Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220719

Explaining Contextualized Word Embeddings in Biomedical Research – A Qualitative Investigation

Marko MILETIC[a] and Murat SARIYAR[a,1]
[a] *Bern University of Applied Sciences, Switzerland*

Abstract. Contextualized word embeddings proved to be highly successful quantitative representations of words that allow to efficiently solve various tasks such as clinical entity normalization in unstructured texts. In this paper, we investigate how the Saussurean sign theory can be used as a qualitative explainable AI method for word embeddings. Our assumption is that the main goal of XAI is to produce confidence and/or trust, which can be gained through quantitative as well as quantitative approaches. One important result is related to the fact that the differential structure of language as explained by Saussure corresponds to the possibility of adding and subtracting word embeddings. On the other hand, these mathematical structures provide insights into the inner workings of natural language.

Keywords. Explainable AI, NLP, attention, BERT, saussure

1. Introduction

Word embeddings proved to be highly successful quantitative representations of words that allow to efficiently solve various tasks such as text categorization, and programming chatbots [1,2]. Unlike sparse vector, which are computed, for example, based on one-hot encoding for word co-occurrence matrices, embeddings are dense short vectors with a dimension ranging from 50 to 1000 [3]. These dimensions require interpretation to infer meanings from them. One most fascinating property is that the addition and subtraction of these vectors are semantically meaningful, e.g., the vectors for France + capital city yield the vector Paris, even though there seem to be no full justification for that fact [4].

An important differentiation is between methods for static and for dynamic contextual embeddings. Whereas the former ones assume a fixed context, e.g., Word2Vec [5], the latter ones generate different embeddings according to the different contexts of the word, e.g., Bidirectional Encoder Representations from Transformers (BERT) [6]; hence, it is primarily the context that is encoded, which allows for word sense disambiguation. Self-supervision is essential in most approaches: an artificial task is solved, e.g., a word c that occurs near the of the target word w acts as the one to be predicted by w and the learned weights in this task are the word embeddings. How to explain that these word embeddings capture semantic content with low-dimensional vectors consisting of real values, which on their own have no clear interpretation?

[1] Corresponding Author, Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

Explainable Artificial Intelligence (XAI) is the subject field that provides methods and tools to gain insight into the mechanism and results of machine learning methods [7]. Usually, quantitative methods, such as the Local Interpretable Models (LIME [8]), are used. Here, we will use qualitative approach: the linguistic structuralism of Saussure [9]. It describes language as the effect of differences between signs. These signs are composed of a signifier (sound-image) and a signified (reference or concept). According to Saussure, the visible side of a sign (the individual signifier) is arbitrarily connected to a signified and responsible for its meaning through placement in a complex system of signification. For instance, the signifier "tree" for the concept 'tree' conveys its meaning only because of its differences to other signifiers such as "bush", "leave", "branch", "trunk", and so on. Translations between languages can only be accurate if these differences are isomorphic between the two languages. The theoretical congruence with word embeddings seems obvious. In the following, we will investigate the benefits of using Saussurean sign theory for explaining the mechanism of word embeddings.

2. Methods

We assume that the main goal of XAI is to produce confidence and/or trust. Trust is a reliance on one's own expectations with respect to certain mechanisms in view of the involved risks and alternatives. Confidence on the other hand is a reliance without taking alternatives or the risks into consideration [10]. In case of contextualized word embeddings for clinical texts, it will have a relevant impact on the acceptability of an AI system, whether confidence or trust can be achieved. Clinicians often want to be able to assess the risk of unintuitive and wrong classifications of the underlying AI algorithm. Table 1 from [11] will be applied to address the research question stated above by sequentially investigating the listed goals of XAI. As a method, we will refer to BioBERT [12], which is an attention-based and contextualized language representation model, pre-trained on general & biomedical domain corpora. The pre-trained model provides contextual word embeddings for words as well as sub-words (allowing to recognize new words), which can be fine-tuned for specific tasks.

Table 1. Properties of different XAI goals with one example for each of the goals.

Goals of XAI	Black box	Alternatives	Examples
Confidence	yes	Not Considered	"The system is already in use in hospital X"
Layman trust	Opened slightly	Considered	List of features that have significant impacts (LIME)
Expert trust	Opened fully	Considered fully	Results of relevance propagation for deep nets

3. Results

Using or producing word embeddings is state-of-the art in natural language processing. This fact alone could generate some confidence but should be enriched with the description of concrete use cases, in which high accuracy values are achieved. One prominent example in medicine is entity normalization (EN) with SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms), which is the task of mapping a named entity mentioned in a clinical text to a corresponding SNOMED-CT code. For example, Viral pneumonia (disorder) is mapped to the ID: 75570004. Using the BioBERT model, state-of-the-art performance for biomedical EN is achieved by relying on contextualized word embeddings [3]. In contrast to static word embeddings, they allow for word sense disambiguation. For example, the sentence "The viral pneumonia of her cat had no effect on the patient" should not lead to a code mapping, which BioBERT is able to avoid. This dependence of word meanings on the context is nothing else as the differential structure of the language described by Saussure (see the tree-example in Section 1), miraculously captured by numeric vectors produced by deep neural networks. Due to the dynamic differential structure of language, it is also possible that a word is replaced with a completely different one, which just assume the same differential structures and leads to (nearly) the same overall meaning of a sentence, e.g. "The viral pneumonia of her **bumblebee** had no effect on the patient". A static word embedding cannot learn that this kind of replacement is not always possible.

