

Research and Applications

Better synonyms for enriching biomedical search

Lana Yeganova,¹ Sun Kim ,^{1,2} Qingyu Chen,¹ Grigory Balasanov,¹ W. John Wilbur,¹ Zhiyong Lu¹

¹National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

²Present address: Amazon Alexa AI, Seattle, WA.

Corresponding Author: Zhiyong Lu, PhD, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA (zhiyong.lu@nih.gov)

Received 23 January 2020; Revised 20 May 2020; Editorial Decision 9 June 2020; Accepted 20 August 2020

ABSTRACT

Objective: In a biomedical literature search, the link between a query and a document is often not established, because they use different terms to refer to the same concept. Distributional word embeddings are frequently used for detecting related words by computing the cosine similarity between them. However, previous research has not established either the best embedding methods for detecting synonyms among related word pairs or how effective such methods may be.

Materials and Methods: In this study, we first create the BioSearchSyn set, a manually annotated set of synonyms, to assess and compare 3 widely used word-embedding methods (word2vec, fastText, and GloVe) in their ability to detect synonyms among related pairs of words. We demonstrate the shortcomings of the cosine similarity score between word embeddings for this task: the same scores have very different meanings for the different methods. To address the problem, we propose utilizing pool adjacent violators (PAV), an isotonic regression algorithm, to transform a cosine similarity into a probability of 2 words being synonyms.

Results: Experimental results using the BioSearchSyn set as a gold standard reveal which embedding methods have the best performance in identifying synonym pairs. The BioSearchSyn set also allows converting cosine similarity scores into probabilities, which provides a uniform interpretation of the synonymy score over different methods.

Conclusions: We introduced the BioSearchSyn corpus of 1000 term pairs, which allowed us to identify the best embedding method for detecting synonymy for biomedical search. Using the proposed method, we created PubTermVariants2.0: a large, automatically extracted set of synonym pairs that have augmented PubMed searches since the spring of 2019.

INTRODUCTION

In a traditional search setting, relevant documents may fail to be retrieved if the surface word forms are different from those used in a query. These differences may reflect varying degrees of morphological relationships, from abbreviations to inflections or derivations (eg, needle/needles, autoimmune/autoimmunity) to lexically unrelated synonym pairs (eg, youths/adolescents, vigilant/attentive). This

is particularly magnified in the biomedical domain, known to be rich in synonyms and closely related terms. Therefore, incorporating the detection of semantically similar terms into the search process, including same-stem and different-stem synonyms, can improve retrieval.¹⁻⁷

Pedersen et al⁸ define semantic relatedness as a more general notion, with semantic similarity being a special case of relatedness that

is tied to the “likeness.” For instance, the terms “genotypical” and “phenotypical” are frequently used in the same context and are closely related; however, they are not synonyms, and expanding a query containing 1 term to include the other is unfavorable. “Cornification” and “keratinization,” in comparison, are semantically similar, and retrieving documents containing “keratinization” when searching with “cornification” is beneficial.

Automated estimation of the degree of semantic relatedness—in particular, distinguishing terms that frequently co-occur together from terms carrying the same or very close meaning—is a big challenge in the biomedical domain. The Unified Medical Language System (UMLS; <https://www.nlm.nih.gov/research/umls>) and Medical Subject Headings (MeSH; <https://www.ncbi.nlm.nih.gov/mesh>) are commonly used resources for synonymy.^{9,10} However, controlled vocabularies often treat related terms as synonyms. For example, “trophoblasts” and “syncytiotrophoblasts” belong to the same UMLS concept, but do not convey synonymous meaning. Our goal in this study is to develop solutions for identifying pairs of terms that carry substantially the same meaning and can be treated interchangeably in biomedical search.

Word embeddings have been widely used in natural language processing (NLP) applications, due to their ability to capture rich semantic word representations.¹⁰ Many studies have explored word embeddings for both measuring the semantic similarity of words in the biomedical domain^{7,8,11–13} and for deep learning model inputs for various downstream tasks.^{14–17} It has been observed that the choice of word embedding significantly influences the performance of downstream NLP tasks, with no clear correlation being established between the 2. Moreover, no single method has been found to be superior on a range of NLP applications.

Here, we consider 3 word-embedding methods—word2vec,¹⁸ fastText¹⁹ and GloVe²⁰—to better understand their ability to distinguish synonyms from semantically related terms for application in biomedical information retrieval. We observe that cosine similarity between vectors is common, but is not an ideal choice for measuring the synonymy relationship between word vectors, as it frequently can not distinguish relatedness from synonymy. To address the issue, we combine cosine similarity with isotonic regression²¹ and convert the cosine similarity score into the probability of 2 words being synonyms. These normalized scores represent a more interpretable probabilistic measure of synonymy.

To summarize, these are the contributions of this study, which—to the best of our knowledge—is the first study addressing synonym interchangeability in biomedical search. First, we introduce the BioSearchSyn corpus: a new, manually annotated synonym data set for building and evaluating methods for identifying semantic similarity. Second, we examine and compare word-embedding approaches in their ability to detect synonymous words among related term pairs in BioSearchSyn. Third, we use that set to learn how to transform the cosine similarity measure into a probability of terms being synonyms. And finally, we apply the proposed method to create a large-scale, data-driven resource of about 125 000 term variant pairs that are used to improve search results in PubMed. The BioSearchSyn annotation data set, PubTermVariants2.0, and word embeddings are freely available at <ftp://ftp.ncbi.nlm.nih.gov/pub/lu/Synonyms>.

