and hyperparameters, different neural network architectures, and different ways of handling multi-word phrases [29,35], and we did not test all possible implementations. Thus, we do not claim that our novel metrics necessarily perform better than neural embeddings in general. It should also be noted that we decided to include only the most important, informative words and phrases in our vocabulary (excluding the rare ones), in contrast to the word2vec based embeddings that encompass a much larger vocabulary, We feel that the relatively prevalent words and phrases are likely to be the most valuable for representing biomedical articles. However, it is worth exploring the effects of different vocabulary restrictions, especially when applying our vectors to different corpora, domains and tasks. To our knowledge, no one has studied the effects of using similarly large stop lists for word2vec based embeddings, so it is conceivable that some of the differences that we observe in our metrics vs. word2vec embeddings reflect the differences in vocabulary size.

Limitations of this study. The biomedical term similarity datasets provided for download are corpus-dependent and may not directly apply to other domains (e.g., clinical notes or social media). Although the inclusion of many bigrams, trigrams, and abbreviations should reduce the problem of term ambiguity to some extent, the vector representations of words represent an overall average across word instances and different word senses. The implicit term metrics have relatively few parameters, which we have attempted to set at near-optimal values. However, the choice of 300-dimensions for the vector representations, and the particular weighting schemes for similarity of vectors, may be regarded as best guesses and might be tuned further for optimal performance in specific tasks.

The value of the novel implicit term and text passage metrics remains to be explored for other types of text modeling tasks not examined here, for example, for information retrieval or sentiment analysis. Word2vec-based metrics have been very actively explored for a variety of extrinsic tasks such as named entity recognition, part of speech tagging, ranking PubMed articles for semantic relatedness [19,27], and word sense disambiguation [38]. One should be cautioned that performance on one set of tasks does not necessarily correlate with performance on other tasks [39,40]. It remains to be investigated whether our implicit term similarity metrics have any added value in modeling the latter applications.

Finally, note that in our vector representation schemes, the vector components are an ranked set of terms – NOT an unranked set of numbers, as is the case in most, if not all, other vector representations used in the text mining and computer science communities. On the one hand, this underscores the novelty of our contribution, but on the other hand, it raises the question of whether it will be difficult to handle our vectors computationally, e.g., as features for machine learning frameworks which incorporate order in the 300 dimensional implicit term vectors.

Certainly the 300 dimensional implicit term vectors can be approximated as an unranked set of terms ("bag of words"), and encoded as 1-hot vectors in the vector space consisting of 97,754 dimensions (the number of unique implicit terms in our full term model). Both standard machine learning frameworks, such as SVM and neural networks, will accept 1-hot vectors such as these as input. But this method ignores the ranking of the terms in our vectors. It is possible to weight the terms based on their ranking, for example, giving each term a weight of (300 – rank)/300 scores each feature between zero and one. However, this method provides no obvious way to sensibly combine implicit term features with the weights of features derived from other methods, such as ngrams or word2vec. Our claim is that the discriminative value of implicit term vectors is complementary to these other methods. Therefore, bag of terms representation, whether weighted or un-weighted, is probably not the best way to incorporate these vectors into a machine learning framework.

We have been experimenting with two more advanced approaches of incorporating implicit term features, in combination with other types of features, into machine learning frameworks. The first approach is using the implicit term vectors to generate pairwise article similarity scores (see Methods), which are used as features for input into machine learning frameworks [41]. This approach can accommodate features that arise from multiple vector types which can produce alternative text-based similarity scores, such as word2vec-based vectors, as well as nontextual types of article similarity such as being published in similar journals or being indexed by similar MeSH terms [18,19,41]. In unpublished studies to be presented elsewhere (Cohen et al, in preparation), we have applied this approach to classification of PubMed articles among five publication types related to clinical studies. A combination of article similarity scores were used as features including implicit term vectors, word2vec-based pvtopic vectors, important words, journal similarity, and title and abstract bigrams, inputting all these features into an SVM classifier. For all five publication types, the best performance was observed when including all of the feature types. Removing either the implicit term vector-based similarity score, or the pvtopic similarity score, resulted in decreased performance, showing that the contributions of implicit vectors, word2vec, and bigrams to publication type classification are complementary, and not redundant.

