Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200197

The Impact of Specialized Corpora for Word Embeddings in Natural Langage Understanding

Antoine NEURAZ[a,b], Bastien RANCE[a], Nicolas GARCELON[a], Leonardo Campillos LLANOS[b], Anita BURGUN[a], Sophie ROSSET[b]
[a] *INSERM, UMR 1138 Team 22, Paris Descartes, Paris, France*
[b] *LIMSI, CNRS, Université Paris Saclay*

Abstract. Recent studies in the biomedical domain suggest that learning statistical word representations (static or contextualized word embeddings) on large corpora of specialized data improve the results on downstream natural language processing (NLP) tasks. In this paper, we explore the impact of the data source of word representations on a natural language understanding task. We compared embeddings learned with Fasttext (static embedding) and ELMo (contextualized embedding) representations, learned either on the general domain (Wikipedia) or on specialized data (electronic health records, EHR). The best results were obtained with ELMo representations learned on EHR data for the two sub-tasks (+7% and +4% of gain in F1-score). Moreover, ELMo representations were trained with only a fraction of the data used for Fasttext.

Keywords. Natural Language processing, Contextual word embeddings, Natural language understanding

1. Introduction

The recent advances in language representation such as static word embeddings [1,2], and contextual word embeddings (e.g. ELMo[3]) led to a significant performances improvement in natural language understanding (NLU). These methods can take advantage of large text corpora to learn a representation of the vocabulary in the vector space according to words semantics and syntactic use[1]. These approaches are unsupervised: they do not require learning datasets with manual annotations. However, the training corpus must be large enough to learn a correct representation. For example, it is customary to use a full dump of Wikipedia® to learn the models[1–3]. In the biomedical domain, the vocabulary is very specific. Moreover, when looking at clinical reports in electronic health records (EHRs), the syntax is also different from articles from sources similar to Wikipedia. A recent study from Wang *et al.*[4] evaluated the performances of embeddings learned on Wikipedia, biomedical publications and clinical notes in English : learning embeddings using clinical notes from EHRs increased the performances. Would this improvement also be true for contextual word embeddings? Nonetheless, it may be challenging to have access to the large number of documents requested (*e.g.* more than 100k patients in the study of Wang *et al.*) or even have access to embedding matrices learned on such corpus, due to privacy issues. It is possible that using more advanced methods to learn the word representations such as

ELMo compared to static embeddings would allow researchers to use embeddings learned on general domain corpora and still obtain good performances.

The aim of this work is to explore the impact of the method and data source of embeddings in the context of a weakly supervised NLU task in French. For this task in a restricted domain (biomedicine), we focused on the two NLU tasks: slot-filling and intent classification.

More specifically, this task aims at providing NLU in a virtual assistant in French for clinicians to explore biological tests results information in the patient's record in natural language (e.g. *Donne moi le dernier résultat de créatinine* 'Give me the last result of creatinin'). The set of queries that a physician may have about characteristics and results of a patient is broad and diverse. Therefore, enabling queries in natural language may help accessing information more efficiently. Given that no public dataset is available for this task in French, we used a weakly supervised approach by generating training data from question templates, terminologies and paraphrases as described in a previous work [5].

We compared the performances of two types of word representations: static word embeddings (Fasttext) and contextualized word embeddings (ELMo). For each method, we also compared two different learning sets in French: #1 Wikipedia or #2 a set of 1M clinical notes from our local EHR.

2. Methods

2.1. Embeddings

We compared two types of embeddings: #1 continuous skip-gram model with sub-word information as implemented in fastText[2], #2 embeddings from language models (ELMo) where the vectors are learned from the internal states of a deep bidirectional language model[3]. As a baseline, we use a continuous skip-gram model learned only on the training set (no external dataset).

We also compared the performances of these methods when learned on either a general domain dataset or a specialized dataset. The general domain dataset (hereafter, Wiki) is made of a dump of the French version of Wikipedia plus the French dataset of CommonCrawl. For this dataset, we used pre-learned models for French downloaded from fasttext website[1] for Fasttext and from github[2] for ELMo.

The specialized dataset (hereafter, EHR) is constituted of a random set of 1M clinical notes from the clinical data warehouse of Necker – Enfants malades hospital, a French AP-HP childrens hospital in Paris[6]. This dataset contains 162M tokens and a vocabulary of size 92k. For Fasttext, we used vectors of 300 dimensions and a window-size of 5. For ELMo, we kept only a subset of the EHR data (24M tokens) due to the high training time of ELMo. To compare embeddings from different sources, we kept the hyper-parameters as described in [3].

