Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210845

A Comparison of Word Embeddings to Study Complications in Neurosurgery

Gleb DANILOV[a,1], Konstantin KOTIK[a], Michael SHIFRIN[a], Yulia STRUNINA[a], Tatyana PRONKINA[a], Tatyana TSUKANOVA[a], Timur ISHANKULOV[a], Maria SHULTS[a], Elizaveta MAKASHOVA[a], Yaroslav LATYSHEV[a], Rinat SUFIANOV[a], Oleg SHARIPOV[a], Anton NAZARENKO[a], Nikolay KONOVALOV[a], and Alexander POTAPOV[a]
[a] *Laboratory of Biomedical Informatics and Artificial Intelligence, National Medical Research Center for Neurosurgery named after N.N. Burdenko, Moscow, Russian Federation*

Abstract. Our study aimed to compare the capability of different word embeddings to capture the semantic similarity of clinical concepts related to complications in neurosurgery at the level of medical experts. Eighty-four sets of word embeddings (based on Word2vec, GloVe, FastText, PMI, and BERT algorithms) were benchmarked in a clustering task. FastText model showed the best close to the medical expertise capability to group medical terms by their meaning (adjusted Rand index = 0.682). Word embedding models can accurately reflect clinical concepts' semantic and linguistic similarities, promising their robust usage in medical domain-specific NLP tasks.

Keywords. Neurosurgery, complications, NLP, word embeddings, clustering

1. Introduction

Word embeddings enable capturing useful semantic properties and linguistic relationships between words which might be important for information extraction, classification, and more complex natural language processing (NLP) tasks. In our opinion, word embeddings might help study distinct clinical concepts and discover relationships between them as soon as the underlying models accurately reflect clinical semantics. Thus, it seems reasonable to test this capability of word embeddings in modeling the known relationships between clinical concepts well-recognized by medical experts.

The specific domain we focused on was complications in neurosurgery. There are well-known types of complications accepted by many experts. However, the wide spectrum has not been described and agreed upon between neurosurgeons. Furthermore, the formal definitions of complications in neurosurgery are vague. Therefore, we hypothesized that word embeddings learned from narrative clinical notes can contribute to understanding the spectrum of complications in neurosurgery and a more rigorous definition of this concept. Our study aimed to compare the capability of different word

[1] Corresponding author, Gleb Danilov, N.N. Burdenko Neurosurgery Center, 4th Tverskaya-Yamskaya str. 16, Moscow 125047, Russian Federation; E-mail: glebda@yandex.ru.

embeddings to capture the semantic similarity of clinical concepts related to complications in neurosurgery at the level of medical experts.

2. Methods

To accomplish the research task, all unstructured textual data potentially containing the information on complications in neurosurgery were obtained from the electronic health records (EHR) of the National Medical Research Center of Neurosurgery named after academician N.N. Burdenko (Moscow, Russia) for the period between 2000 and 2017. The source documents reflected the initial assessment, past medical history, history of present illness, laboratory tests, physical and neurological examination, studies, medication, operative reports, daily notes, discharge summaries, etc. All the texts were typed in by doctors and other medical personnel on a keyboard. The corpus was preprocessed as follows: all the characters except for Cyrillic symbols and single spaces removed; texts tokenized with a space separator; stop-words, meaningless tokens (single letters, artifacts, etc.) and words occurred less than 6 times in the corpus eliminated; spelling corrected with the method we proposed in our previous work and tokens lemmatized (1). A medical expert then screened the resulted vocabulary of unique lemmas to select maximum terms potentially related to any adverse events (with broad inclusion criteria to capture diseases, symptoms, syndromes, accidents, medical errors, etc.).

All the words in the initial corpus were substituted by their lemmas to train word embeddings with Word2vec, GloVe, FastText, and pointwise mutual information (PMI) algorithms (2–5). When appropriate, we varied model type (CBOW/skip-gram), context window size (5-20), vector size (50-300), and the number of iterations over data across the models. The unprocessed clinical corpus was used to train RoBERTa (Robustly Optimized BERT Pretraining Approach) masked language model (6). It was trained during 5-10 epochs using base architecture. Different aggregated techniques were applied to get word embeddings of the vocabulary: mean average and maximum calculation of representations from the encoder-layers. The intersection of all sets of nouns showing positive cosine similarity with the word "complication" in every vector space obtained was further screened and labeled (when possible) by the type of clinical entity (symptom, syndrome, disease), body system, organ involved and ICD10 code for each term. A fully labeled subset of nouns was grouped by 4 aforementioned features to shape benchmark clusters. A k-means clustering algorithm was then applied to cluster each set of word embeddings with k equal to the number of benchmark clusters. Finally, we judged the clustering quality comparing to benchmark clusters using an adjusted Rand index.

The data were processed, and most word embeddings were learned within the R programming environment (version 4.0.3) in RStudio Server IDE (version 1.3.1093) using *tidyverse, tidytext, dplyr, Matrix, text2vec, word2vec, widyr, irlba, SnowballC, furrr and fossil* packages. FastText and RoBERTa vector representations were obtained with Python programming language (version 3.6.10) in Jupyter Notebook (version 6.1.4) using *fasttext* and HuggingFace *tokenizers* and *transformers* libraries.