A second potential approach is also promising. It may be possible to use our implicit term vectors as input directly to an embedding layer or a sequence encoding neural network, such as LSTM, which would transform them into a new low dimensional vector that preserves term ranking, and whose components consist of numbers. The resulting numerical vector could then be used in a deep learning classifier, or extracted and used in a traditional machine learning method such as SVM. We are experimenting with this approach as part of applying deep learning methods to publication type classification.

5. Implementation

In order to foster further research on biomedical term similarity and article similarity metrics, we have provided free, public access to our datasets on the Arrowsmith project website http://arrowsmith.psych.uic.edu/arrowsmith_uic/word_similarity_metrics.html. These data are being released under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike CC BY-NC-SA International Public License 4.0. Specifically:

1. The basic word model, consisting of word vector representations and pre-computed word-word similarity measures, is available as datasets that can be downloaded.
2. The full term model, consisting of term vector representations for words, abbreviations, bigrams, and trigrams, and pre-computed term-term similarity measures, is available as datasets that can be downloaded.
3. All PubMed articles (having English abstracts, for which the title + abstract contains at least 25 words) have been represented as 300-dimensional vectors by both the implicit weighted score and by pvtopic. The vector representations are available for download, and as new articles appear weekly, we plan to incrementally add new vectors to the dataset.
4. A query interface at http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/word_sim_metrics.cgi permits the user to enter any word or term (from either the basic or full model) and view the 300 most similar words or terms as pre-computed according to the direct odds ratio, implicit shared terms, implicit weighted score, and word2vec. The interface also shows normalized values by giving the percentile value corresponding to each score (e.g., a given similarity score may be greater than 98% of all similarity scores in the dataset).
5. Another query interface at http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/doc_sim.cgi permits the user to enter any two PubMed article IDs (PMIDs) and view the weighted and unweighted implicit similarity scores. Note that the unweighted implicit shared terms scores range from 0 to 300.
6. Finally, a query interface at http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/pvtopic_sim.cgi permits the user to enter any two PubMed article IDs and view the pvtopic similarity scores. Note that the pvtopic similarity scores range from -1 to $+1$.

Acknowledgements

Our studies are supported by NIH grants R01LM10817 and P01AG03934. We thank Ruixue Wang (Wuhan University) for computing some of the word2vec word similarity scores. Thanks, too, to Keven Bretonnel Cohen for suggesting a more vivid introductory paragraph. A preprint of this paper was first deposited into arXiv [42].

Declaration of interest statement

The authors declare that they have no conflicts of interest.

Appendix A. Supplementary material

Several biomedical term relatedness and term similarity benchmarks were downloaded from http://rxinformatics.umn.edu/SemanticRelatednessResources.htm and shown in Supplementary files 1–5. Most of the words and phrases listed in the benchmarks had exact matches to our dataset of words and phrases used for calculating direct and implicit similarity metrics. When exact matches were not found, terms were manually mapped to the closest words or multi-word terms in our dataset. (For example, some brand names for drugs were mapped to their generic counterparts.) For the words that failed to map to any terms in our dataset, the involved term pairs were removed from consideration. Similarly, for the words that failed to map to the word2vec corpus, the involved term pairs were removed from consideration when computing word2vec similarity. Term mappings to our dataset are shown in Supplementary Files 1–5.

In order to assess whether the Spearman rank correlations showed statistically significant differences for different metrics, we divided the benchmark dataset randomly into five equal subsets and computed the Spearman correlations for each subset. This gave a mean, SD, and SEM for each metric, and allowed us to compare different metrics by unpaired, two-tailed t-test using an online t-test calculator (https://www.graphpad.com/quickcalcs/ttest1.cfm).

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2019.103096.

