MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190533

Do You Need Embeddings Trained on a Massive Specialized Corpus for Your Clinical Natural Language Processing Task?

Antoine Neuraz[a,b,c], Vincent Looten[b,d], Bastien Rance[b,d], Nicolas Daniel[d], Nicolas Garcelon[a,e], Leonardo Campillos Llanos[c], Anita Burgun[a,b,d], Sophie Rosset[c]

[a] *Institut National de la Santé et de la Recherche Médicale (INSERM), Centre de Recherche des Cordeliers, UMR 1138 Equipe 22, Paris Descartes, Sorbonne Paris Cité University, Paris, France*
[b] *Department of Medical Informatics, Necker-Enfants Malades Hospital, Assistance Publique des Hôpitaux de Paris (AP-HP)*
[c] *LIMSI, CNRS, Université Paris Saclay*
[d] *Hôpital Européen Georges Pompidou, AP-HP, Université Paris Descartes, Sorbonne Paris Cité, Paris, France*
[e] *Institut Imagine, Paris Descartes Université Paris Descartes-Sorbonne Paris Cité, Paris, France*

Abstract

We explore the impact of data source on word representations for different NLP tasks in the clinical domain in French (natural language understanding and text classification). We compared word embeddings (Fasttext) and language models (ELMo), 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 one of the two tasks(+7% and +8% of gain in F1-score).

Keywords:
Natural language processing, electronic health records.

Introduction

State of the art methods in natural language processing (NLP) usually include learning a statistical representation of the vocabulary such as word embedding (e.g., word2vec, Glove, Fasttext [1]) or language models (e.g. ELMo [2]). One of the advantages of these approaches is that they are unsupervised: they do not require the learning datasets to be manually annotated with a ground truth. However, in the biomedical domain, the vocabulary is very specific. Would the representations learned from general domain corpus be usable for NLP tasks in the biomedical domain? Wang *et al.* [3] showed that learning embeddings using clinical notes from EHRs increased the performances. 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.

This work explores the impact of the method and data source of embeddings on different tasks. We compared the performances of two types of word representations: Fasttext and ELMo. We also compared two different learning sets in French: Wikipedia or a set of 1M clinical notes from our local EHR. We evaluated the performances of these approaches on two different NLP tasks in French: natural language understanding and text classification.

Methods

Embeddings

We compared two types of embeddings: 1) continuous skip-gram model with sub-word information (*i.e.,* each word is represented as a bag of character n-grams), as implemented in fastText[1] and 2) embeddings from language models (ELMo) where the vectors are learned from the internal states of a deep bidirectional language model as described in [2]. 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 (French Wikipedia, Wiki) or a specialized dataset (1m clinical notes from a French AP-HP childrens hospital in Paris, EHR).

Tasks

Natural language understanding in a virtual assistant (VA task) This task aims to provide natural language understanding (NLU) in a virtual assistant 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', Comment a évolué l'hémoglobine depuis 2 ans 'How has the hemoglobin evolved for the last 2 years'). 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 access information more efficiently.

Given that no public dataset is available for this task in French, we generated a training set of 16,000 questions from templates, terminologies and paraphrases as described in a previous work [4]. This dataset contains 144k words (mean length of a question is 9). We also generated a 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.

For this study, we focused on the slot filling task (VA-sequence) consisting of a sequence labeling task aiming at

identifying whithin the question the lab test 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.

Text classification of absconding behaviour in nurses' notes (ABS-task) Absconding, where a patient leaves the **hospital without informing staff, is a rare but serious adverse event that can affect the safety of both the patient and people outside the hospital. Identifying hospital encounters with absconding patients can be challenging in most hospital information systems, given that those events are not systematically collected in a structured format.**

We collected nurses' notes from the clinical data warehouse of the European Hospital Georges Pompidou, an AP-HP hospital located in Paris. We collected a set of 2,605 notes that were manually classified by two experts by whether it mentions an actual absconding or not.

Models

As we aim to evaluate the impact of embeddings methods and data sources, we fixed the model used for the tasks and only let vary the embeddings.

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

Sequence Labeling task (VA-sequence) For the sequence labeling task (VA-sequence), we used a recurrent neural network (RNN) based on long short term memory units (LSTM). The architecture is based on a 2 layers bidirectional LSTM of 256 units.

CNN for Text Classification (ABS-task) For the classification tasks we used a convolutional neural network (CNN). After the embedding layer, the model contains a 1 dimensional convolutional layer of 250 units, kernel size of 3, ReLU activation, followed by a max pooling layer and a dense fully connected layer with softmax or sigmoid activation depending on the number of classes to predict.

We used a weighted F1-score to evaluate the results of the different models with 5 fold cross-validation.

Results

Sequence labeling task (VA-sequence) All the results for this task are presented in Table 1. 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]).

Text classification On this task, all the models give similar results with a low variation around the baseline (F1-score of 0.89 (95%CI [0.88-0.90])) as shown in Table 2.

Table 1 – Results for the NLU task for the virtual assistant: sequence labeling (Bi-LSTM)

Method	F1-score [95%CI]
Baseline (only training set)	0.62 [0.61-0.64]
Fasttext on Wiki	0.69 [0.67-0.70]
Fasttext on EHR	0.67 [0.61-0.73]
ELMo on Wiki	0.69 [0.67-0.70]
ELMo on EHR	**0.76 [0.74-0.77]**

Table 2 – Results for the classification of absconding in nurses' notes (CNN)

Method	F1-score [95%CI]
Baseline (only training set)	0.88 [0.87-0.89]
Fasttext on Wiki	0.85 [0.84-0.86]
Fasttext on EHR	0.87 [0.86-0.88]
ELMo on Wiki	**0.89 [0.88-0.90]**
ELMo on EHR	0.84 [0.83-0.85]

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] 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).

[2] 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.

[3] 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.

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

Address for correspondence
Antoine Neuraz, antoine.neuraz@aphp.fr