BioReddit: Word Embeddings for User-Generated Biomedical NLP

Marco Basaldella and **Nigel Collier**
Department of Theoretical and Applied Linguistics
University of Cambridge
Cambridge, UK
{mb2313, nhc30}@domain

Abstract

Word embeddings, in their different shapes and evolutions, have changed the natural language processing research landscape in the last years. The biomedical text processing field is no stranger to this revolution; however, researchers in the field largely trained their embeddings on scientific documents, even when working on user-generated data. In this paper we show how training embeddings from a corpus collected from user-generated text from medical forums heavily influences the performance on downstream tasks, outperforming embeddings trained both on general purpose data or on scientific papers when applied to user-generated content.

1 Introduction

In the Natural Language Processing community, user-generated content, i.e. data from social media, user forums, review websites, and so on, has been the subject of many studies in the past years; the same holds for the biomedical domain, where there has been a great effort on the applications of NLP techniques for biomedical scientific publications, patient records, and so on. However, the intersection of the two fields is still in its infancy, even when dealing with relatively basic NLP tasks. For instance, in the field of user-generated biomedical natural language processing (hence UG-BioNLP), to the best of our knowledge there are no publicly available corpora for Named Entity Recognition (NER) akin in size and purpose e.g. to the CoNLL 2003 dataset. (Tjong Kim Sang and De Meulder, 2003), making it hard to compare systems effectively. Moreover, while there have been experiments on training word embeddings with biomedical data, we are not aware of any publicly available word embeddings trained on UG-BioNLP data.

For this reason, we decided to investigate the impact of using purpose-trained word embeddings in the Bio-UG field. In order to train such embeddings, we collected a dataset from Reddit, scraping posts from medical-themed subreddits, both on general health topics such as '`r/AskDocs`', or on disease-specific subreddits, such as '`r/cancer`', '`r/asthma`', and so on. We then trained word embeddings on this corpus using different off-the-shelf techniques. Then, to evaluate the embeddings, we collected a second dataset of 4800 threads from the health forum *HealthUnlocked*, which was annotated for the NER task. Then, we analyzed the performance of the embeddings on the tasks of NER and of adverse effect mention detection. For NER, we used Conditional Random Fields as a baseline. We compared them against Bidirectional LSTM-CRFs (Lample et al., 2016), on which we analyzed the impact of using our custom-trained word embeddings against embeddings trained on general purpose data and scientific biomedical publications when evaluating on our purpose-built HealthUnlocked dataset and on the PsyTar and CADEC corpora. Finally, we evaluated the performance of a simple architecture for adverse reaction mention detection on the PsyTAR corpus. We conclude the paper explaining our intentions for future research, in other to obtain other results that confirm the preliminary findings we present in this work.

2 Related Work

The benefit of using in-domain embeddings for the biomedical domain has already been proven effective. For example, (Pakhomov et al., 2016) and (Wang et al., 2018) found that using clinical notes or biomedical articles for training word embeddings has generally a positive impact on down-

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 34–38
Hong Kong, November 3, 2019.
https://doi.org/10.18653/v1/D19-62

stream NLP tasks. (Nikfarjam et al., 2015) trained embeddings on user-generated medical content and used them successfully on the pharmacovigilance task; however, they trained the embeddings an adverse reaction mining corpus, hence making them too task-specific to be considered useful on generic UG-BioNLP tasks.

3 Datasets

3.1 BioReddit

To train our embeddings on user-generated biomedical text, we choose to scrape data from the discussion website Reddit. The website is organized by forums, called *subreddits*, where the discussion is restricted to a topic, e.g. general news, computer science, and so on. There is a great number of health-themed subreddits, where users from all around the world discuss their health problems or ask for medical advice, which is ideal for training our embeddings.

We also evaluated the micro-blogging platform Twitter as a possible source for the embeddings, but we quickly discarded it due to its unstructured nature. On Twitter, in fact, information is not pre-aggregated by subject, and one has to search for the required posts by searching for keyword or *hashtag*. This, along with the restrictive limits imposed by Twitter APIs, makes it hard to find relevant content, so we decided to continue with Reddit instead.

We designed a scraping script that downloaded discussions from 68 health themed subreddits. We selected subreddits where users

- could ask for advice, e.g. `/r/AskDocs`, `/r/DiagnoseMe`, `r/AskaPharmacist`,
- discuss a specific illness, e.g. `r/cancer`, `r/migraine`, `r/insomnia`,
- can discuss on any health-related topic, e.g. `r/health`, `r/HealthIT`, `r/HealthInsurance`.

