Journal of the American Medical Informatics Association, 27(10), 2020, 1538–1546
doi: 10.1093/jamia/ocaa136
Advance Access Publication Date: 30 June 2020
Research and Applications

Research and Applications

Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts

Yuqing Mao and Kin Wah Fung

National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Corresponding Author: Kin Wah Fung, MD, MS, MA, Building 38A, Rm9S918, MSC-3826 National Library of Medicine 8600 Rockville Pike Bethesda, MD 20894, USA; kwfung@nlm.nih.gov

Received 10 February 2020; Revised 3 June 2020; Accepted 4 June 2020

ABSTRACT

Objective: The study sought to explore the use of deep learning techniques to measure the semantic relatedness between Unified Medical Language System (UMLS) concepts.

Materials and Methods: Concept sentence embeddings were generated for UMLS concepts by applying the word embedding models BioWordVec and various flavors of BERT to concept sentences formed by concatenating UMLS terms. Graph embeddings were generated by the graph convolutional networks and 4 knowledge graph embedding models, using graphs built from UMLS hierarchical relations. Semantic relatedness was measured by the cosine between the concepts' embedding vectors. Performance was compared with 2 traditional path-based (shortest path and Leacock-Chodorow) measurements and the publicly available concept embeddings, cui2vec, generated from large biomedical corpora. The concept sentence embeddings were also evaluated on a word sense disambiguation (WSD) task. Reference standards used included the semantic relatedness and semantic similarity datasets from the University of Minnesota, concept pairs generated from the Standardized MedDRA Queries and the MeSH (Medical Subject Headings) WSD corpus.

Results: Sentence embeddings generated by BioWordVec outperformed all other methods used individually in semantic relatedness measurements. Graph convolutional network graph embedding uniformly outperformed path-based measurements and was better than some word embeddings for the Standardized MedDRA Queries dataset. When used together, combined word and graph embedding achieved the best performance in all datasets. For WSD, the enhanced versions of BERT outperformed BioWordVec.

Conclusions: Word and graph embedding techniques can be used to harness terms and relations in the UMLS to measure semantic relatedness between concepts. Concept sentence embedding outperforms path-based measurements and cui2vec, and can be further enhanced by combining with graph embedding.

Key words: UMLS, semantic relatedness, medical terminologies, deep learning, word embedding, graph embedding

INTRODUCTION

Semantic relatedness refers to the notion of whether 2 concepts are closely related in meaning. The concepts do not need to be the same type of concepts (ie, belong to the same semantic type) to be considered related. For example, "heart," an anatomic entity is closely related to "heart failure," a disorder. Studies using human judges have shown that there is considerable agreement on semantic relatedness of most concept pairs.[1] Automated measurement of semantic relatedness has become an important tool in multiple areas including information retrieval, text mining, and natural language processing

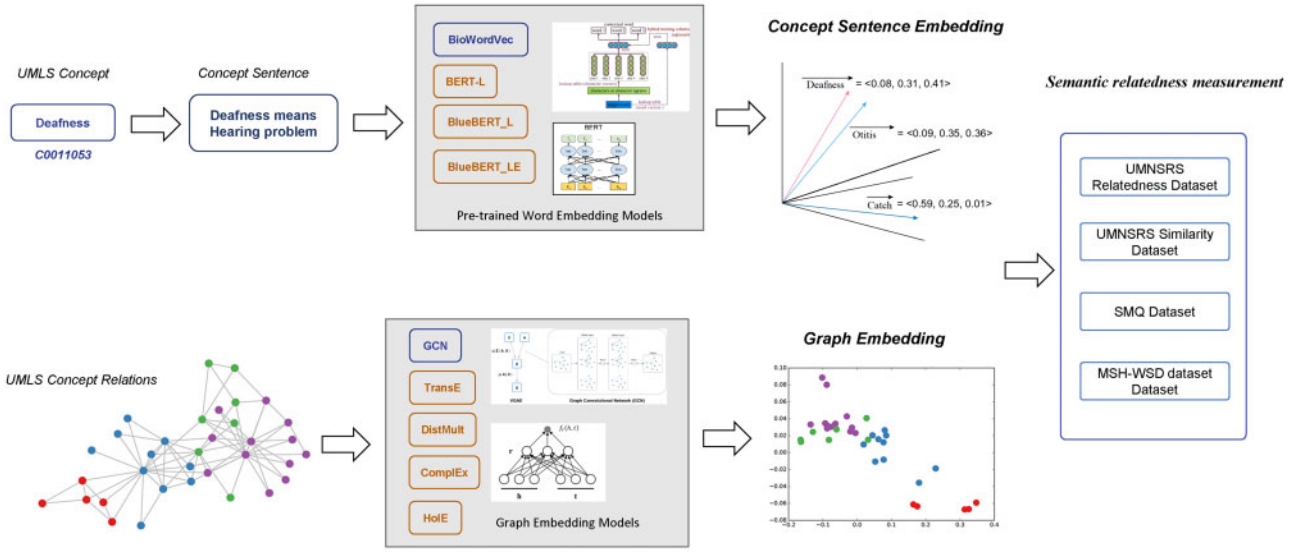

Figure 1. Overview diagram of the proposed method. GCN: graph convolutional network; MSH-WSD: MeSH Word Sense Disambiguation; SMQ: Standardized MedDRA Query; UMLS: Unified Medical Language System.