[1] https://fasttext.cc

[2] https://github.com/HIT-SCIR/ELMoForManyLangs

2.2. Task

We evaluated the impact of the different embeddings approaches on the task of NLU in a virtual assistant (VA task). Given that no public dataset is available for this task in French, we used the dataset generated from templates, terminologies and paraphrases as described in a previous work [5]. The training dataset contains 16,000 questions, 144k words (mean length of a question is 9), the development set of 4,000 questions for the tuning of the models. For the evaluation, we collected from physicians in our hospital, a set of 178 questions that they would like to ask in such a system. This set of questions has been manually annotated.

The slot filling task (VA-sequence) consists of a sequence labeling task aiming at identifying, in the question, the labtest mention (e.g *créatinine* 'creatinin') and the date- related information (e.g. *22/04/2012*, *depuis 4 semaines* 'for 4 weeks'). In the training set, the number of distinct lab mentions was 336 with a length ranging from 1 to 11 tokens and a median length of 2. In terms of vocabulary, there is only an overlap of 28% between the training set and the test set for labtest mentions. The date labels include exact dates, relative dates and time ranges.[5]

The intent classification task (VA-classification) is divided into 4 sub-tasks corresponding to 4 axes of classification. For each utterance, we assign one label per axis. Two axes concern the results of the lab exams (*i.e.* the type of result (5 categories) and interpretation of the result (5 categories). The two latter concern temporal aspects (*i.e.* the time of result (3 categories) and constraints on time (4 categories)).[5]

2.3. Models

We evaluated 5 different configurations of embeddings: continuous skip-gram model of 300 dimensions (D) learned on the training set; fasttext of 300D learned on Wiki; fasttext of 300D learned on EHR; ELMo of 1024D learned on Wiki; ELMo of 1024D learned on EHR.

For the sequence labeling task (VA-sequence), we used a recurrent neural network (RNN) based on bidirectional long short term memory units (biLSTM)[7]. We used two layers of biLSTM of size 256.

For the classification tasks we used a convolutional neural network (CNN)[8]. The model contains a 1D convolutional layer of 250 units, kernel size of 3, ReLU activation, followed by a max pooling layer and a dense fully connected layer.

All the models were implemented using Keras, with a Tensorflow backend, and the optimizer was Adam. We used dropouts after the embedding layer and before the final dense layer as well as L2-regularization on the convolutional layers to limit overfitting. We used a weighted F1-score (harmonic mean of the precision and the recall) to evaluate the results of the different models. To estimate the variability of the performances, we used 10 repetitions of 5 fold cross-validation on the test set.

3. Results

All the results are presented in Table 1. For the sequence labeling task (VA-sequence), the best results are obtained with ELMo learned on EHR with a F1-score of 0.76 (95%CI [0.74-77]) compared to Fasttext on EHR (0.67, 95%CI [0.61-0.73]). Interestingly, we

show similar results between ELMo on Wiki (0.69, 95%CI [0.67-0.70]) and Fasttext on EHR (0.67, 95%CI [0.61-0.73]). To note, the distance between the training set and the test set, estimated using the perplexity of the n-gram model, is 194.

Table 1. Results for the NLU task for the virtual assistant: sequence labeling (Bi-LSTM) and intent classification (CNN)

Method	**Sequence labeling** F1-score [95%CI]	**Intent classification** Mean F1-score [95%CI]
Baseline (only training set)	0.62 [0.61-0.64]	0.62 [0.58-0.65]
Fasttext on Wiki	0.69 [0.67-0.70]	0.52 [0.49-0.55]
Fasttext on EHR	0.67 [0.61-0.73]	0.66 [0.63-0.69]
ELMo on Wiki	0.69 [0.67-0.70]	0.49 [0.46-0.52]
ELMo on EHR	**0.76 [0.74-0.77]**	**0.70 [0.67-0.73]**

For the intent classification task (VA-classification), again, the best results come from the ELMo model learned on EHR (F1-score 0.70, 95%CI [0.67-0.73]) compared to the second-best Fasttext on EHR (F1-score 0.66, 95%CI [0.63-0.69]). In this task, the models learned on Wiki performed poorly: Fasttext on Wiki with a F1-score of 0.52 (95%CI [0.49-0.55]) and ELMo on Wiki 0.49 (95%CI [0.46-0.52]) with a baseline at 0.62 (95%CI [0.58-0.65]).