THE BIOSEARCHSYN CORPUS

The BioSearchSyn corpus provides human judgements on the presence or absence of semantic similarity for 1000 pairs of closely related terms (single words). Our motivation for creating the synonym

set was an objective to evaluate the performance of computational methods at estimating the degree of semantic relatedness between 2 terms with respect to a biomedical search. In particular, the goal is to distinguish terms that are synonyms—that is, carrying the same or very close meanings—from those that are only closely related and/or frequently co-occur together. We define *term1* and *term2* to be semantically similar with respect to a biomedical search if documents matching *term2* and not *term1* are found useful by a searcher whose query includes *term1*. We refer to such pairs of terms as interchangeable or synonymous. Here, we describe the process of collecting candidate term pairs and provide annotation guidelines with justification.

Selecting candidate term pairs

The BioSearchSyn set consists of 2 components: term pairs that stem to the same form (eg, autoimmune/autoimmunity) and term pairs that do not stem to the same form (eg, youths/adolescents). We refer to the term pairs that stem to the same form as same-stem synonyms, and to the pairs that do not stem to the same form as different-stem synonyms. We make a distinction between the same-stem and different-stem synonyms, as they exhibit different characteristics. In the literature, same-stem synonyms are also referred to as “term variants.”

Same-stem synonyms

To collect candidate same-stem synonyms, we started by collecting term pairs (requiring each term to appear in at least 10 articles) that stem to the same form,²² and applied a hypergeometric (HG) test²³ to decide whether the observed co-occurrence of 2 terms is above random. On the pairs that passed the HG test, we computed the morpho-semantic similarity score (MS), following Wilbur and Smith,²⁴ and retained those that scored high (above 0.9). This set contains over 82 000 word pairs. For manual annotation, we sampled 200 random pairs from those that passed the HG and MS tests, 200 random term pairs that passed the HG test but failed the MS test, and 100 random pairs from the pool of term pairs that stemmed to the same form but did not pass the HG test.

Different-stem synonyms

This set includes term pairs that do not stem to the same form. To that end, we collected PubMed terms that appear in 50 or more PubMed documents. For each term, we generated a candidate synonym, following the distributional semantics model,²⁵ and filtered out pairs that appeared in PubMed as collocations. This selection process resulted in roughly 30 000 term pair candidates. From that set, we sampled 500 term pairs and ensured there were at least 10 PubMed articles where candidate synonym *term2* appears without the original term *term1*.

Web annotation tool

Judging term pairs selected from PubMed (<https://pubmed.gov>) is a challenging task, significantly different from judging term pairs in general text. The reason is the abundance of low-frequency tokens in biomedical literature. PubMed contains close to 5 million unique tokens, of which about half appear in 3 or fewer documents. Unlike evaluating the similarity between common English terms that judges would likely be more familiar with, judging biomedical terminology requires context to help better understand the modalities of terms, and determine whether 1 term can be substituted for another in a given context.

Term Pair Review

Term 1: **astemizole**

Term 2: **ebastine**

Question: Can **ebastine** be replaced by **astemizole** in the lower PubMed Abstract?

Explicit Answer

Yes for all

No for all

Antonyms

[Set 1](#)

[Set 2](#)

[Set 3](#)

[Set 4](#)

[Set 5](#)

[Set 6](#)

[Set 7](#)

[Set 8](#)

[Set 9](#)

[Set 10](#)

PMID: 2565265

Title: Pharmacological modulation of cutaneous reactivity to histamine: a double-blind acute comparative study between cetirizine, terfenadine and **astemizole**

Abstract:

In a double-blind study performed in 81 healthy volunteers, 10 mg cetirizine and 60 mg terfenadine given orally in a single administration significantly inhibited histamine. **Astemizole** (10 mg) was completely ineffective. The inhibitory effect of cetirizine was potent and regular whereas 6/28 (21%) volunteers did not respond. The difference observed between cetirizine and terfenadine might be due to differences in the metabolism of the two drugs after administration: terfenadine is rapidly metabolized whereas cetirizine is directly active without the need for biotransformation and, indeed is poorly metabolized.

PMID: 1683840

Title: Pharmacological modulation by cetirizine and **ebastine** of the cutaneous reactivity to histamine.

Abstract:

The peripheral H1-inhibiting effects of cetirizine 10 mg and **ebastine** 10 mg were compared at the skin level after single oral administration. The study was performed in 12 subjects under double-blind randomized crossover conditions. Both drugs were significantly effective up to 24 h. The suppressive effect of cetirizine was significantly more marked.

Figure 1. Illustration of the synonym annotation tool. For each pair of terms, the annotator is asked to judge whether the terms can be used interchangeably in the context of presented PubMed documents.

Given a pair of terms, *term1* and *term2*, our goal is to decide whether they can be used interchangeably in a biomedical search. To make the decision process rigorous, for each pair of terms we selected 10 pairs of PubMed documents. In each pair, 1 document contains *term1* and the second contains *term2* but not *term1*. The document containing *term2* is selected randomly, while the document containing *term1* is computed to be the closest to the *term2* document. Here, the closeness of the 2 documents is computed using the vector retrieval model and a cosine similarity metric based on *tf-idf* (term frequency-inverse document frequency) term weighting.²⁶ A web page was created to assist with the annotation process; for each pair of terms, we presented 10 pairs of abstracts, with each abstract pair in a different tab. The illustration in [Figure 1](#) is a screenshot of the tool.