References

[1] T. Pedersen, S.V.S. Pakhomov, S. Patwardhan, C.G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, J. Biomed. Inform. 40 (3) (2007) 288–299.
[2] Y. Mrabet, H. Kilicoglu, D. Demner-Fushman, TextFlow: a text similarity measure based on continuous sequences, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, Vol. 1, pp. 763–772.
[3] M.E. Lesk, Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, in: Proceedings of the SIGDOC Conference 1986, Toronto, June, 1986.
[4] S. Mohammadi, S. Kylasa, G. Kollias, A. Grama, Context-specific recommendation system for predicting similar pubmed articles, InData Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on 2016 Dec 12, IEEE, pp. 1007–1014.
[5] T.E. Workman, G. Rosemblat, M. Fiszman, T.C. Rindflesch, A literature-based assessment of concept pairs as a measure of semantic relatedness, AMIA Annu. Symp. Proc. 16 (2013) (2013) 1512–1521.
[6] D. Trivison, Term co-occurrence in cited/citing journal articles as a measure of document similarity, Inf. Process. Manage. 23 (3) (1987) 183–194.
[7] T.K. Landauer, P.W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse processes. 25 (2–3) (1998) 259–284.
[8] D. Lin, Automatic Retrieval and Clustering of Similar Words, COLING-ACL (1998) 768–774.
[9] J. Lin, W.J. Wilbur, PubMed related articles: a probabilistic topic-based model for content similarity, BMC Bioinf. 8 (1) (2007) 423.
[10] C. Van Gysel, M. de Rijke, E. Kanoulas, Neural vector spaces for unsupervised information retrieval. arXiv preprint arXiv:1708.02702, 2017 Aug 9.
[11] K.W. Boyack, D. Newman, R.J. Duhon, R. Klavans, M. Patek, J.R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, K. Börner, Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches, PLoS One 6 (3) (2011) e18029.
[12] C. Corley, R. Mihalcea, Measuring the semantic similarity of texts, in: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment 2005 Jun 30, Association for Computational Linguistics, pp. 13–18.
[13] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, In Advances in neural information processing systems 2013, pp. 3111–3119.
[14] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[15] W. Wei, R. Marmor, S. Singh, S. Wang, D. Demner-Fushman, T.T. Kuo, C.N. Hsu, L. Ohno-Machado, Finding related publications: extending the set of terms used to assess article similarity, AMIA Summits Translational Sci. Proc. 2016 (2016) 225.
[16] H. Luo, Z. Liu, H. Luan, M. Sun, Online learning of interpretable word embeddings, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1687–1692.
[17] A. Subramanian, D. Pruthi, H. Jhamtani, T. Berg-Kirkpatrick, E. Hovy, SPINE: SParse Interpretable Neural Embeddings. arXiv preprint arXiv:1711.08792. 2017 Nov 23.
[18] J.L. D'Souza, N.R. Smalheiser, Three journal similarity metrics and their application to biomedical journals, e115681, PLoS One 9 (12) (2014), https://doi.org/10.1371/journal.pone.0115681.
[19] N.R. Smalheiser, G. Bonifield, Two similarity metrics for medical subject headings (MeSH): an aid to biomedical text mining and author name disambiguation, J. Biomed. Discov. Collab 6 (7) (2016) e1, , https://doi.org/10.5210/disco.v7i0.6654.
[20] V.I. Torvik, M. Weeber, D.R. Swanson, N.R. Smalheiser, A probabilistic similarity metric for Medline records: a model for author name disambiguation, AMIA Annu. Symp. Proc. 1033 (2003).
[21] V.I. Torvik, N.R. Smalheiser, Author name disambiguation in MEDLINE, ACM Trans. Knowl. Discov. Data 3 (3) (2009) pii: 11.
[22] V.I. Torvik, N.R. Smalheiser, M. Weeber, A simple Perl tokenizer and stemmer for biomedical text, Unpublished technical report, accessed from http://arrowsmith.psych.uic.edu/arrowsmith_uic/tutorial/tokenizer_2007.pdf December 2017.
[23] Accessed from http://arrowsmith.psych.uic.edu/arrowsmith_uic/data/stopwords_pubmed, January 15, 2019.
[24] W. Zhou, V.I. Torvik, N.R. Smalheiser, ADAM: another database of abbreviations in MEDLINE, Bioinformatics 22 (22) (2006) 2813–2818.
[25] ADAM: Another Database of Abbreviations in MEDLINE, Accessed from http://arrowsmith.psych.uic.edu/arrowsmith_uic/adam.html, December 2017.
[26] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, pp. 1188–1196.
[27] K. Hashimoto, G. Kontonatsios, M. Miwa, S. Ananiadou, Topic detection using paragraph vectors to support active learning in systematic reviews, J. Biomed. Inform. 62 (2016) 59–65, https://doi.org/10.1016/j.jbi.2016.06.001.
[28] Y. Ling, Y. An, M., Liu, S. Hasan, Y. Fan, X. Hu, Integrating extra knowledge into word embedding models for biomedical nlp tasks, in Neural Networks (IJCNN), 2017 International Joint Conference on. IEEE 2017.
[29] S. Henry, C. Cuffy, B.T. McInnes, Vector representations of multi-word terms for