We collected all the posts from these subreddits from the beginning of 2015 to the end of 2018. After that, we cleaned the corpus for bot-generated content, e.g. bots automatically suggesting to seek professional medical advice. We obtained a corpus with 300 million tokens and a vocabulary size of 780,000 words. While the number of tokens is considerably lower than the size of other word embedding training datasets, which could be two orders of magnitude bigger, the vocabulary is quite big; for example, GloVE (Pennington et al., 2014) was trained with a 1.2 million big vocabulary and 27 billion tokens when using Twitter, and on a 600,000 word vocabulary and 6 billion tokens when using Wikipedia.

3.2 HealthUnlocked

In order to evaluate our embeddings, as a first step, we decided to focus on the Named Entity Recognition task. We obtained 4800 forum threads from HealthUnlocked[1], a British social network for health where users can discuss their health with people with similar conditions and obtain advice from professionals.

We annotated the dataset by marking the entities belonging to seven categories, namely: *Phenotype*, *Disease*, *Anatomy*, *Molecule*, *Gene*, *Device*, and *Procedure*. We describe in detail the categories in Table 1.

Since the dataset is collected from patients' discussions, the language used is far from technical. For example[2],

- an user describes paresthesia of arm as *"a tickling sensation in my arms"*;
- another patient, to describe her swollen abdomen, writes that she *"looked six months pregnant"*,
- another user writes that *"her mood is low"*, to explain her depression.

All these phrases, while expressed in layman's language, describe very specific symptoms. For this reason, we developed a set of annotation guidelines where the annotators were asked to mark *any possible mention* of an entity belonging to the seven categories above, even if not expressed with technical language. After running a pilot annotation task on a small set of discussions, we fine tuned the annotation guidelines, and we asked PhD-qualified biomedical experts to annotate 4800 threads from the forums. After the annotation, the files were shuffled and split in train, test, and development set, obtaining 8750, 2526, and 1250 sentences respectively. The number of annotations per category and per set is described in Table 1.

3.3 PsyTAR

The PsyTAR dataset *"contains patients expression of effectiveness and adverse drug events as-*

[1] https://healthunlocked.com/

[2] Please note that we use feminine pronouns to preserve the privacy of the patients.

Category	Description	Train	Dev	Test
Anatomy	Any anatomical structure, organ, bodily fluids, tissues, etc.	1060	146	308
Device	Any medical device used in diagnosis, therapy or prevention.	276	26	82
Disease	Any disorder or abnormal condition.	1234	203	363
Gene	Any molecule carrying genetic information.	342	47	87
Molecule	Any chemical substance.	1791	240	544
Phenotype	Any abnormal morphology, physiology or behaviour.	2963	421	872
Procedure	Any medical procedure used in diagnosis, therapy or prevention.	1158	163	294

Table 1: Description and statistics of the HealthUnlocked dataset used for the experiments.

sociated with psychiatric medications." (Zolnoori et al., 2019). The dataset contains 6000 sentences annotated for mentions and spans (i.e. NER) of Adverse Drug Reactions, Withdrawal Symptoms, Drug Effectiveness, Drug Ineffectiveness, Sign/Symptoms/Illness, and Drug Indications. Each entity is grounded against the Unified Medical Language System (UMLS) and SNOMED Clinical Terms. The source of the corpus is the drug review website Ask a Patient[3]. The language used is very simple, without the use of specialist terms, and with no guarantee of grammatical/spelling correctness.

3.4 CADEC

The CADEC corpus (Karimi et al., 2015) is a corpus of consumer reviews for pharmacovigilance. It is sourced from Ask a Patient too and it is annotated for mentions of concepts such as drugs, adverse reactions, symptoms and diseases, which are linked against SNOMED and MedDRA.

4 Experiments

4.1 Embeddings

Using the dataset described in Section 3.1, we trained three word embedding models, namely GloVe (Pennington et al., 2014), ELMo (Peters et al., 2018), and Flair (Akbik et al., 2018). We choose these models due to their popularity, performance, and relative low resource requirements. In particular, GloVe requires just hours to be trained on a CPU, while ELMo and Flair obtained state-of-the-art results in the NER task at the time of their publication, and both models can be trained in relatively short time ($\sim$1 week) using 1 or 2 GPUs. As general purpose and PubMed embeddings, we use the ones provided or recommended by the respective architecture authors; unfortunately, we are not aware of any GloVe PubMed pre-trainer embeddings available in the public domain. Using our BioReddit dataset, we trained all the embeddings with their default parameters, as described in their respective papers.

[3] https://www.askapatient.com/

Algorithm	P	R	F
CRF	69.7	60.1	64.5
GloVe-Default	69.6	68.3	68.9
GloVe-BioReddit-50	68.7	65.7	67.2
GloVe-BioReddit-100	70.2	71.7	70.9
GloVe-BioReddit-200	72.1	70.3	71.2
ELMo-Default	72.3	72.8	72.5
ELMo-PubMed	73.7	73.7	73.7
ELMo-BioReddit	73.9	76.7	75.3
Flair-Default	75.0	75.8	75.4
Flair-PubMed	75.8	75.1	75.4
Flair-BioReddit	76.5	76.2	76.4

Table 2: Performance of different embeddings technique on NER, when trained and evaluated on the dataset described in Section 3.2.