In the example shown, the annotator is presented with a pair of candidate synonym terms: “astemizole” and “ebastine.” The decision about the pair is made on the basis of 10 pairs of PubMed documents where the terms are presented in context. If in 4 out of 10 pairs of documents the terms are judged to be interchangeable, it is statistically sufficient to claim that the pair is synonymous at a useful level. In the next section, we provide the detailed justification of our 4-out-of-10 judging strategy.

Justification of our 4-out-of-10 judging strategy

We developed a rigorous statistical approach as a guide in deciding whether a pair of terms is synonymous in the context of a biomedical search. Let us assume that it is a satisfactory result if a person retrieves with a variant *w* of a term *u* and finds at least 1 hit in a sample of 10 retrieved documents that carries the same meaning for *w* as *u*. Here, we calculate how this usefulness standard relates to our 4 of 10 positives test, used to decide when we should accept that *w* is synonymous with *u* at a useful level.

Given the probability of a positive on any given draw is *p*, for a random sample of size 10 the probability of no positive hit is $(1 - p)^{10}$. Suppose we want this probability to be less than 0.3. Then:

$$\log(1 - p) < \log(0.3)/10$$

$$p > 1 - \exp\left(\frac{\log(0.3)}{10}\right) = 0.1134$$

Given a word *w*, now suppose we have a candidate synonym or variant form, *u*. The question we ask is: what is the probability that in a randomly chosen document containing *w* but not *u*, the meaning of *w* is substantially the same as *u* in the nearest document containing *u*? We want this probability to exceed 0.1134. We will have

Table 1. BioSearchSyn set statistics

	Positives	Negatives	Total
Same-stem synonym pairs	399	101	500
Different-stem synonym pairs	209	291	500

$p > .1138$ with 98% confidence if we examine 10 randomly chosen examples and find the statement to be true in 4 or more cases. In this scenario, we can assume with 98% confidence that if w is substituted for w , among the top 10 hits there will be 1 where w has substantially the same meaning as w in the closest document with w , 70% of the time. In other words, for about 98 out of 100 of the terms we have judged and found to have 4 or more positives out of 10, it will be true that $p > .1134$. For these 98 pairs, it will be true that at least 70% of the time for a retrieval of 10 cases of w documents, at least 1 will have the same meaning for w as w . This analysis is performed for a single-term query. For queries containing 2 or more tokens, the odds of success are significantly higher, because additional query tokens focus the context of retrieval.

Manual annotation processes

Using this approach, a group of 12 scientists with backgrounds in biomedical informatics annotated 1000 candidate synonym pairs. Each term pair received 2 independent annotations. Term pairs that were not agreed upon underwent a second round of reviews. At that stage, a decision about the term pair was reached. The resultant synonymy set, which we refer to as the BioSearchSyn set, consists of 2 subsets: 500 same-stem pairs and 500 different-stem pairs. Table 1 presents the statistics in terms of the numbers of positive and negative pairs in the same-stem and different-stem synonym sets. In Table 2 we show examples of same-stem and different-stem candidate pairs found in the BioSearchSyn set.

In computing candidate synonym pairs, we learned that generating different-stem term pairs is significantly more challenging than generating candidate synonyms that stem to the same form. This results in a much smaller pool of candidate term pairs with different stems, as compared to same-stem pair candidates. At the same time, we observed that about 80% of candidate term pairs in the same-stem set are judged to be synonyms, compared to only 41.8% of pairs in the different-stem set. Different-stem pair candidates are computed using distributional similarity models, which can not distinguish well between related terms (astemizole/ebastine), synonyms (ecologically/environmentally), and antonyms (hypertonicity/hypotonicity), due to similar language used with these term groups. For example, distinct genes and gene functions appear in the same context, resulting in unrelated genes being predicted as candidate synonyms. As a result, different-stem pairs that are judged to be synonyms are enriched in general terms: for example, inmates/prisoners, purchased/bought, and handoffs/handovers. On the contrary, different-stem term pairs with specific biomedical meanings (cochlo-dinium/osteoporosis) are mostly judged as not synonyms and should not be replaced for the purposes of search.

Another interesting example of such term pairs is fever/hyperthermia. While closely related and perceived as synonymous by many people, the underlying biological processes of fever and hyperthermia are different, and they are not perceived as synonymous by PubMed's readership. Fever is an internally regulated rise in temperature, while hyperthermia is an unregulated rise. A user searching

Table 2. Examples of synonym and non-synonym term pairs in the BioSearchSyn set

Data Sets	Word pairs		Label
Same-stem synonyms	xenogeneic	xenogenic	Y
	nanorobot	nanorobotic	Y
	inhibitors	inhibiting	Y
	calculator	calculus	N
Different-stem synonyms	comparators	comparing	N
	handoffs	handovers	Y
	adversities	hardships	Y
	ecologically	environmentally	Y
	hypertonicity	hypotonicity	N
	cochlo-dinium	osteoporosis	N
	astemizole	ebastine	N

Note: N: non-synonym term pair; Y: synonym term pair.

with “fever” expects papers about infections, while a user searching with “hyperthermia” likely expect papers about heat stroke.

