Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Representing document-level semantics of biomedical literature using pre-trained embedding models

Jon Stevens
AbbVie Information Research
jon.stevens@abbvie.com

Brandon Punturo
AbbVie Information Research,
University of Michigan School of Information
bpunt@umich.edu

Derek Chen
University of Michigan School of Information
dinganc@umich.edu

Mike Kim
University of Michigan School of Information
miketkim@umich.edu

Jacob Zimmer
University of Michigan School of Information
zimmerja@umich.edu

Abstract

We present two novel tasks aimed at capturing document-level semantics, i.e., high-level topical or thematic content, of biomedical scientific publications. We use these tasks to evaluate whether word and sequence embedding models pre-trained on biomedical literature can be used to derive meaningful document-level semantic representations for these publications. We evaluate approaches from two broad categories: **(1) lexical pooling**, or vectorizing documents purely based on aggregation of lexical items, which includes the NCBI's BioWordVec model and Tf-idf-based vectorizations, both with and without word pre-filtering based on biomedical ontologies, **(2) sequence embedding**, which includes the NCBI's BioSentVec model and BioBERT. For both of our tasks, lexical pooling outperformed sequence embedding, and the best overall method was mean pooling of BioWordVec word embeddings. We also include baselines trained on non-biomedical English to show that training on biomedical literature is warranted. The methods discussed here have potential applications for clustering, comparing, analyzing and recommending scientific literature in the biomedical domain.

1 Background

The last several years of NLP research have seen a number of breakthroughs leveraging the concept of transfer learning (Pan & Yang 2010), particularly in the form of pre-trained embedding models. Such models provide low-dimensional vectorized representations of text which are informed by large corpora but can be applied to small-data NLP tasks. Example architectures include, for static word embeddings, word2vec/skip-grams (Mikolov et al. 2013), GloVe (Pennington et al. 2014) and fasttext (Bojanowski et al. 2017), for sentence and paragraph embeddings, doc2vec (Le & Mikolov 2014) and sent2vec (Pagliardini et al. 2018), and for context-sensitive word and sequence embeddings, ULM-Fit (Howard & Ruder 2018) and BERT (Devlin et al. 2018), the latter adapting the transformer architecture of Vaswani et al. (2017).

These models have found use within the biomedical domain, particularly for processing scientific literature, where the NCBI's PubMed and MedLine databases provide a large, freely available data source for model training. Most recently, the NCBI has released two pre-trained embedding models, both trained on millions of biomedical abstracts and clinical notes (Chen et al. 2018): (1) BioWordVec, a static word embedding model based on fasttext, which uses character-level information to enhance word embeddings, particularly those of rare words, and (2) BioSentVec, a sentence

embedding model based on sent2vec, which learns to embed words and n-grams and average them to create a single sentence embedding, optimized on the task of predicting missing words.

Lee et al. (2019) have fine-tuned the base BERT model on millions of biomedical texts, including those from MedLine. This model, dubbed BioBERT, can also serve as both a contextual word embedding model and a document embedding model, if one pools the penultimate layer of transformer outputs (Xiao 2018).

Going beyond end-to-end embedding models, we may also incorporate BioNLP's long tradition of utilizing curated biomedical vocabularies and ontologies to extract insights from literature via text mining (see Fleuren & Alkema 2015 for an overview). In our case, we use ontologies to refine some models by giving higher weight to biomedically relevant terms in the text.

While great progress has been made in the evaluation of biomedical word embeddings (see e.g. Chiu et al. 2016; Wang et al. 2018), these evaluations have been aimed, naturally, at word- and phrase-level semantics, focusing on either word similarity or downstream tasks which do not require good representations of the overall thematic or topical content of each document. Moreover, intrinsic evaluations of embedding quality tend to rely on subjective scoring or ranking, where the assumptions about the semantic space into which the documents are embedded are unclear.

2 Overview

The aim of this paper is to compare different methods for using pre-trained models to created embedded representations of scientific publications from the MedLine database, and to evaluate their ability to capture document-level semantics in a useful way. To this end, we introduce two tasks that leverage document-level semantics: prediction of academic departments from MedLine titles/abstracts, and pairwise correlation of model-derived document similarity (measured as cosine similarity of the document embeddings) with document similarity derived from gold-standard Medical Subject Headings (so-called MeSH terms). On these tasks we compare methods derived from the BioWordVec, BioSentVec and BioBERT models, as well as n-gram Tf-idf vectorization and two embedding models pre-trained on general English. For the BioWordVec and Tf-idf methods we also evaluate the addition of biomedical ontologies to pre-select only words with biomedical relevance.

Having high-quality embedded document representations has a host of potential applications, both in the biomedical domain and in similar domains, including efficient document clustering, similarity scoring and the construction of knowledge graphs for easier discovery of scientific literature.