semantic relatedness, J. Biomed. Inform. (2017), https://doi.org/10.1016/j.jbi.2017.12.006.
[30] S.V. Pakhomov, T. Pedersen, B. McInnes, G.B. Melton, A. Ruggieri, C.G. Chute, Towards a framework for developing semantic relatedness reference standards, J. Biomed. Inform. 44 (2) (2011) 251–265, https://doi.org/10.1016/j.jbi.2010.10.004.
[31] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, G.B. Melton, Semantic similarity and relatedness between clinical terms: An experimental study, AMIA Annu. Symp. Proc. 13 (2010) (2010) 572–576.
[32] S.V. Pakhomov, G. Finley, R. McEwan, Y. Wang, G.B. Melton, Corpus domain effects on distributional semantic modeling of medical terms, Bioinformatics 32 (23) (2016) 3635–3644.
[33] T.H. Muneeb, S.K. Sahu, A. Anand, Evaluating distributed word representations for capturing semantics of biomedical concepts, Proc. ACL-IJCNLP 30 (2015) 158.
[34] B. Chiu, G. Crichton, A. Korhonen, S. Pyysalo, How to train good word embeddings for biomedical NLP, Proc. BioNLP16 12 (2016) 166.
[35] Y. Zhu, E. Yan, F. Wang, Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec, BMC Med. Inform. Decis Mak. 17 (1) (2017) 95, https://doi.org/10.1186/s12911-017-0498-1.
[36] Author-ity Exporter, Accessed from http://abel.lis.illinois.edu/cgi-bin/exporter/search.pl December 2017.
[37] J. Liu, X. Ren, J. Shang, T. Cassidy, C.R. Voss, J. Han, Representing documents via latent keyphrase inference, in: Proceedings of the 25th international conference on World Wide Web 2016 Apr 11, International World Wide Web Conferences Steering Committee, pp. 1057–1067.
[38] S. Tulkens, S. Šuster, W. Daelemans, Using distributed representations to disambiguate biomedical and clinical concepts, arXiv preprint arXiv:1608.05605, 2016 Aug 19.
[39] B. Chiu, A. Korhonen, S. Pyysalo, Intrinsic evaluation of word vectors fails to predict extrinsic performance, in: Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP 2016 Aug 7, pp. 1–6.
[40] Q. Ai, L. Yang, J. Guo, W.B. Croft, Analysis of the paragraph vector model for information retrieval, in: Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval 2016 Sep 12, ACM, pp. 133–142.
[41] N.R. Smalheiser, A.M. Cohen, Design of a generic, open platform for machine learning-assisted indexing and clustering of articles in PubMed, a biomedical bibliographic database, Data Inf. Manage. 2 (2018) 21–36, https://doi.org/10.2478/dim-2018-0004.
[42] N.R. Smalheiser, G. Bonifield, Unsupervised low-dimensional vector representations for words, phrases and text that are transparent, scalable, and produce similarity metrics that are complementary to neural embeddings, arXiv 2018 arXiv:1801.01884v2.