MATERIALS AND METHODS

Detecting synonymy

A typical way to evaluate the closeness of 2 words is to measure the cosine of the angle between the word embeddings representing the 2 words. Given the vector embeddings A and B , the cosine similarity is measured as $A \cdot B / \|A\| \|B\|$. The cosine similarity score is often used to compare how well different embedding algorithms capture word similarity and relatedness, as well as for other benchmark tasks. We used cosine similarity scores as our method.

We trained word2vec, fastText, and GloVe on all PubMed documents. Documents are pre-processed using a customized Natural Language Toolkit Treebank tokenizer. For word2vec, we used the skip-gram model. The parameters in word2vec and fastText were $1e-4$, 10, and 0.05 for sub-sampling, negative sampling, and learning rates, respectively. The minimum and maximum numbers of characters were set to 2 and 3 for fastText; $x_{max} = 100$ for GloVe; and the number of training epochs was set to 50 for all models. Also, for all models, we disregarded words that appeared fewer than 5 times in the PubMed corpus. With these parameters fixed, during training we varied the vector size (vec) and the window size (win), and tested 4 combinations: $vec = 100$ and $win = 5$; $vec = 100$ and $win = 10$; $vec = 200$ and $win = 5$; and $vec = 200$ and $win = 10$. The hyperparameters were selected based on the existing evaluation on word embeddings;¹⁹ the tested values of the vector dimension and window size in our study are a subset of those in Bojanowski et al.¹⁹

We observed 2 problems using cosine similarity for identifying synonyms for a biomedical search. The first is the absence of consensus regarding the embedding method that captures synonymy the best. Second, embeddings are good at identifying related words and words that frequently co-occur, but they possess no mechanism to distinguish synonyms from words that are related but do not carry the same meaning. As a result, models that provide state-of-the-art performance over multiple benchmarks do not necessarily work well for the task of finding synonyms for a biomedical search. What we find important for a search is to be able to include more synonyms among closely located words in Euclidean space. Based on the data we have collected, we have observed that the performance of the

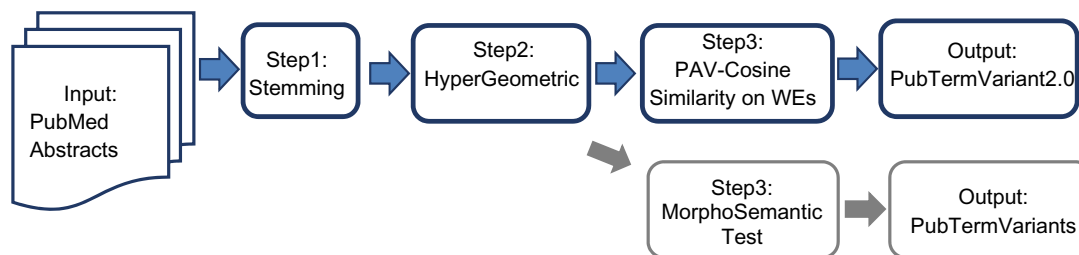


Figure 2. Graphical representation of data-driven workflow for generating the PubTermVariants2.0 resource. The workflow is compared to that of PubTermVariants. Both methods start with stemming and the hypergeometric test. The morphosemantic test previously used for computing PubTermVariants is now replaced with the proposed probabilistic word similarity score. Abbreviations: PAV, pool adjacent violators; WEs, word embeddings.

methods differs in how well cosine scores rank synonym pairs above non-synonym pairs, and also in how accurately the cosine scores reflect the probability a pair of words are synonyms.

Because cosine scores vary in how well they reflect the probability that a word pair are synonyms, we have found it convenient to use a regression method to convert cosine scores to probabilities of synonymy. The pool adjacent violators (PAV) algorithm²¹ is an isotonic regression algorithm that derives a monotonically non-decreasing estimate, which assigns a maximal likelihood to the data. We use it as a method for converting a cosine-similarity score into a probability that 2 terms are synonyms. The estimate is derived based on the cosine similarity scores of word pairs and their corresponding labels in the BioSearchSyn corpus. PAV allows one to make a more accurate estimate of the probability of 2 words being synonymous, compared to the raw cosine-similarity score. While the PAV function is simple and straightforward, it is very useful and has been successfully used in biomedical applications.^{27,28} Another option would have been logistic regression. We compare isotonic and logistic regression on our data in the [Supplementary Appendix](#). While the performance of the 2 methods is comparable, isotonic regression seems to perform slightly better in thresholding to obtain useful data.

Computing synonymy at PubMed scale

We integrated the proposed methodology into a data-driven, large-scale, automatic synonym identification approach to produce a better resource of term pairs that have the same meaning and can be used interchangeably in PubMed searches. We focused on term pairs that stem to the same form. The proposed technique produced PubTermVariants2.0: an improvement to the previous PubTermVariants⁶ resource.

Figure 2 presents a graphical representation of the workflow for computing term pairs. Like computing PubTermVariants,² the pipeline starts by reading the entire PubMed and extracting space-separated tokens appearing in 5 or more articles. Only those that are 5 characters or longer are retained, based on the observation that shorter words tend to be more ambiguous and are frequently abbreviations.⁶ The Porter stemmer²² is used for stemming, and words that stem to the same form are paired. Then, the hypergeometric distribution and the *P*-value test are used for every pair of words to determine whether the observed co-occurrence of 2 words is likely to be by chance.²³ We applied the hypergeometric test to the collected pairs and selected those that passed the test (≤ 0.01). Both steps were performed as was done for the PubTermVariants pipeline. The novel modification replaces the morphosemantic test with the proposed probabilistic word similarity computed on word2vec embeddings.