(NLP). Some examples are information retrieval,[2] information extraction,[3] automatic spelling correction,[4] machine translation,[5] document classification,[6] and word sense disambiguation.[7] Semantic similarity is a special kind of semantic relatedness that refers to the "likeness" between concepts. Semantic similarity usually applies to concepts belonging to the same semantic type and can be linked by hierarchical relationships within a taxonomy or similar artifacts.[1]

In the biomedical domain, measurement of semantic relatedness can generally be divided into knowledge-based and distributional methods.[1] Knowledge-based methods rely on existing knowledge sources such as dictionaries and taxonomies. The most common approaches involve path finding measures based on hierarchical relationships.[8] Distributional methods rely on the distribution of concepts in a corpus to compute relatedness. Based on the distributional hypothesis, concepts that are related are more clustered in their distributional space compared with unrelated concepts.[9] Corpora used generally come from clinical documents or the scientific literature. Knowledge-based and distributional methods have their strengths and weaknesses. Knowledge-based methods are generally more practical to use because clinical documents may not be readily available and the processing of large corpora is computationally expensive.[6] Moreover, the results of distributional methods have been shown to vary according to the corpora used, which may be a problem for standardization and benchmarking.[10,11]

The Unified Medical Language System (UMLS) is a commonly used knowledge source in semantic relatedness measurement.[12] UMLS is a medical terminology resource developed by the U.S. National Library of Medicine that incorporates hundreds of biomedical vocabularies.[13] The vast collection of terms and their relations in the UMLS provides a rich source of material to support various semantic relatedness measurements. Traditionally, most of these methods focus on the relations more than the terms. With the advent of deep learning, new methods have emerged such as word embedding and graph embedding, which can be leveraged to exploit both the terms and relations in the UMLS for semantic relatedness measurement. In this study, we use publicly available word embedding models to generate concept sentence embeddings based on the UMLS terms. Instead of the traditional path counting approach, we apply graph embedding models to learn the context of the UMLS relations. We compare our results to some published path-based semantic related measurements and corpus-based concept embeddings. As far as we know, this is the first study that combines word and graph embedding to measure semantic relatedness, using the UMLS as the sole knowledge source.

Background

Word embedding

Word embedding is a common technique used in NLP and machine learning. Basically, it is the process of transforming words and phrases into vectors of real numbers. The essential requirement of word embedding is that words with similar meaning should have a similar representation. Most of the word embedding models are based on the theory of distributional semantics—the meaning of words in a text can be estimated by looking at the distribution of words around them. Word2vec[14] is a widely adopted predictive embedding model that uses neural networks for learning the word embedding. Because embedding of rare words could be poorly estimated, the FastText model[15] has been proposed to address this issue by making use of subwords. This enhancement is particularly relevant to biomedicine as many rare words can be broken down into more commonly used subwords (eg, "deltaproteobacteria" is made up of the subwords "delto," "proteo," and "bacteria"). BioWordVec[16] is a word embedding model based on FastText and trained on unlabeled texts from biomedical literature combined with information from the MeSH (Medical Subject Headings) terminology. BioWordVec has achieved significantly better performance than other methods in several NLP tasks.

While BioWordVec generates the same embedding for a word regardless of context, contextualized word representations can generate different word embeddings for a word depending on context. The most recent incarnation of context sensitive embedding is BERT (Bidirectional Encoder Representations from Transformers),[17] a multilayer bidirectional transformer encoder that can learn deep bidirectional representations. BERT has shown promise in various NLP tasks. Because BERT is pretrained on general English corpora, attempts have been made to improve its performance in biomedicine by adding bio-

Table 1. Performance of the different measures on UMNSRS-Relatedness dataset

Semantic relatedness measurement	Spearman correlation
Shortest path	0.3093
Leacock-Chodorow	0.3093
cui2vec	0.4603
BERT-L	0.3120
BlueBERT_L	0.3707
BlueBERT-LE	0.3838
BioWordVec	0.5770
Graph Embedding (GCN)	0.3461
Graph Embedding (TransE)	0.1232
Graph Embedding (HolE)	0.3287
Graph Embedding (DistMult)	0.3205
Graph Embedding (CompIEx)	0.1965
BlueBERT_LE+Graph Embeddings	0.4095
BioWordVec+Graph Embeddings (GCN)	0.5904[a]
Cui2vec+BlueBERT-LE	0.3691
Cui2vec+Graph Embeddings (GCN)	0.4048

GCN: graph convolutional network.
[a]Best performing method.

Table 2. Performance of the different measures on UMNSRS-Similarity dataset

Semantic similarity measurement	Spearman correlation
Shortest path	0.3538
Leacock-Chodorow	0.3538
cui2vec	0.5411
BERT-L	0.3454
BlueBERT_L	0.4021
BlueBERT-LE	0.4169
BioWordVec	0.6182
Graph Embedding (GCN)	0.4037
Graph Embedding (TransE)	0.1883
Graph Embedding (HolE)	0.3949
Graph Embedding (DistMult)	0.3716
Graph Embedding (CompIEx)	0.3238
BlueBERT_LE+Graph Embeddings	0.4326
BioWordVEc+Graph Embeddings (GCN)	0.6288[a]
Cui2vec+BlueBERT-LE	0.5235
Cui2vec+Graph Embeddings (GCN)	0.5434

GCN: graph convolutional network.
[a]Best performing method.

medical texts in its training. BlueBERT from the National Center for Biotechnology Information is one such effort that utilized PubMed articles and clinical texts that achieved good results.[18,19]