For gaining layman trust, feasible alternatives should be described together with some high-level explanations of the inner workings of BioBERT. MedCAT is an example for an alternative of contextualized word embeddings. In its self-supervision, it uses the unique names in SNOMED CT (one of the names assigned to each concept should be unique), find those names in the corpus, learns the embedding together with its context (choosing a certain window), and is then applied by comparing the resulting contextualized word embedding with the word embedding of new concept candidates [12]. Such a contextualization is rather limited. In contrast to that, BioBERT uses the masked-language and next-sentence prediction tasks of BERT together with huge amount of data. For the former task, 15% of all words in the corpus are replaced by the string "[MASK]" and the neural network must find the correct word (some subtleties are ignored here). For the next-sentence prediction, in 50% of all sentence pairs the second sentence is replaced with a random one, and the neural network must predict the right sequence. These two tasks reflect much more how the differential structure of language is working according to Saussure: by a combination of paradigmatic selection (which signifier?) and a syntagmatic combination (how to proceed?). For instance, BioBERT learns that "The viral pneumonia of her cat had no effect on the patient" should not be followed by "She wants to marry the doctor to regain her health in one week".

For an expert, it is important to describe the inner mechanisms BioBERT in detail, which is beyond the scope of this paper. However, we will provide a hint how one might approach the task of gaining expert trust by referring to the attention mechanism in BioBERT. Attention is applied to word embeddings and is used in three places: in the encoder (input sequence pays attention to itself), in the decoder (the target sequence pays attention to itself) and in the encoder-decoder (the target sequence pays attention to the input sequence). Focusing on the encoder part: self-attention means that a sequence of word embeddings is compared with itself to learn how they should be adapted to capture the meaning of the whole sequence, consisting of 512 tokens. Due to the multi-head attention, several aspects of the context can be learned in parallel, not only with respect to the meaning, but also to the grammar of the sequence. In each head, three weight matrices are learned: query weights (Q), key weights (K) and value weights (V). These are used to learn how to adapt the word embeddings. Based on the context, Q decides on what to focus (left), K decides how to do the comparison with the other tokens (right), and V is responsible for the magnitude of changes of the word embedding values. It is just their (differential) positions in the respective formula that give these matrices such roles. From a Sausserean point of view, explaining what Q, K, V corresponds to is on the same level as explaining why the signifier "tree" is used rather than "okul" for denoting trees. One can point to the whole differential structure that is provided by the huge corpus for learning the BioBERT model, but a thorough understanding seems rather difficult.

4. Discussion

Coming back to our research question: the theory of Saussure provides aspects that allow to reflect the quantitative results of contextualized word embedding in a qualitative way. The differential structure of language and its complexity seem to be representable via dense numerical vectors that are produced by self-supervision. On the other hand, these mathematical structures provide insights into the inner workings of natural language. However, as the word "reflection" indicates, the theory of Saussure does not explain why contextualized word embeddings and approaches based on them perform so good. It can be used to gain confidence and layman trust, but expert trust often requires more. Opening the black box of deep learning methods is a subject field for itself, which primarily must utilize quantitative methods to digest the huge amount of information that is captured in the trained neural networks. The solutions of quantitative XAI could then be assessed and enriched by qualitative approaches.

References

[1] Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. Journal of Biomedical Informatics 2019;100:100057.

[2] Pola S, Sheela Rani Chetty M. Behavioral therapy using conversational chatbot for depression treatment using advanced RNN and pretrained word embeddings. Materials Today: Proceedings 2021; in press.

[3] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234–1240.

[4] Lavrač N, Podpečan V, Robnik-Šikonja M. Text Embeddings, in: N. Lavrač, V. Podpečan, and M. Robnik-Šikonja (Eds.), Representation Learning: Propositionalization and Embeddings. Cham: Springer International Publishing 2021;55–82.

[5] Di Gennaro D, Buonanno A, Palmieri FAN. Considerations about learning Word2Vec. Journal of Supercomputing 2021;77:12320–12335.

[6] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, and T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA. Association for Computational Linguistics 2019;4171–4186.

[7] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9:e1312.

[8] Stiglic G, Kocbek P, Fijacko N et al. Interpretability of machine learning-based prediction models in healthcare. WIREs Data Mining and Knowledge Discovery 2020;10:e1379.

[9] De Saussure F. Course in General Linguistics. Columbia University Press 2011.

[10] Luhmann N. Trust and Power. Chichester: Wiley 2017.

[11] Sariyar M, Holm J. Medical Informatics in a Tension Between Black-Box AI and Trust. Stud Health Technol Inform. 2022;289:41–44.

[12] Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, Mascio A, Zhu L, Folarin AA, Roberts A, Bendayan R. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. Artificial Intelligence in Medicine. 2021 Jul 1;117:102083.