3. Results

To create a clinical corpus, 588 text fields from 78 tables of the EHR database were identified. The corpus was compiled of 13 060 326 narrative text records containing data for 90 688 complete cases of neurosurgical treatment. Text preprocessing and tokenization produced 229 019 413 raw word tokens ending up with 40 121 unique lemmas. Of these, the expert selected 5 853 terms, potentially relevant for the concepts of complications/adverse events. A total of 84 vector spaces with different word embedding engines and varying learning parameters were obtained in the study. After finding the intersection of all sets of nouns showing a positive cosine similarity to the word "complication" in every vector space, it was possible to completely label 258 words, which were grouped into 40 benchmark clusters by 4 features (see the "Methods" section). The results of vector clustering with the k-means algorithm in benchmarking are shown selectively for 10 types of word embeddings in Table 1.

Table 1. Benchmarking of word embeddings clustering assessed with ARI. CBOW – continuous bag of words, SG – skip-gram, NI – number of iterations, ARI – adjusted Rand index.

	Model	CBOW/SG	Window size	Vector size	NI	ARI
1	FastText	skipgram	10	100	-	0.682
2	FastText	cbow	5	200	-	0.677
3	RoBERTa	-	-	-	-	0.330
4	GloVe	-	10	100	20	0.157
5	GloVe	-	10	300	20	0.116
6	PMI	-	10	100	-	0.081
7	Word2vec	cbow	10	300	10	0.013
8	Word2vec	skipgram	10	300	10	0.013
9	Word2vec	cbow	10	100	10	0.005
10	Word2vec	skipgram	10	100	10	0.005

Figure 1 shows a word cloud of medical terms semantically similar to the word “complication” in a high dimensional space produced by the best FastText model in our experiment (ARI = 0.682) and projected in 3-dimensional space by the t-SNE algorithm (perplexity = 8, learning rate = 10) with TensorBoard Embedding Projector. All the terms in Russian were automatically translated with the https://translate.yandex.ru/ service for international presentation purposes. Some of the terms containing misspellings were transliterated. The best word embedding approach demonstrates a reasonable spatial distribution of the related concepts that occur in the context of the "complication" term.

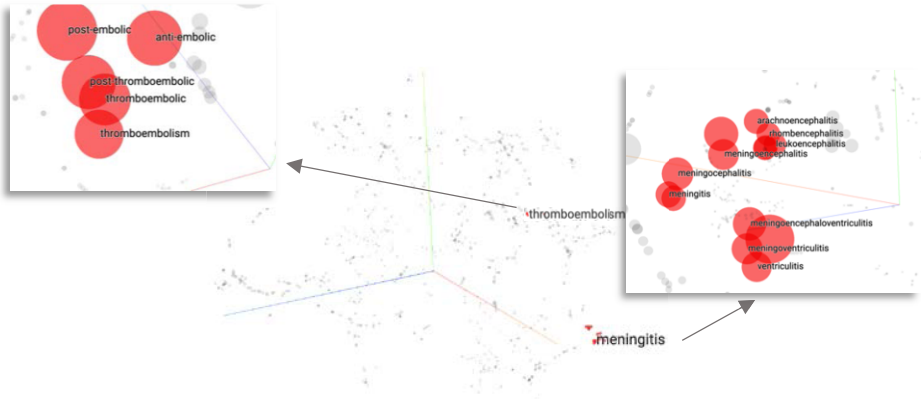

Figure 1. Word clusters of complications in neurosurgery derived from high dimensional FastText word embeddings and represented in 3 dimensions by the t-SNE algorithm with TensorBoard Embedding Projector (http://projector.tensorflow.org/). For example, word clusters for intracranial inflammatory complications are scaled right, and thromboembolism is shown left.

4. Discussion

In our study, models leveraging sub-word information from morphologically rich Russian language performed better compared to those treating words as atomic units. Interestingly, the BERT-based model demonstrated worse results than FastText in our domain-specific task, possibly due to the isolation of words from their contexts. Source words misspellings, expert-dependent benchmark cluster labeling, and a fixed set of models might be the limitations of our study. Generally, our results support those of the authors from other medical domains (7). Y. Wang et al. (2018) showed that word embeddings trained from EHR and medical literature can capture the semantics of medical terms better, and find semantically relevant medical terms closer to human experts' judgments than those trained from general domain data (8). The authors also importantly concluded that no global ranking of word embeddings for all biomedical NLP applications exists (8).

5. Conclusion

Word embedding models can accurately reflect clinical concepts' semantic and linguistic similarities, promising their robust usage in medical domain-specific NLP tasks. *This project was supported by the RFBR grants 18-29-01052 (data preprocessing) and 18-29-22085 (adverse events clustering).*

References

[1] Danilov G, Shifrin M, Strunina U, Pronkina T, Potapov A. An Information Extraction Algorithm for Detecting Adverse Events in Neurosurgery Using Documents Written in a Natural Rich-in-Morphology Language. Stud Health Technol Inform. 2019 Jul;262:194–7.

[2] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings. International Conference on Learning Representations, ICLR; 2013.

[3] Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. 2014. p. 1532–43.

[4] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Trans Assoc Comput Linguist. 2017 Jul;5:135–46.

[5] Silge J. Tidy word vectors, take 2! [Internet]. 2017 [cited 2020 Aug 20]. Available from: https://juliasilge.com/blog/word-vectors-take-two/

[6] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019 Jul; Available from: http://arxiv.org/abs/1907.11692

[7] Kalyan KS, Sangeetha S. SECNLP: A survey of embeddings in clinical natural language processing. Vol. 101, Journal of Biomedical Informatics. Academic Press Inc.; 2020. p. 103323.

[8] Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform. 2018 Nov;87:12–20.