Graph embedding

While deep learning effectively captures hidden patterns of Euclidean data, there is an increasing number of applications where data are represented in the form of graphs. Motivated by convolutional neural network (CNN), recurrent neural network, and auto-encoders from deep learning, new generalizations and definitions of important operations have been rapidly developed over the past few years to handle the complexity of graph data.[20,21] For example, a long short-term memory structure can be used to directly encode graph-level semantic information.[22] The topic of graph neural networks has received growing attentions recently.[23,24] A number of authors generalized well-established neural network models like CNN that apply to regular grid structure to work on arbitrarily structured graphs.[25] Kipf and Welling[26] first presented a simplified graph neural network model, called graph convolutional networks (GCNs), which are an efficient variant of CNN on graphs. GCNs have been an effective tool to create node embeddings that aggregate local information in the graph neighborhood for each node. GCN models can also impose the same aggregation scheme when computing the convolution for each node, which can be considered a method of regularization, and improves efficiency.[27] GCNs have achieved state-of-the-art results on a number of benchmark graph datasets[26,28] in various application areas, such as social networks[29] and natural language processing.[30]

However, most of the existing research on GCNs has focused on learning representations of nodes in simple undirected graphs. For more general and pervasive class of graphs, knowledge graph embedding is one of the recent active research areas. Many knowledge graph embedding approaches have been proposed. Translation-based approaches, such as TransE[31] and its variants, model relations as translating operations on the low-dimensional embedding of the entities. Semantic matching–based approaches, such as Holographic Embeddings model (HolE),[32] DistMult,[33] and its extension in the complex space (ComplEx),[34] compute the score of each triple via similarity-based score function.

MATERIALS AND METHODS

Generation of concept embeddings

Word embedding

For word embedding models, we used BioWordVec and 2 publicly available flavors of BERT, BERT-Large (BERT-L) and BlueBERT-Large (BlueBERT-L). Because BERT allowed users to enhance it by training with additional corpora, we created a third flavor, BlueBERT-LE, by enhancing BlueBERT-L by the concept definitions in the UMLS. We used the 2019AB version of the UMLS and extracted 283 491 English definitions from the MRDEF table. Overall, only 5.66% of UMLS concepts had definitions. We performed sentence segmentation on the definitions with the NLP toolkit spaCy (http://spacy.io). We then used the script provided by the BERT developers to enhance BlueBERT-L using the same vocabulary, sequence length, and other configurations provided by Devlin et al.[17] The script performed "masked language model" and "next sentence prediction" on the UMLS definition corpus. It masked out 15% of the words in the input, ran the entire sequence through a deep bidirectional transformer encoder, and then predicted only the masked words. It also trained the model to learn relationships between sentences using the actual next sentence that came after a sentence (a positive example), or just a random sentence from the corpus (a negative example). Note that the UMLS definitions were only used to enhance BlueBERT-L and not in the direct generation of the concept embeddings.

One common way of using BERT is to couple it with a fine-tuning step using training data for a specific task. However, it is also possible to use BERT alone to generate contextualized word embeddings without additional training data, which is how we used BERT in this study. To get the individual vectors for each token in a sentence, a pooling operation is needed to combine some of the layer vectors. Using appropriate pooling strategy (sum, mean, concatenation, etc.) and layers (last 4 hidden layer, last hidden layer, all layers, etc.) to derive a fixed sized embedding can yield results not far behind BERT with task specific fine-tuning.[17]

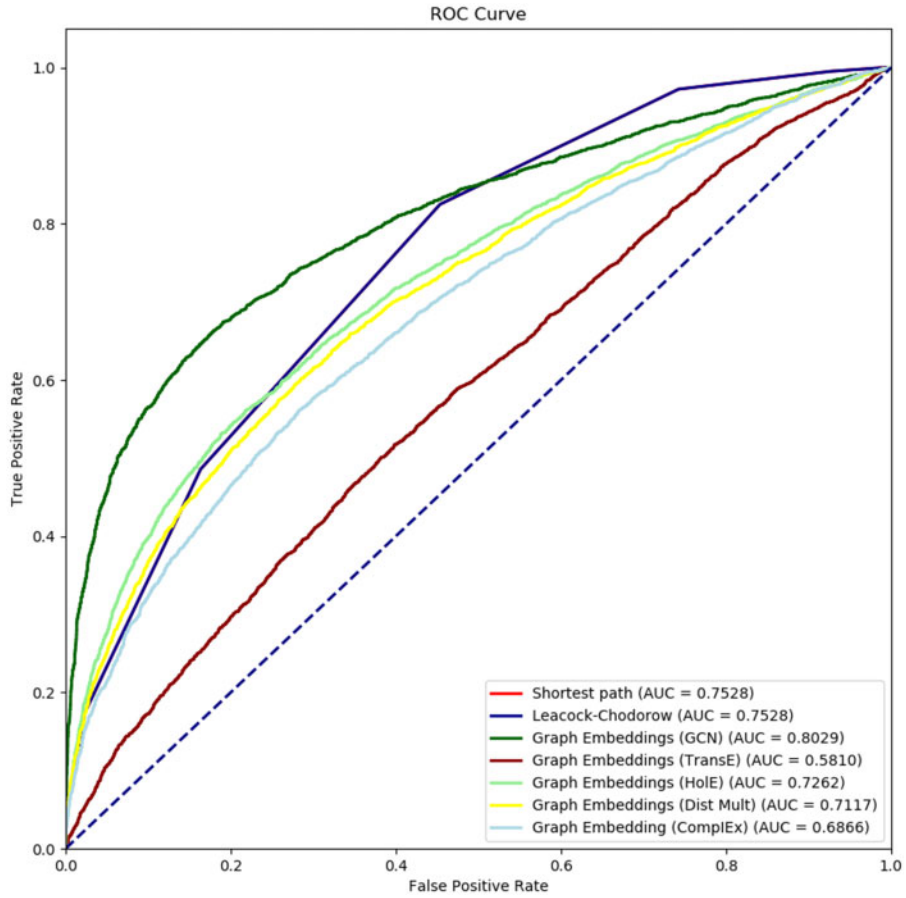

Figure 2. Receiver-operating characteristic curves of the different knowledge-based measures on the Standardized MedDRA Query dataset. AUC: area under the curve; GCN: graph convolutional network.

