Semantic Classification for Diagnostic Decision Support

Jalyoung Joe (jalyoung.joe@gmail.com)
Rosalind Franklin University of Medicine and Science

Research Article

Keywords: clinical decision support, natural language processing, transformer, universal sentence encoder, semantic classification, amyotrophic lateral sclerosis, monomelic amyotrophy, inclusion body myositis

Posted Date: June 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-610851/v1

Abstract

Background: Clinical history and exam findings obtained during the initial evaluation of a patient presenting with neuromuscular complaints are often pivotal in the formulation of diagnostic and therapeutic approach. However, disorders that have diverse overlapping manifestations may pose a diagnostic challenge for clinicians. Although case reports can be valuable sources for guiding clinical decisions, retrieval of applicable information from them is not straightforward.

Methods: In this paper, we propose a feature-extraction-based method to improve the efficacy and effectiveness of this process. The accuracy of the method in linking clinical presentations to correct diagnoses is examined using 30 case reports of amyotrophic lateral sclerosis, monomelic amyotrophy, and inclusion body myositis, obtained from PubMed Central Open Access Subset.

Results: The results show that feature extraction not only augments semantic classification with explainability but also improves the performance.

Conclusions: If developed further, the approach can be used to provide clinicians with decision support in challenging clinical situations.

Introduction

Medicine encompasses rare disorders with diverse and overlapping clinical manifestations that necessitate use of differential diagnostic approach. An example is amyotrophic lateral sclerosis (ALS), a heterogeneous disease characterized by a progressive degeneration of upper and lower motor neurons that can lead to a variety of symptoms [1]. There are a number of disorders with different treatment options and prognostic outcomes that can mimic amyotrophic lateral sclerosis [2,3]. Yet, in the absence of a specific biomarker for the disease, ALS remains a clinical diagnosis [4]. Although diagnostic errors may have disastrous consequences, over 40% of patients with amyotrophic lateral sclerosis are initially diagnosed incorrectly [5].

Case reports can provide clinicians with a rich representation of the spectrum a disorder can manifest throughout its clinical course, and thereby serve as a valuable source of guidance for clinicians in face of challenging cases [6]. Clinicians may use case reports during the formulation of differential diagnosis to identify and differentiate potential causes that can account for the patient's symptoms. However, retrieval of pertinent information from narrative texts that comprise case reports through manual inspection of individual reports can be a prohibitively time-consuming task for busy clinicians.

Advancements in natural language processing may greatly increase the clinical utility of case reports by facilitating information extraction from text. One of the recent breakthroughs in natural language processing was the introduction of transformer-based natural language models in 2017 [7]. Among the many language processing tasks with potential clinical utility that transformer-based models have excelled at is that of semantic classification [8]. The sentence embeddings in the form of fixed-length

vectors innately facilitate the text comparison process, and similarities between texts can be trivially obtained by applying distance metrics directly to the embeddings [9].

However, mere vector similarity between the text embeddings has limited applicability in the clinical context. Clinical reasoning involves attending to key differentiating elements to draw possible explanations for a given clinical presentation in accordance with contemporary clinical criteria [10]. The text similarity methods that do not distinguish features of importance from the rest and do not offer any explanations are difficult to be used in clinical context.

Although language models can be fine-tuned for individual target tasks to overcome such limitations, it is not possible to develop models for every subfield of medicine. Medicine is a rapidly evolving field comprised of distinct sub-fields each with vastly different terminologies and focuses. At present the performance of the transformer models can be measured only empirically, and studies show that fine tuning of general language models such as BERT does not necessarily improve their performance on target tasks [11].

In this paper, it is demonstrated that feature extraction using vector difference between Universal Sentence Encoder (USE) embeddings can not only augment semantic classification with interpretability but also improve the classification accuracy.