3 Vectorization Methods

The goal of each method is to produce a single vectorized representation of a MedLine document (here taken to be title + abstract text) which can be used to (a) determine similarity of two documents, and (b) serve as a featurization technique for machine learning models on small-data downstream tasks. To this end, we compare a number of methods belonging to two broad categories. The first is lexical pooling – documents are vectorized in a "bag of words" manner, by pooling vectorized lexical items. The lexical pooling methods tested are:

- **BioWordVec** pooling: for each word in the document text, obtain the BioWordVec embedding, then average pool all word embeddings to obtain a single document embedding
- **BioWordVec+**: BioWordVec pooling with word pre-filtering based on noun phrase dependency parsing and biomedical vocabularies (see Fig.1): pre-process the document text by extracting likely entities and biomedically relevant terms, and then pool only these word embeddings to create the document embedding. The pre-processing steps are as follows:[1]

1. Using established biomedical ontologies such as MedDRA and ChEMBL, extract all the phrases from the document which refer to concepts in these ontologies as well as the preferred names for any such concepts
2. Using a pre-trained English dependency parser, identify and extract all of the noun phrases in the document

[1] Scibite's TERMite platform was used to apply the ontologies, and the spaCy dependency parser was used to extract the noun phrases.

3. Concatenate the ontology phrases and concept names with the extracted noun phrases to create the input for vectorization

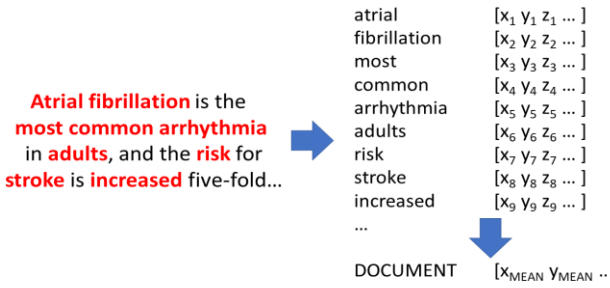

Figure 1: Illustration of BioWordVec pooling with entity extraction, one of the document vectorization methods we assess.

- High-dimensional **Tf-idf** vectorization of word and bigram tokens[2]
- **Tf-idf+**: Tf-idf with word pre-filtering based on noun phrase dependency parsing and biomedical vocabularies

As a non-biomedical comparison we pool **fasttext** embeddings pre-trained on Wikipedia and Common Crawl.

The second category is sequence embedding, where words and/or sequences or words are embedded in context:

- **BioSentVec** sentence pooling: pass each sentence in the document to BioSentVec, then average pool the resulting embeddings[3]
- **BioBERT**: pool transformer outputs from the penultimate layer of Lee et al. (2019)'s fine-tuning of BERT on biomedical literature

As a non-biomedical comparison we pool the penultimate layer of **BERT-base**.

4 Tasks and Data

For quantifying document semantics, the ideal embedding is one where the vector space represents a conceptual or thematic space that is anchored to identifiable concepts and topics in the relevant domains. That is, we want documents with similar embeddings to lie at similar points in a real-world conceptual space. Here we focus on two examples of such spaces: (1) MeSH terms (which include diseases, drugs, chemicals and many general topics and themes), and (2) at a coarser level of granularity, academic disciplines (e.g. cardiology, psychiatry).

MeSH headings provide a human-curated gold standard for medical publication semantics. Several approaches such as DeepMeSH (Peng et al. 2016) have been employed to try to solve the problem of automated MeSH indexing – many scientific articles lack MeSH annotations, either because they are not available on MedLine, or because they are too new to have been annotated. Rather than try to predict MeSH terms individually, we are using them as a gold-standard evaluation metric for our embedding methods, aimed at determining how well the vector space maps onto a known conceptual space in this domain.

To evaluate how well our document representations map onto the MeSH space, we vectorize the MeSH terms associated with a corpus of documents using inverse document frequency to penalize ubiquitous, less informative terms. Then, for random pairs of documents from that corpus, we correlate two metrics: (1) the cosine similarity of the MeSH vectors, and (2) the cosine similarity of our text-based vectors. For our corpus we randomly selected 10,000 pairs of documents (title + abstract) from MedLine. The intuition behind this intrinsic assessment is that the greater the correlation between the two similarity metrics being compared, the greater the extent to which those vectorizations encode the same conceptual space.

For our other task, we evaluate how well the methods do as text featurization methods for the task of learning to classify academic disciplines from document text. We have constructed a data set of over 2,000 recent faculty publications from the Zucker School of Medicine at Hofstra University, freely available at http://academicworks.medicine.hofstra.edu, along with the label of the department from which the publication originated. The department labels (e.g. cardiology, dermatology, neurology – 36 in all) serve as a proxy for academic

[2]In our initial tests we found that including bigrams slightly outperformed unigrams only across the board, and thus we only report these numbers.

[3] Being based on sent2vec, BioSentVec is trained to embed sentences, and therefore performs slightly better when sentences are pooled, compared to when the entire title + abstract are embedded directly.

	TF-IDF	TF-IDF+	BioWordVec	BioWordVec+	BioSentVec	BioBERT	fasttext+	BERT-base
Linear Discriminant Analysis	0.26	0.32	0.54	**0.56**	0.39	0.39	**0.47**	0.38
Linear SVC	0.55	**0.56**	0.52	0.54	0.46	0.33	0.41	0.42
Multiclass Logistic Regression	0.55	0.55	0.49	0.54	0.49	0.34	0.37	0.42

Table 1: Results on department classification task (weighted F1).