The PAV-modified scores are computed by applying the PAV regression function, which converts the cosine similarity score into the likelihood of 2 terms being related. Target performance is set to 92% precision or higher, based on the manually annotated BioSearchSyn corpus.

RESULTS

Cosine similarity as a synonymy measure on the BioSearchSyn set

The performance of the 3 word-embedding methods is examined in this section. In training word embeddings, we applied 2 choices for vector size and 2 choices for the window size, resulting in 4 combinations: $vec = 100$ and $win = 5$; $vec = 100$ and $win = 10$; $vec = 200$ and $win = 5$; and $vec = 200$, $win = 10$. We tested these 4 parameter settings and found very little difference in performance for each method. These 4 parameter settings were also tested in a downstream task of sentence retrieval, and the best performing setting was found to be $vec = 100$ and $win = 10$. Therefore, for the rest of the paper we will be discussing results associated with vector size 100 and window size 10.

Figure 3 depicts the precision-recall graphs for word2vec, fastText, and GloVe. Here, we observe that for the same-stem synonyms, word2vec overall outperforms fastText and GloVe. However, for the different-stem synonyms, fastText seems to be better for high-precision results, but GloVe shows better results if higher recall is preferred.

These analyses lead us to the conclusion that cosine similarity scores are not comparable across the different methods. The results are mixed, and no single method stands out as preferred for all purposes. However, our interest is in skimming off the top-ranked material with the highest probability of being true synonyms; for this purpose, word2vec appears to perform best on the same-stem data and fastText on the different-stem data. One may ask whether we have sufficient data to show how well these methods perform. To answer this question, we applied the information retrieval measure of average precision to the rankings produced by word2vec and fastText to rate their performance, and we compare that performance with expected performance if the scores were randomly shuffled between the data points for each method. Average precision is a good performance measure for our purposes, because it is sensitive to getting positive points in the top ranks. Based on 10 000 random shuffles of each data set, we derive a mean and a 95% confidence interval using the percentile method applied to the very symmetric distributions that are produced. Results are in the second row of Table 3. We see that the

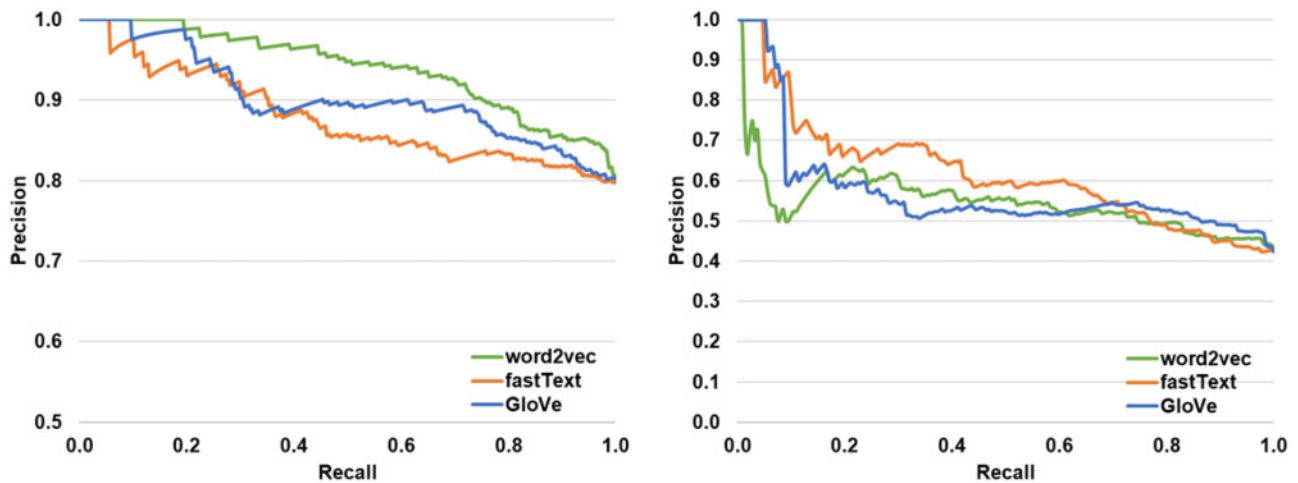


Figure 3. Precision-recall graph for word2vec (green), fastText (orange), and GloVe (blue) for the same-stem synonym set on the left and different-stem synonym set on the right. The embeddings were obtained with a vector size of 100 and window size of 10.

Table 3. Average precision of cosine scores

	Same-Stem Pairs, word2vec	Different-Stem Pairs, fastText
Average precision	0.9396	0.6164
Random shuffled mean average precision (95% CI, reshuffled)	0.8003 (0.7659–0.8343)	0.4248 (0.3842–0.4700)
Bootstrap resampled mean average Precision (95% CI, resampled)	0.9395 (0.9185–0.9579)	0.6178 (0.5491–0.6855)

Note: Data are ranked for same-stem and different-stem pairs, compared with the mean average precisions and CIs obtained when the cosine similarity scores are randomly shuffled between data points. Results for both sets of word pairs are far above the upper 95% confidence limits. The resampled confidence limits show where, with 95% confidence, we can expect performance of a method on the whole data space to lie. CI: confidence interval.