Methods

3.1 Data

The data for the present study come from PubMed Central Open Access Subset, which is a collection of articles that are available under a Creative Commons or similar licenses [12]. The first step in obtaining the data was to search the subset for relevant case reports. Keywords and filters described in Table 1 were used to collect scientific articles containing clinical descriptions of amyotrophic lateral sclerosis (ALS), monomelic amyotrophy (MMA), and inclusion body myositis (IBM). The articles were manually inspected to identify case reports that contain adequate details about the clinical course. Once the case reports had been obtained, the clinical history and exam findings were extracted from the cases. A total of 30 cases were used for analysis described in the following section.

3.2 Analysis

For analysis, USE [13] trained with transformer model was used to obtain embeddings of the text data [14]. To examine whether informative features can be extracted from the embeddings using vector subtraction, cosine similarities between the following pairs were measured:

1) Difference between a pair of sentence embeddings that contain a discriminatory feature of interest.

2) Average of embeddings for each clinical disorder.

The results are displayed in Table 2.

3.3 Evaluation

To examine whether feature extraction can be used to improve accuracy of semantic classification, the USE embeddings of clinical cases were projected to a 4-dimensional feature vector subspace using the sentence pairs shown in Table 2. The performance of a nearest mean model using the feature vectors for classifying case reports by diagnosis was compared to that using that using the raw USE embeddings with k-fold cross validation (k = 5), and the results are shown in TABLE 3.

Results

Table 2 shows interesting results. For instance, although the description of weakness was highly variable and irregular across the case reports, USE was able to capture the difference in distribution of the weakness between the disorders within the embeddings. The results verify that differences between sets of documents can be extracted using vector subtraction.

The results shown in Table 3 show that feature extraction can be used on the embeddings to improve accuracy of semantic classification. The improvement may be attributed to the removal of redundant dimensions in the original embeddings that encode nondiscriminatory features.

Discussion

In this paper, we demonstrated that features of interest can be extracted from USE embeddings and used to improve the performance of semantic classification. The idea behind the proposed method is to extend and take advantage of the generalizability of transformer-based models that is pretrained on massive text corpora. The proposed method is simple and flexible. It can easily be applied to a range of clinical tasks other than formulation of differential diagnosis. As the feature vectors encode each clinical element with simple numerical values that are easily interpretable, the method may also be applied to clinical text summarization tasks. The method may in the future potentially be applied to electronic health records by healthcare organizations to identify and track hidden trends and biases that are difficult to analyze using conventional data analysis methods.

Conclusion

Features of interest can be extracted from sentence embeddings with simple vector arithmetic operations to provide clinicians with decision support in challenging clinical situations.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The Python script and the data used for this study will be made available by the author (https://github.com/pushkin05/bert).

Competing interests

The author has no competing interests as defined by BMC, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Funding

The author received no specific funding for this work.

Authors' contributions

J.J. conceived of the presented idea, designed the study, performed the experiments, analyzed the data, and wrote the manuscript.

Acknowledgements

The author would like to thank the reviewers for their invaluable comments and suggestions.

References

1. Hardiman, Orla, Ammar Al-Chalabi, Adriano Chio, Emma M. Corr, Giancarlo Logroscino, Wim Robberecht, Pamela J. Shaw, Zachary Simmons, and Leonard H. Van Den Berg. "Amyotrophic lateral sclerosis." *Nature reviews Disease primers* 3, no. 1 (2017): 1-19.
2. Cortese, Rosa, Simonetta Gerevini, Franca Dicuonzo, Stefano Zoccolella, and Isabella Laura Simone. "Hirayama disease: the importance of an early diagnosis." *Neurological Sciences* 36, no. 6 (2015):

1049.