Table 3. Performance of the word embedding models in the MSH-WSD dataset

Embedding method	Accuracy
BERT-L	0.753
BlueBERT_L	0.796
BlueBERT-LE	0.805
BioWordVec	0.784
Majority sense (baseline)	0.549

MSH-WSD: MeSH Word Sense Disambiguation.

For each UMLS concept, we generated a "concept sentence" made up of the concept's English preferred term, followed by other English terms, excluding duplicate words and terms that were case or word order variants of the preferred term. Two special tokens were inserted—the classifier token [CLS] at the beginning, and the word "means" between the preferred term and other terms. For instance, the concept sentence for CUI: C0162275 is "[CLS] Ketonuria means Excess of ketones present in urine Acetonuria (finding) Ketoacidurias Ketonaciduria Ketonurias Acetonurias." The 3 BERT models were then applied to the concept sentence. We combined the embedding of the [CLS] token and the average embedding of each word in the preferred term to generate the concept sentence embedding. We used BioWordVec to generate word embeddings for each word in the concept sentence and took the average of all embeddings as the concept sentence embedding. There were 2 variations of BioWordVec (with window sizes of 2 and 20, respectively), suitable for either intrinsic or extrinsic tasks, and we used them accordingly (see Evaluation). In the set of UMLS concepts that we used in our evaluation, the number of tokens in the concept sentences ranged from 3 to 526, with an average of 69 tokens. There was an upper limit of 512 tokens for the BERT processor, and a small fraction (0.17%) of the concept sentences needed to be truncated.

Graph embedding

We used the GCN and 4 knowledge graph embedding approaches (TransE, HolE, DistMult, and CompIEx) to generate the graph embeddings. We represented the UMLS as graphs with concepts as nodes and relations as edges. We only used the hierarchical (parent-child) relations from SNOMED CT (Standardized Nomenclature for Medicine Clinical Terms) and MedDRA (Medical Dictionary for Regulatory Activities). We picked SNOMED CT and MedDRA because SNOMED CT was the most comprehensive clinical terminology in the UMLS and MedDRA was the source of the terms in one of our reference standards for evaluation. For MedDRA, we also included the "classified_as" relations. The "classified_as" relations in MedDRA were not, strictly speaking, parent-child relations. They represented the narrow-to-broad relationship between the lower level terms and preferred terms in MedDRA. We extracted hierarchical relations from SNOMED CT and MedDRA from the

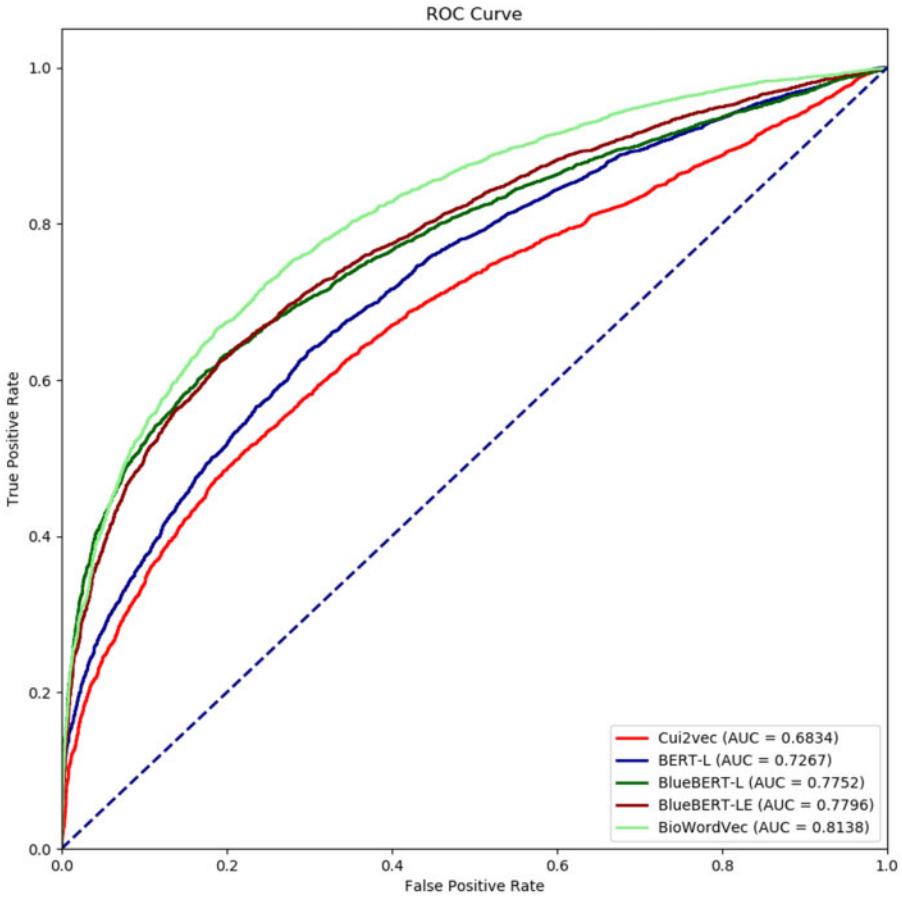

Figure 3. Receiver-operating characteristic curves of the different distributional-based measures on the Standardized MedDRA Query dataset. AUC: area under the curve.