Table 4. Bootstrap resamplings and average precision

Same-Stem Pairs		Different-Stem Pairs	
Data set	Average Precision (95% CI)	Data set	Average Precision (95% CI)
word2vec-GloVe*	0.0370 (0.0190–0.0564)	fastText-GloVe	0.0699 (–0.0026 to 0.1416)
word2vec-FastText*	0.0619 (0.0321–0.0942)	fastText-word2vec	0.0453 (–0.0242 to 0.1157)
GloVe-FastText	0.0249 (–0.0087–0.0594)	GloVe-word2vec	0.0246 (–0.0171 to 0.0649)

Note: There were 10 000 bootstrap resamplings for each data set. For each sample, we computed the average precision for each of the 3 embedding methods. This allowed us to compare the performance of the methods as the difference in average precision on each sample and compute the mean and 95% confidence interval for each comparison. We found that word2vec is better than GloVe or fastText on the same-stem data. *Significant difference at the 5% level.

CI: confidence interval.

observed average precisions of the data sets are well above the upper bound of the 95% shuffled confidence limits. This is strong evidence of the correlation between the human judgments and the cosine scores produced by embeddings for the pairs.

The bootstrapped (resampled) confidence intervals in the third row of Table 3 are also important, as they show that with 95% confidence the performance of the word2vec method on the whole data set of same-stem pairs can be expected to lie between 92 and 96% average precision. This is excellent performance and justifies our use of the method. It also shows that a different random sample of that data than the sample we actually judged would not be expected to yield results much different from what we see. A good share of the variation in samples obtained from random resampling is due to the variation in the relative numbers of positive and negative points, and a lower number of positive points will be reflected in lower av-

erage precision for a sample, but also in a lower expected average precision when scores are randomly shuffled and lower rank-shuffled confidence limits.

The bootstrap resampled confidence intervals for the fastText method on the different-stem pairs yield many of the same conclusions. The performance of the fastText method is clearly well above random; however, a performance in the 60% average precision range is not sufficient to provide a significant number of high-quality predictions to enhance searches.

One may ask whether we can show statistically that word2vec performs better than fastText and GloVe on the same-stem data, and what can be said for the different-stem data. We created the same 10 000 resamplings with replacement used above, and computed the differences between the average precisions of 2 different methods for each sample. We did this for the same-stem data and

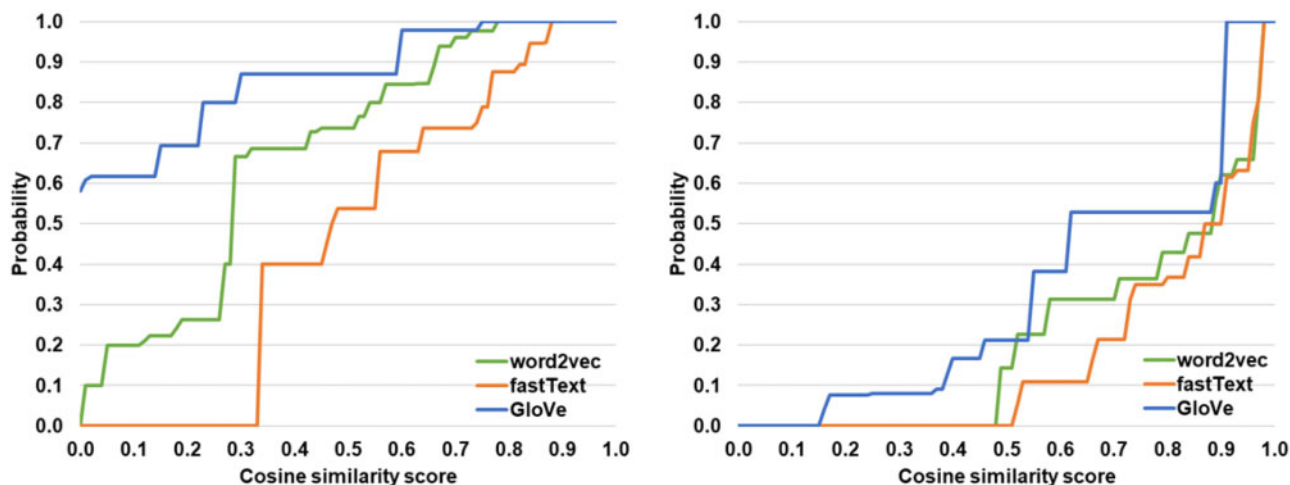


Figure 4. Cosine similarity: probability mapping graph for word2vec (green), fastText (orange), and GloVe (blue) for the same-stem synonym set on the left and different-stem synonym set on the right. PAV functions were trained on embeddings with a vector size of 100 and window size of 10. Abbreviation: PAV, pool adjacent violators.

the different-stem data. The distributions obtained are very symmetric, and we applied the percentile method to estimate 95% confidence intervals. The results for all comparisons are in [Table 4](#).

PAV transformation of cosine similarity scores

As mentioned above, the cosine similarity measures how close 2 words are in vector space, but it does not accurately reflect the likelihood of synonymy between a pair of words. Here, we address this issue by transforming the cosine similarity score into a probability of terms being synonyms. We show the results of applying the PAV transformation to cosine similarity scores. Results for word2vec, fastText, and GloVe are shown in [Figure 4](#). In the figure, the X-axis represents the cosine similarity score between a pair of terms, and the Y-axis represents the probability that 2 terms in a pair with that score are synonyms. These probabilities are computed using the BioSearchSyn set.