3. Dabby, Ron, Dale J. Lange, Werner Trojaborg, Arthur P. Hays, Robert E. Lovelace, Thomas H. Brannagan, and Lewis P. Rowland. "Inclusion body myositis mimicking motor neuron disease." *Archives of neurology* 58, no. 8 (2001): 1253-1256.
4. Hardiman, Orla, Leonard H. Van Den Berg, and Matthew C. Kiernan. "Clinical diagnosis and management of amyotrophic lateral sclerosis." *Nature reviews neurology* 7, no. 11 (2011): 639-649.
5. Belsh, Jerry M., and Philip L. Schiffman. "Misdiagnosis in patients with amyotrophic lateral sclerosis." *Archives of internal medicine* 150, no. 11 (1990): 2301-2305.
6. Nissen, Trygve, and Rolf Wynn. "The clinical case report: a review of its merits and limitations." *BMC research notes* 7, no. 1 (2014): 1-7.
7. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).
8. Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang. "How to fine-tune BERT for text classification?." In *China National Conference on Chinese Computational Linguistics*, pp. 194-206. Springer, Cham, 2019.
9. Chandrasekaran, Dhivya, and Vijay Mago. "Evolution of Semantic Similarity—A Survey." *ACM Computing Surveys (CSUR)* 54, no. 2 (2021): 1-37.
10. Cunha, Burke A. "The master clinician's approach to diagnostic reasoning." The American journal of medicine 130, no. 1 (2017): 5-7.
11. Lin, Chen, Steven Bethard, Dmitriy Dligach, Farig Sadeque, Guergana Savova, and Timothy A. Miller. "Does BERT need domain adaptation for clinical negation detection?." *Journal of the American Medical Informatics Association* 27, no. 4 (2020): 584-591.
12. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
13. https://tfhub.dev/google/universal-sentence-encoder-large/5
14. Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant et al. "Universal sentence encoder." *arXiv preprint arXiv:1803.11175* (2018).
15. Hirayama, Keizo. "Juvenile muscular atrophy of distal upper extremity (Hirayama disease)." *Internal medicine* 39, no. 4 (2000): 283-290.
16. Kiernan, Matthew C., Steve Vucic, Benjamin C. Cheah, Martin R. Turner, Andrew Eisen, Orla Hardiman, James R. Burrell, and Margaret C. Zoing. "Amyotrophic lateral sclerosis." The lancet 377, no. 9769 (2011): 942-955.
17. Needham, Merrilee, and Frank L. Mastaglia. "Sporadic inclusion body myositis: a review of recent clinical advances and current approaches to diagnosis and treatment." *Clinical Neurophysiology* 127, no. 3 (2016): 1764-1773.

Tables

TABLE 1. Keywords and filters used for search of PubMed Central Open Access Subset for case reports of amyotrophic lateral sclerosis, monomelic amyotrophy, and inclusion body myositis

Keywords and filters	PMCIDs of case reports used in this study
cc by license[filter] AND (lateral sclerosis[title]) AND (case[title])	PMC7253073, PMC6699726, PMC5762189, PMC5513085, PMC5372255*, PMC5116211*, PMC4602962, PMC3541770, PMC3240922
cc by license[filter] AND (hirayama[title])	PMC7460763, PMC7441664, PMC7310178, PMC6528527, PMC5715571, PMC5209606, PMC3625547
cc by license[filter] AND (monomelic amyotrophy[title])	PMC4794906
cc by license[filter] AND (inclusion body[title])	PMC7674586, PMC7402895, PMC7493364, PMC7017927, PMC4533788*, PMC6649988, PMC5420916, PMC4574179, PMC3623469, PMC3832753
* The article contains more than one clinical case.	

TABLE 2. Cosine similarity between difference vectors and average embeddings

Sentence 1	Sentence 2	ALS	MMA	IBM
a = “young’	a = “elderly”	-.09	.03	-.10
b = “in the hands”	b = “in the legs”	.00	.11	-.12
b = “and dysphagia”	b = “without dysphagia”	.00	-.04	.01
b = “and fasciculations”	b = “without fasciculations”	-.03	-.07	-.02
Sentence = A/an {a} patient with weakness {b}.				

TABLE 3. Comparison of performance of nearest mean models for classification using k-fold cross-validation (k = 5)

	TP	FP	TN	FN	F1
Original embeddings	16	14	46	14	0.53
Feature vector	22	8	52	8	0.73