MRREL table. The resulting graph contained 406 240 nodes and 899 151 edges.

The graph encoder conducted unsupervised learning for relationships, linking a prediction with the GCN-based Variational Graph Auto-Encoders model[35] or a knowledge graph embedding model by using the UMLS concepts and relations as input values. When a concept (node) was used as input to the pretrained graph embedding model, the model returned the relational learning representation as the embedding vector, based on the latent learning representation captured from the training graph data. For example, after performing forward propagation through the GCN, the embedding vector closely resembled the community structure of the graph. Therefore, the distance between the embedding vectors for the concepts "Influenza" and "Pneumonia" (which were siblings) was much smaller than the distance between the vectors for concepts "Influenza" and "Atherosclerosis" (separated by 5 hops). For all graph embedding models, we used the common settings suggested in the literature: the number of epochs was 200, the embedding dimension was 200, the optimizer was the Adam optimizer,[36] and the learning rate was 0.001. As in Bordes et al,[31] we used the margin-based pairwise ranking loss with regard to TransE, HolE, DistMult, and CompIEx. An overview diagram of our method is shown in Figure 1.

Evaluation

Embedding methods are generally evaluated by intrinsic or extrinsic tasks. Intrinsic tasks involve the measurement of semantic relatedness between words, sentences or concepts. Extrinsic tasks involve downstream text processing tasks such as information retrieval or word sense disambiguation (WSD). We evaluated our embeddings using both intrinsic (semantic relatedness between UMLS concepts) and extrinsic tasks (WSD), with comparison with other published methods.

Semantic relatedness measurement

We used 3 reference standards of semantic relatedness. The first was the manually annotated UMNSRS-Relatedness dataset, which consisted of 587 term pairs of UMLS concepts, with their corresponding relatedness scores manually judged by domain experts from the University of Minnesota Medical School.[37] After excluding obsolete concepts and concepts that were not applicable to some of the measurements, there were 473 pairs of concepts in this reference standard. The second was the UMNSRS-Similarity dataset, which consisted of 566 term pairs of UMLS concepts, with their corresponding similarity scores manually judged.[37] After excluding obsolete concepts and concepts that were not applicable to some of the measurements, there were 480 pairs of concepts in this reference standard.

The third reference standard was based on the Standardized MedDRA Queries (SMQs). The SMQs were created to improve adverse drug reaction signal detection by grouping together MedDRA terms that were related to a specific adverse reaction. We used the same method as[38] and randomly selected 5000 term pairs that

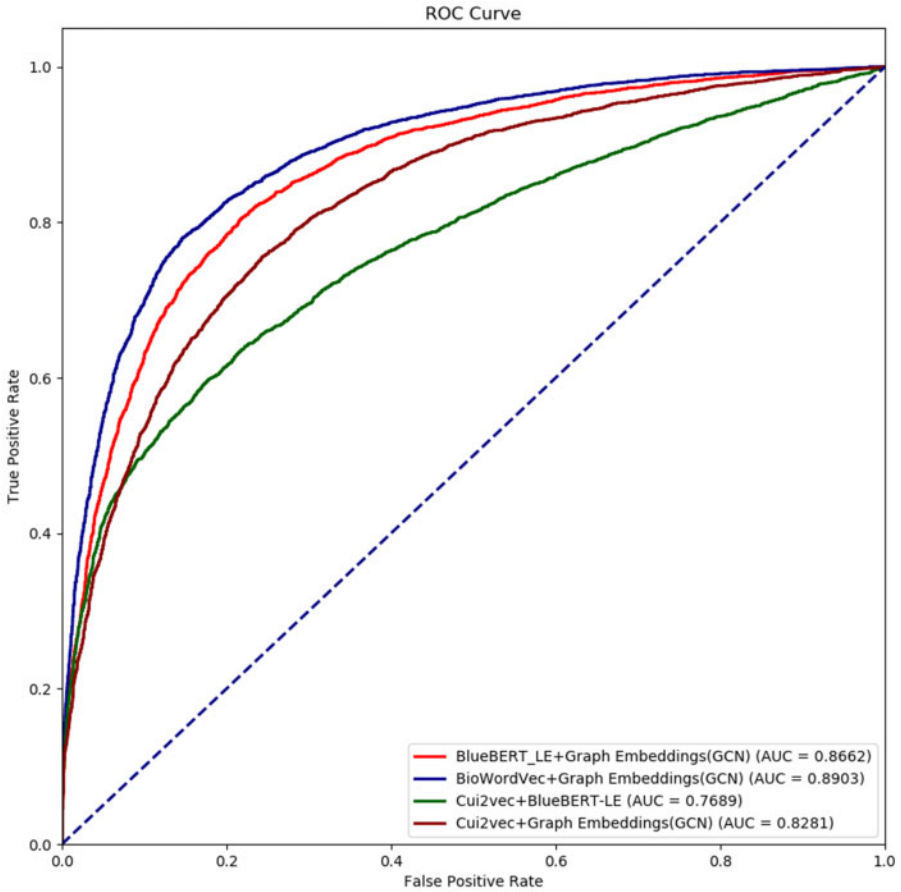

Figure 4. Receiver-operating characteristic curves of the different combined-embedding measures on the Standardized MedDRA Query dataset. AUC: area under the curve; GCN: graph convolutional network.

existed in the same SMQ category as positive examples and 5000 term pairs that existed in different SMQ categories as negative examples. Because we used the MedDRA relationships in our graph embeddings, which might confer unfair advantage, we also repeated the graph embeddings without the MedDRA relationships to see the difference.