4. Discussion

The use of pre-learned embeddings may improve substancially the results on certain downstream tasks. As shown by Wang *et al.* [4] on various tasks in English, we show increased performances on two of three tasks in French when using pre-learned embeddings on a large corpus of clinical notes. The higher improvement was obtained on the VA-sequence task suggesting that tasks requiring extended lexicons may benefit the most of the pre-embeddings. Indeed, the identification of labtest mentions highly depends on the vobabulary available during training and we show that 72% of the mentions in the test set are absent from the training set. In this scenario, the representations learned on a large specialized dataset helped the model to obtain a better generalization when dealing with previously unseen mentions.

The test set shows 29% of out of vocabulary words overall. It may also partly explain the improvements observed on the classification task also (VA-classification). Interestingly, in the VA-classification task, both embeddings (Fasttext and ELMo) learned on Wiki did worsen the results compared to the baseline. This noise might come from the lack of specialized vocabulary in the Wiki corpus and a different usage of specialized words that may vary depending on the biomedical context and that are not taken into account in the general domain.

Sheikhshab *et al.*[9] evaluated ELMo embeddings on named entity recognition tasks in the biomedical domain. They also showed improved results when using an in-domain corpus to train ELMo (*i.e.* articles from Pubmed). It would be interesting to evaluate ELMo embeddings learned on a corpus of biomedical article in French but collecting such a corpus may be challenging.

This study has some limits. First, the embeddings learned on EHR came from a single hospital. This may cause some bias in the results given that the semantic particularities of the different practices may have an impact on the suitability of the embeddings for the different tasks. One possible solution to tackle this issue would be to

learn embeddings from multiple sites. However, due to privacy issues, it is not possible to transfer massive amount of hospital data. Two other approaches might be used to workaround this issue and should be explored for the biomedical domain: #1 the use of federated learning to train the models without moving the data[10]; #2 learning one embedding per site and aligning the different embeddings using unsupervised techniques. Finally, it is not obvious to determine the size of the training corpus for an embedding. It is a common practice to use all the articles from Wikipedia for example. But do we need that amount of data in a specialized domain? Comparing the impact of the size of the embedding corpus on downstream task will be of great interest.

5. Conclusions

Depending on the task, embeddings learned on large corpora can have a significant impact on NLP tasks in the biomedical domain in French. Moreover, learning these embeddings on clinical notes will increase the performances compared to general domain. As it may not be feasible to access a large corpus of clinical notes, it is still profitable to use advanced methods such as ELMo learned on general domain and obtain reasonable results. When the task does not rely on a large specialized vocabulary, the impact of external embeddings might be reduced.

References

[1] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2013: pp. 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf (accessed March 31, 2017).

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Enriching word vectors with subword information, *ArXiv Preprint ArXiv:1607.04606*. (2016). https://arxiv.org/abs/1607.04606 (accessed February 14, 2017).

[3] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, Deep Contextualized Word Representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018: pp. 2227–2237. doi:10.18653/v1/N18-1202.

[4] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu, A comparison of word embeddings for the biomedical natural language processing, *Journal of Biomedical Informatics*. **87** (2018) 12–20. doi:10.1016/j.jbi.2018.09.008.

[5] A. Neuraz, A. Burgun, L.C. Llanos, and S. Rosset, Natural language understanding for task oriented dialog in the biomedical domain in a low ressources context, *Machine Learning for Health (ML4H) Workshop at NeurIPS 2018*. (2018).

[6] N. Garcelon, A. Neuraz, R. Salomon, H. Faour, V. Benoit, A. Delapalme, A. Munnich, A. Burgun, and B. Rance, A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse, *Journal of Biomedical Informatics*. **80** (2018) 52–63. doi:10.1016/j.jbi.2018.02.019.

[7] A. Graves, and J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Networks*. **18** (2005) 602–610. doi:10.1016/j.neunet.2005.06.042.

[8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE*. **86** (1998) 2278–2324. doi:10.1109/5.726791.

[9] G. Sheikhshabbafghi, I. Birol, and A. Sarkar, In-domain Context-aware Token Embeddings Improve Biomedical Named Entity Recognition, (n.d.) 5.

[10] D. Zhang, X. Chen, D. Wang, and J. Shi, A Survey on Collaborative Deep Learning and Privacy-Preserving, in: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), 2018: pp. 652–658. doi:10.1109/DSC.2018.00104.