As shown in [Figure 4](#), raw cosine similarity scores do not provide accurate estimates of the likelihood of synonymy. The cosine similarity score has a very different meaning for the 3 different methods. However, when we normalize scores using the PAV, we obtain a more interpretable probabilistic measure of synonymy. We also observe that the probability of a set of terms being a synonym pair is not a linear function of the cosine score coming from the embeddings.

PubTermVariants2.0

We implemented the pipeline described in [Figure 2](#) to obtain PubTermVariants2.0. In addition to collecting term pairs from PubMed documents, we explored MeSH and UMLS as additional sources for obtaining same-stem pairs. MeSH is a controlled vocabulary for indexing and searching biomedical literature. Frequently, a MeSH heading is associated with MeSH entry terms. These are terms found to be synonymous to the MeSH heading. We collected pairs of single-word MeSH headings and single-word MeSH entry terms that represent a singular-plural relationship, keeping only those that appear in PubMed and are 5 characters or longer.

The resulting PubTermVariants2.0 set contains 125 072 term pairs, compared to 71 839 term pairs of 5 characters or longer in the

original PubTermVariants set. About 90% of PubTermVariants (64 389 pairs) also appear in the new set, while about 10% are ruled out by the probabilistic similarity constraint when set to target 92% precision. Two annotators worked on a sample of 20 term pairs from the ruled-out portion, and found that only 75% of pairs are true synonyms, which supports the effectiveness of the probabilistic similarity constraint. An example of a correctly ruled-out pair is *sulfusate/suffusion*, where *sulfusate* is the compound that is being *suffused*, while *suffusion* is the process of permeating or infusing something with a substance. Other correctly ruled-out examples from the original set of pairs are *perineural/perineuritis*, *mushrooming/mushrooms*, and *mineralizer/minerals*. We also considered a random sample of 20 term pairs that are in PubTermVariants2.0 and not in PubTermVariants. Two annotators manually annotated these pairs and labeled all of them positive. An example of a term pair that was correctly added is *bronchiolitides/bronchiolitis*, with *bronchiolitides* being the plural form of *bronchiolitis*. As of the spring of 2019, PubTermVariants2.0 replaced PubTermVariants to support indexing and search functionality for PubMed.^{29,30} We further used the proposed method with a PAV threshold of 92% to score the 500 same-stem term pairs of the BioSearchSyn set, based on 5-fold cross-validation. The 500 same-stem pairs are composed of 200 term pairs that pass both the HG and MS tests, 200 term pairs that pass the HG test but not the MS test, and 100 term pairs that do not pass the HG test. Of the 200 pairs that pass both the HG and MS tests, 120 pairs scored above the 92% threshold; among these, 118 (98.33%) are labeled positive. Of the 200 pairs that pass the HG test, but not the MS test, 56 pairs scored above the 92% threshold; among these, 52 (92.86%) are labeled positive. Of the 100 pairs that do not pass the HG test, only 6 score above the threshold, all of which are labeled positive in the gold standard. This provides support for the assertion that PubTermVariants2.0 is at least 92% positives.

DISCUSSION

Because the goal of the method is to provide a high-accuracy data set for use in PubMed searches, in the error analysis we concentrated on false positives: 2 pairs from the set that pass both the HG and MS tests (*nitronates/nitrone* [score 0.936]; and *grove/groves* [score 0.949]) and 4 pairs from the set that pass HG but not the MS test

(organisation/organize [score 0.936]; comparators/comparing [score 0.936]; publishable/publishers [score 0.949]; prompt/prompting [score 0.949]). The term pair of nitronates/nitrone represents different chemical groups, despite the lexical similarity. On one hand, the term “comparing” is used abundantly in PubMed literature, as many research studies rely on comparing a proposed method to others. “Comparator,” on the other hand, is often a physical instrument for performing a specific comparison, and expanding the term comparing to term comparator is not justified. Compare/comparator is a case of semantic drift, and we would place organisation/organize, publishable/publishers, and prompt/prompting in the same category. Of these examples, publishable/publishers is the closest to a case of synonymy, but it is too distant to be useful for searches. The case of grove/groves appears to be different: most frequently these terms are synonyms having a singular plural relationship. However, grove also appears as a misspelling of groove, and this appears to have contributed to judging the pair as negative. Word-embedding scores are high for all these word pairs, since the words in each pair tend to appear together in PubMed documents; for example, there are about 500 documents in PubMed that include both “comparators” and “comparing.”

CONCLUSION

In this work, we address the identification of biomedical synonyms and their usage in biomedical searches. We introduce a corpus of 1000 term pairs, the BioSearchSyn set, which provides a reliable benchmark for examining the semantic similarity of terms in the context of biomedical searches. Using the BioSearchSyn corpus, 3 word-embedding approaches were compared in their ability to detect synonyms among related-term pairs using the cosine similarity between word vectors. We observed that cosine similarity scores are not comparable across the different methods: the same score has very different meanings for the 3 methods. We proposed converting the raw cosine similarity score into a probability measure of terms being related using PAV: the isotonic regression algorithm. This normalized score provides a probabilistic measure of synonymy and allows a uniform interpretation of the different word-embedding scores.

Based on our findings, we created PubTermVariants2.0: a set of 125 072 synonymous, biomedical term pairs from PubMed that has been recently incorporated into the PubMed search. With the objective to provide a highly accurate resource, PubTermVariants2.0 includes only single-word, same-stem synonyms. Future work will focus on improving the ability to detect different-stem synonyms, as well as multiword synonyms, for inclusion in the knowledge base. In addition to its use in biomedical searches, PubTermVariants2.0 may be of use in a range of biomedical natural language processing tasks, including semantic similarity and summarization tasks.