For semantic relatedness scores, we computed the cosine between each pair of UMLS concepts' embedding vectors in the standard way using their dot product and magnitude. We used the concept sentence embeddings generated by BioWordVec and 3 flavors of BERT, the concept graph embeddings generated by GCN, and 4 knowledge graph embedding models individually. We also used the combination (concatenation) of the top-performing sentence and graph embeddings to see whether a combined approach would perform better.

For comparison with our concept embeddings, we used a set of publicly available UMLS concept embeddings called cui2vec that was generated based on large corpora of clinical and bibliographical data.[39] In cui2vec, the corpora were first normalized against the UMLS concepts containing SNOMED CT concepts, then the word embedding tool Word2vec was used to generate the embeddings for the UMLS concepts. Cui2vec provided concept embeddings for 108 477 UMLS concepts that could be recognized in the corpora. We used the cosine of cui2vec concept embeddings in the same way to measure semantic relatedness.

We also compared our results with 2 traditional knowledge-based (path-based) semantic relatedness measurements that did not involve deep learning: the shortest path measure[40] and Leacock-Chodorow measure.[41] The shortest path measure used the reciprocal of the shortest distance between 2 UMLS concepts. The Leacock-Chodorow method used the negative log of the shortest distance between 2 concepts divided by the depth of the path. The paths were calculated from the same graphs used by the graph embedding models.

To compare between various semantic relatedness measurements, for the UMN datasets, we calculated the Spearman correlation between the semantic relatedness scores generated by the various methods and the scores in the reference standards. For the SMQ reference standard, we used the area under the curve (AUC) from the receiver-operating characteristic plot.

Word sense disambiguation

We used the MeSH WSD (MSH-WSD) corpus[42] as reference standard. The MSH-WSD corpus consisted of 203 ambiguous terms (106 regular terms, 88 acronyms, 9 could be either), each term could be associated with multiple UMLS concepts (senses). For each of these concepts, up to 100 MEDLINE abstracts containing the ambiguous terms were retrieved. The challenge was to find the correct concept (sense) for each abstract.

For each abstract, we generated the word embeddings for the "subsentence" around the ambiguous term, with a window size of 6, using the 3 flavors of BERT and BioWordVec. For BERT, a single

embedding could be generated for the whole subsentence. For BioWordVec, we took the average of the embeddings for words constituting the subsentence. To find the correct UMLS concept, we picked the one with the highest cosine value between the subsentence embedding and the UMLS concept sentence embeddings generated by the same embedding method. We did not use graph embeddings in this task because there was no straightforward way to generate graph embeddings from the abstracts.

RESULTS

UMNSRS-Relatedness dataset

The results are summarized in Table 1. BioWordVec outperformed all other methods used individually. The combination of BioWordVec and graph (GCN) embeddings had the best performance overall. Cui2vec outperformed all BERT embeddings, but the combination of cui2vec with GCN resulted in worse performance. The performance of BlueBERT-LE was the best among the 3 flavors of contextual word embedding, and it was slightly better than graph embeddings. GCN was the best among the 5 graph embedding models, and was better than the 2 traditional path-based measurements.

UMNSRS-Similarity dataset

The results are summarized in Table 2. The Spearman correlations were higher for all methods compared with the UMNSRS-Relatedness dataset, but the overall trend was very similar. Unlike the relatedness dataset, cui2vec combined with GCN was better than cui2vec alone, but was still worse than BioWordVec. GCN performed better than BlueBERT-L in the similarity dataset but was still not as good as BlueBERT-LE.

SMQ dataset

The receiver-operating characteristic curves of the various measurements are shown in Figures 2-4. The curves of the 2 path-based measurements (purplish line in Figure 2) were overlapping completely because they generated the same results for most data points. Their curves were also more angular because many concept pairs had the same score. Similar to the UMN datasets, BioWordVec was the best performing single measurement. Unlike the UMN datasets, cui2vec's performance was the worst among all methods. GCN graph embedding outperformed the best BERT embedding (BlueBERT-LE), which in turn outperformed path-based measurements. Again, combined word and graph embeddings had the best performance overall.

Among the 10 000 randomly selected concept pairs (both positive and negative examples) used in our evaluation, only 6 pairs were connected by parent-child relationships in MedDRA. When we omitted MedDRA relations from the graph embeddings, the AUC of GCN dropped slightly from 0.8029 to 0.7960, which did not affect the overall order.

Word sense disambiguation

The accuracy scores obtained by our models using the different word embeddings are shown in Table 3. BlueBERT-LE and BlueBERT-L outperformed BioWordVec in the WSD task. All embedding methods significantly outperformed the majority sense baseline 0.549 for the MSH-WSD dataset, which was achieved by assigning the most frequent concept to every instance.[43]

DISCUSSION

In this study, we employed the latest deep learning techniques in word and graph embedding to generate semantic relatedness measurements between UMLS concepts. Our key requirements are the use of publicly available, off-the-shelf tools, and no additional resources except the UMLS. The best results were obtained by combined word and graph embeddings, which significantly outperformed some existing corpus-based concept embeddings and path-based measurements.