TF-IDF	TF-IDF+	BioWordVec	BioWordVec+	BioSentVec	BioBERT	fasttext+	BERT-base
0.04	0.20	**0.34**	0.31	0.27	0.07	0.10	**0.27**

Table 2: MeSH correlation results (Spearman's ρ)

discipline, a very coarse-grained measure of what the documents are about. The driving intuition is that the document embedding method that succeeds most in capturing document-level topical information should be better at separating out these classes. Each vectorization method serves as input to a number of classification models optimized on this task. The data set is relatively small, and thus the utility of transfer learning comes into play – we expect pre-trained models to succeed insofar as they add knowledge obtained from a much larger corpus. Moreover, if the densely embedded feature vectors succeed, they do so with much greater computational efficiency than Tf-idf, requiring learning from only a few hundred features, rather than over 100,000.[4]

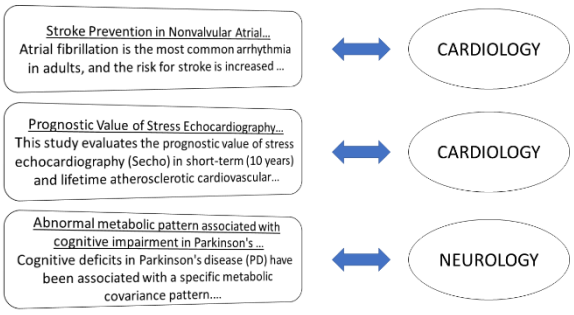

Figure 2: Example documents (left) and labels (right) from the Hofstra dataset.

5 Results

For the department prediction task, a number of classification layer architectures were tested using each of the five document vectorization techniques outlined above, of which the best performing models – linear support vector classifier (Tang 2013), multiclass logistic regression and linear discriminant analysis – are reported. Hyperparameters were optimized separately for each model and input type using random search.

The results are shown in Table 1. We see that the lexical pooling methods generally outperform the sequence embedding methods, with the best results coming from a linear SVC on Tf-idf+ document vectors and linear discriminant analysis on BioWordVec+ document vectors. The non-biomedical fasttext model does not perform as well as BioWordVec, but surprisingly, BERT-base does outperform BioBERT in two of three cases.

Results of the MeSH correlation are given in Table 2. Here BioWordVec is the "winner", i.e., these results suggest that these are the embeddings that best map onto the semantic space carved out by the expert-curated MeSH vocabulary. In this experiment, the ontology-based pre-filtering only introduced an advantage for the Tf-idf vectorizations. All correlations were statistically significant.

6 Discussion

Of the methods we compared, average pooling of biomedically trained word embeddings seems best suited to capture the document-level semantics of biomedical documents. BioWordVec+ performs similarly on the department classification task to its sparse Tf-idf counterpart, which performs surprisingly well and better than all the others. At the

[4] We plan to make both of our data sets available to the public. For questions about access, please contact jon.stevens@abbvie.com

same time, BioWordVec pooling yields the closest approximation to MeSH-based document similarity. The broader implication of this is that meaningful embedded representations of biomedical abstracts can be obtained by a simple averaging of word vectors, and that in some cases, improvement can be found by using biomedical ontologies and noun phrase parsing filter out irrelevant words. Our results also reinforce the notion that domain matters – pooling fasttext vectors trained on large amounts of non-biomedical English does not produce as good a result. Document embeddings for scientific literature have numerous practical applications in the biomedical domain, because they are easily obtained, information-dense representations that can be stored in a database, quickly retrieved, and used in document classification models, search and text mining systems, and article recommender systems. Some mysteries remain to be addressed by future work, such as the underperformance of BioBERT, in particular when compared to the BERT-base model.

Acknowledgments

The authors would like to thank Brian Martin, Kevin Chiou, Mehmed Sariyildiz and the rest of the RAIDERS team, Rob Gregg, Sajeev Cherian, Masha Trenhaile, Kamron Mehrayin, and Kevyn Collins-Thompson.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

Chen, Q., Peng, Y., & Lu, Z. (2018). BioSentVec: creating sentence embeddings for biomedical texts. arXiv preprint arXiv:1810.09302.

Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016). How to train good word embeddings for biomedical NLP. In Proceedings of the 15th workshop on biomedical natural language processing (pp. 166-174).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fleuren, W. W., & Alkema, W. (2015). Application of text mining in the biomedical domain. Methods, 74, 97-106.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 328-339).

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196).

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: pre-trained biomedical language representation model for biomedical text mining. arXiv preprint arXiv:1901.08746.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Pagliardini, M., Gupta, P., & Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 528-540).

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H., & Zhu, S. (2016). DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. Bioinformatics, 32(12), i70-i79.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Tang, Y. (2013). Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., ... & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. Journal of biomedical informatics, 87, 12-20.

Xiao, H. (2018). bert-as-service documentation. https://github.com/hanxiao/bert-as-service