FUNDING

This work was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

AUTHOR CONTRIBUTORS

LY and SK are joint first authors. LY, SK, WJW, and ZL conceived the study, designed the experiments, and participated in every aspect of the study. GB provided the web page for annotation, monitored the annotation process, and performed the subsequent analysis of the data. SK and QC computed word

embeddings and compared different embedding approaches. All authors participated in writing and reviewing the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

- Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. *J Biomed Inform* 2009; 42 (2): 390–405.
- Fiorini N, Canese K, Starchenko G, et al. Best match: new relevance search for PubMed. *PLoS Biol* 2018; 16 (8): e2005343.
- Fiorini N, Leaman R, Lipman DJ, Lu Z. How user intelligence is improving PubMed. *Nat Biotechnol* 2018; 36 (10): 937–45.
- Hersh W. Health informatics series. In: Kathryn MJB, Hannah J, eds. *Information Retrieval: A Health and Biomedical Perspective*. Berlin, Germany: Springer; 2009.
- Kim S, Fiorini N, Wilbur WJ, Lu Z. Bridging the gap: incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. *J Biomed Inform* 2017; 75: 122–7.
- Yeganova L, Kim W, Kim S, et al. PubTermVariants: biomedical term variants and their use for PubMed search. In: proceedings from the 15th Workshop on Biomedical Natural Language Processing; 12 August 2016; Berlin, Germany.
- Yu Z, Wallace BC, Johnson T, Cohen T. Retrofitting concept vector representations of medical concepts to improve estimates of semantic similarity and relatedness. In: proceedings of the 16th World Congress on Medical and Health Informatics; 21–25 August 2017; Hangzhou, China.
- Pedersen T, Pakhomov SVS, Patwardhan S, et al. Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform* 2007; 40 (3): 288–99.
- Qu M, Ren X, Han J. Automatic synonym discovery with knowledge bases. In: proceedings from the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 13–17 August 2017; Halifax, Canada.
- Zhang C, Li Y, Du N, Fan W, Yu P. *SynonymNet*: multi-context bilateral matching for entity synonyms. In: proceedings from International Conference on Learning Representations; 6–9 May 2019; New Orleans, LA.
- Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton. Semantic similarity and relatedness between clinical terms: an experimental study. In: proceedings from the AMIA Annual Symposium/AMIA Symposium; November 13–17, 2010; Washington DC.
- Zhu Y, Yan E, Wang F. 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* 2017; 17 (1): 1–8.
- Chiu B, Pyysalo S, Vulić I, Korhonen A. Bio-SimVerb and Bio-SimLex: wide-coverage evaluation sets of word similarity in biomedicine. *BMC Bioinform* 2018; 19 (1): 33.
- Chen Q, Lee K, Yan S, Kim S, Wei C, Lu Z. BioConceptVec: creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLoS Comput Biol* 2020; 16 (4): e1007617.
- Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. In: proceedings from the Seventh IEEE International Conference on Healthcare Informatics; 10–13 June 2019; Beijing, China.
- Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data* 2019; 6 (1): 1–9.

17. Hassanzadeh H, Nguyen A, Verspoor K. Quantifying semantic similarity of clinical evidence in the biomedical literature to facilitate related evidence synthesis. *J Biomed Inform* 2019; 100: 103321.
18. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: proceedings from Neural Information Processing Systems (NIPS); 5–10 December 2013; Lake Tahoe.
19. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 2017; 5: 135–46.
20. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: proceedings from the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics; October 25–29, 2014; Doha, Qatar.
21. Ayer M, Brunk HD, Ewing GM, Reid W, Schlosser-Silverman E. An empirical distribution function for sampling with incomplete information. *Ann Math Stat* 1955; 26 (4): 641–7.
22. Porter MF. An algorithm for suffix stripping. *Program* 1980; 14 (3): 130–7.
23. Larson HJ. *Introduction to Probability Theory and Statistical Inference*. 3rd ed. New York, NY: John Wiley & Sons; 1982.
24. Wilbur WJ, Smith L. A study of the morpho-semantic relationship in Medline. *Open Inf Syst J* 2013; 6 (1): 1–12.
25. Lin D. Automatic retrieval and clustering of similar words. In: proceedings from the 36th Annual Meeting of the Association for Computational Linguistics; 10–14 August 1998; Montreal, Quebec.
26. Witten IH, Moffat A, Bell TC. *Ranking and Information Retrieval, in Managing Gigabytes: Compressing and Indexing Documents and Images*. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 1999.
27. Kim S, Yeganova L, Wilbur WJ. Summarizing topical contents from PubMed documents using a thematic analysis. In: proceedings from the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics; 17–21 September 2015; Lisbon, Portugal.
28. Wilbur WJ, Yeganova L, Kim W. The synergy between PAV and AdaBoost. *Mach Learn* 2005; 61 (1-3): 71–103.
29. Fiorini N, Lipman DJ, Lu Z. Towards PubMed 2.0. *Elife* 2017; 6: e28801.
30. Fiorini N, Canese K, Bryzgunov R, et al. PubMed Labs: an experimental system for improving biomedical literature search. *Database (Oxford)* 2018; bay094. doi: 10.1093/database/bay094.