Our method has several advantages. First, on the one hand, it can be applied to all UMLS concepts because all UMLS have terms and relations, which are used to generate the concept sentence embeddings and graph embeddings. On the other hand, cui2vec only covers UMLS concepts that are recognized in their corpora, which amount to about 100 000 (2.3%) UMLS concepts. Second, our method does not involve resources outside of the UMLS. Compared with methods that rely on additional text corpora, ours is easier to implement and less demanding on processing power and time. In addition, different corpora may produce different results. Our method is solely based on the UMLS and the results should be generally reproducible—an essential requirement for benchmarking and comparative studies. Third, our method has uniformly good performance across various semantic relatedness datasets. As shown in our study, existing semantic-related measurements may vary considerably in their performance based on the task at hand. The UMN datasets consisted of many concept pairs that were not hierarchically related in existing terminologies (eg, pallor and iron), so path-based measurements and graph embeddings did not perform as well as our sentence embeddings or the corpus-based cui2vec. On the other hand, all the concepts from the SMQ dataset came from MedDRA and belonged to the same semantic type (eg, retinoblastoma and eye abnormalities). They were also more likely to be linked, directly or indirectly, by hierarchical relations (not only those from MedDRA). This explains why the path-based measurements and graph embeddings outperformed cui2vec and some sentence embeddings for the SMQ dataset. The BioWordVec sentence embeddings outperformed all others in all datasets, and could be further improved by combining with graph embedding. As an external reference, our best results for the semantic relatedness and semantic similarity datasets were 0.59 and 0.63, respectively, which compared favorably to Pakhomov et al,[10] who used embeddings learned from large biomedical corpora (the best results were 0.58 and 0.62, respectively). On the SMQ dataset, our best performance of 0.89 AUC was significantly better than that of Bill et al (0.827).[38] As for extrinsic tasks, our word embeddings performed reasonably well in the MSH-WSD dataset and the accuracy (0.753-0.805) was comparable to Pakhomov et al (accuracy 0.740-0.777).[10] Contrary to semantic relatedness, BERT outperformed BioWordVec in WSD, which was probably attributable to the context sensitive nature of BERT.

Our study shows that graph embedding is a better way to capture relational information between concepts compared with traditional path counting. While GCN consistently outperforms path-based measurements, the other graph embedding models are less impressive. Our results also show that it is possible to combine word and graph embeddings to enhance performance. Before graph embedding, mathematical and statistical operations on graphs are generally limited, and applying machine learning methods directly to graphs is challenging. Graph embedding transforms graphs into vectors, which are easier to work with in machine learning. This also opens up the possibility of combining graphical data with other

types of data. In our study, combined word and graph embedding uniformly outperforms other methods.

Similar to other studies, our study confirms that word embedding algorithms benefit from additional training with domain specific corpus. The National Center for Biotechnology Information's BlueBERT-Large, which was trained on biomedical texts, outperformed BERT-Large, which was trained on general English corpora. Further training of BlueBERT-Large with UMLS definitions yielded an additional improvement.

We acknowledge the following limitations in our study. In word embedding, the method to generate the concept sentence was empirically decided. We have not experimented with methods to leverage the other characteristics of UMLS terms, such as term types and the nature of the source terminologies (eg, purpose, specific domain). In graph embedding, we restricted our source relationships to hierarchical relationships from only SNOMED CT and MedDRA and did not experiment with the inclusion of more relation types from more sources. However, there is a limit to the graph size that most graph embedding methods can handle. We compared our results with only 2 path-based measurements and 1 set of corpus-based concept embeddings because they were publicly available and relatively straightforward to implement. For reference standards, we only picked the UMN and SMQ datasets, which were publicly available and relatively sizeable. The lack of large-scale reference standards is an often cited limitation in semantic relatedness research.

In the future, we plan to explore the use of additional information carried in the UMLS, and new tools for word embedding (eg, Clinical XLNET)[44] and graph embedding (eg, graph attention auto-encoder, adversarially regularized graph auto-encoder).[45,46] In addition, we would also like to experiment with novel methods to harness lexical and relational information from the UMLS (eg, retrofitting and extended UMLS definitions).[47,48]

CONCLUSION

Deep learning techniques, word and graph embedding, can be leveraged to measure semantic relatedness between UMLS concepts. By using public, off-the-shelf tools and no additional resources outside the UMLS, our methods outperformed some existing path-based measurements and corpus-based concept embeddings.

FUNDING

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

AUTHOR CONTRIBUTIONS

The authors do KWF and YM conceived the study together. YM implemented the various algorithms and analyzed the results. Both contributed to the writing of the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors do not have competing interests.

REFERENCES

1. 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.
2. Srihari R K, Zhang Z, Rao A. Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval* 2000; 2 (2/3): 245–75.
3. Stevenson M,Greenwood MA. A semantic approach to IE pattern induction. In: proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; 2005: 379–86.
4. Budanitsky A, Hirst GJCL. Evaluating wordnet-based measures of lexical semantic relatedness. *Comput Linguistics* 2006; 32 (1): 13–47.
5. Liu X-Y, Zhou Y-M, Zheng R.-S. Measuring semantic similarity in WordNet. In: proceedings of the 2007 International Conference on Machine Learning and Cybernetics; 2007.
6. Garla VN, Brandt CJBb. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC Bioinformatics* 2012; 13 (1): 261.
7. Aouicha MB, Taieb MAHJ. Computing semantic similarity between biomedical concepts using new information content approach. *J Biomed Inform* 2016; 59: 258–75.
8. Zhu G, Iglesias CA. Sematch: Semantic similarity framework for knowledge graphs. *Knowledge Based Syst* 2017; 130: 30–2.
9. Schütze H. Word space. In: proceedings of the 5th International Conference on Neural Information Processing Systems; 1992: 895–902.
10. Pakhomov SVS, Finley G, McEwan R, *et al*. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics* 2016; 32 (23): 3635–44.
11. Wang Y, Liu S, Afzal N, *et al*. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform* 2018; 87: 12–20.
12. McInnes BT, Pedersen T, Pakhomov SV. UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. *AMIA Annu Symp Proc* 2009; 2009: 431–5.
13. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (90001): D267–70.
14. Mikolov T, Sustskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: proceedings of the 26th International Conference on Neural Information Processing Systems; 2013: 3111–9.
15. Bojanowski P, Grave E, Joulin A, *et al*. Enriching word vectors with subword information. *Trans Assoc Comput Linguistics* 2017; 5: 135–46.
16. Zhang Y, Chen Q, Yang Z, *et al*. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data* 2019; 6 (1): 1–9.
17. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019: 4171–86.
18. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: oroceedings of the 18th BioNLP Workshop and Shared Task; 2019: 58–65.
19. Lee J, Yoon W, Kin S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36 (4): 1234–40.
20. Monti F, Boscaini D, Masci J, Rodolà E, Svoboda J, Bonstein MM. Geometric deep learning on graphs and manifolds using mixture model CNNs. In: proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017: .
21. Chen J, Zhu J, Song L. Stochastic training of graph convolutional networks with variance reduction. In: proceedings of the 35th International Conference on Machine Learning; 2018.
22. Song L, Zhang Y, Wang Z, Gildea D. A graph-to-sequence model for AMR-to-text generation. In: proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018.
23. Cai H, Zheng VW, Chang KC-C, *et al*., A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans Knowl Data Eng* 2018; 30 (9): 1616–37.

24. Battaglia PW, Hamrick JB, Bapst V, *et al*. Relational inductive biases, deep learning, and graph networks. *arXiv*:1806.01261; 2018.
25. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: proceedings of the 30th International Conference on Neural Information Processing Systems; 2016.
26. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv*:1609.02907; 2016.
27. Hamilton WL, Ying R, Leskovec J. Representation learning on graphs: methods and applications. *arXiv*:1709.05584; 2017.
28. van den Berg R, Kipf TN, Welling M. Graph convolutional matrix completion. *arXiv*:1706.02263; 2017.
29. Chen J, Ma T, Xiao C. Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv*:1801.10247; 2018.
30. Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. *Proc AAAI Conf Artif Intell* 2019; 33: 7370–7.
31. Bordes A, Usunier N, Garcia-Durán A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: proceedings of the 30th International Conference on Neural Information Processing Systems; 2013: 2787–95.
32. Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs. *Proc AAAI Conf Artif Intell* 2016; 30: 1955–61.
33. Yang B, Yih W-T, He X, Gao J, Deng L. Embedding entities and relations for learning and inference in knowledge bases. *arXiv*:1412.6575; 2014.
34. Trouillon T, Weibl J, Riedel S, Gaussier E, Bouchard G. Complex embeddings for simple link prediction. *Proc Mach Learn Res* 2016; 48: 2071–80.
35. Kipf TN, Welling M. Variational graph auto-encoders. arXiv preprint arXiv:1611.07308 2016.
36. Kingma DP, Ba J. *ADAM:* A method for stochastic optimization. *arXiv*:1412.6980; 2014.
37. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. *AMIA Annu Symp Proc* 2010; 2010: 572–6.
38. Bill RW, Liu Y, Mcinnes BT, Melton GB, Pedersen T, Pakhomov S.Evaluating semantic relatedness and similarity measures with standardized MedDRA queries. AMIA Annu Symp Proc 2012: 2012: 43
39. Beam AL, Kompa B, Schmaltz A, *et al*. Clinical concept embeddings learned from massive sources of multimodal medical data. *arXiv*:1804.01486; 2020.
40. McInnes BT, *et al*. Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. *AMIA Annu Symp Proc* 2011; 2011: 895–904.
41. Leacock C, Chodorow MJ. Combining local context and WordNet similarity for word sense identification. In: *Fellbaum C, Miller G, eds. WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press; 1998; 49 (2): 265.
42. Jimeno-Yepes AJ, McInnes BT, Aronson AR. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC Bioinformatics* 2011; 12 (1): 223.
43. McInnes BT, Pedersen T. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *J Biomed Inform* 2013; 46 (6): 1116–24.
44. Huang K, Singh A, Chen S, *et al*., Clinical XLNet: modeling sequential clinical notes and predicting prolonged mechanical ventilation. *arXiv*:1912.11975; 2019.
45. Salehi A, Davulcu H. Graph attention auto-encoders. *arXiv*:1905.10715; 2019.
46. Pan S, Hu R, Long G, Jiang J, Yao L, Zhang C. Adversarially regularized graph autoencoder for graph embedding. In: proceedings of the 27th International Joint Conference on Artificial Intelligence; 2018; 2609–15.
47. Yu Z, Wallace BC, Johnson T, *et al*. Retrofitting concept vector representations of medical concepts to improve estimates of semantic similarity and relatedness. *Stud Health Technol Inform* 2017; 245: 657–61.
48. Park J, Kim K, Hwang W, *et al*. Concept embedding to measure semantic relatedness for biomedical information ontologies. *J Biomed Inform* 2019; 94: 103182.