4.2 Named Entity Recognition

In order to evaluate our embeddings we use Conditional Random Fields and as a baseline, and then we evaluate our embeddings using a Bidirectional LSTM-CRF sequence tagging neural network (Lample et al., 2016). We refer the reader to the original paper for an explanation on how this architecture works, as the details are outside to the scope of the present paper.

We present our results in Table 2. As expected, all the neural architectures largely improve the results obtained by the CRF and, in line with the literature, Flair performs slightly better than ELMo, which in turn performs better than GloVe. Using our purpose-built embeddings, called *BioReddit* in the Table, we always obtain an improvement with respect to using embeddings trained on general-purpose data (*Default* in Table) or on PubMed, barring the smallest GloVe vectors.

Category	P	R	F
Anatomy	72.2	76.6	74.3
Device	67.2	50.0	57.3
Disease	76.8	80.2	78.4
Gene	80.4	85.0	82.7
Molecule	88.4	88.6	88.5
Phenotype	70.5	66.9	68.6
Procedure	76.6	80.2	78.4

Table 3: Performance on the NER task of the Flair-BioReddit on the HealthUnlocked dataset on the seven categories defined in Section 3.2.

Corpus	Task	Embedding	P	R	F
		Default	65.3	59.7	62.4
PsyTAR	NER	PubMed	65.0	55.3	59.8
		BioReddit	63.7	63.8	63.7
		Default	81.3	69.2	74.8
PsyTAR	ADR	PubMed	77.5	72.6	75.0
		BioReddit	79.5	73.7	76.5
		Default	77.1	76.0	76.5
CADEC	NER	PubMed	77.2	76.1	76.7
		Bioreddit	78.6	77.4	78.0

Table 4: Performance of the Flair embeddings on the NER and Adverse Reaction Mention Detection on the PsyTAR and CADEC corpora.

In Table 3 we provide a per-category breakdown of the best performing embeddings, i.e. Flair embeddings trained on our BioReddit corpus. It's interesting to note how the most difficult categories are *Device* and *Phenotype*. We explain this results by noting that the former is the least represented category in the corpus, while the latter was actually expected to be the hardest category. In fact, looking into the corpus, we found that users are relatively precise when talking about disease names, genes, molecules, and so on, while they don't necessarily describe their symptoms using "proper" medical language.

In Table 4 we see the results we obtain on the NER task on the PsyTAR and CADEC corpora while using Flair embeddings, where BioReddit embeddings always outperform general-purpose and PubMed trained ones. Interestingly, PubMed embeddings behave considerably worse than the others on the PsyTAR corpus, which seems to support the intuition that using a specialized scientific corpus is not always the guarantee of better performance.

4.3 Adverse Reaction Mention Detection

The task of Adverse Reaction Mention Detection (hence ADR) consists in detecting whether in a sentence a user mentions that he is experiencing/experienced an adverse reaction to a drug. For this task, we designed a simple neural architecture, where a bidirectional GRU (Cho et al., 2014) reads a sentence, and a softmax layer on its top performs the binary classification task of detecting wether the input sentence contains an ADR or not. When evaluating on the PsyTAR corpus we again obtain the best performance when using our BioReddit embeddings, followed by the PubMed trained ones and the default ones.

5 Conclusions

In this paper we showed how training ad-hoc embeddings for the task of user-generated biomedical text processing improves the results in the tasks of named entity recognition and adverse reaction mention detection. While preliminary, our results show a strong indication that embeddings trained on biomedical scientific literature only are not guaranteed to be effective when used on user-generated data, since people use "layman terms" which are seldom, if ever, used in scientific literature. As future work, we acknowledge the need to better investigate the results we present here. A good starting point would be to analyze other embedding techniques, in order to investigate if the performance improvement is due to embedding techniques themselves or to the datasets used. Moreover, we need to analyze the performance of our BioReddit embeddings on non-user generated content, as e.g. scientific abstracts, in order to investigate whether they are able to perform effectively on this domain too. Finally, we think that a manual investigation of the results of the downstream tasks is important, to investigate e.g. if the improvement in the ADR task is due to the embeddings helping to classify sentences with more colloquial language. Unfortunately, due to licensing and privacy issues, we are not allowed to release the HealthUnlocked corpus. However, we make available our BioReddit embeddings trained on GloVe, ELMo and Flair at https://github.com/basaldella/bioreddit.
For the sake of reproducibility, we also we make available our PsyTAR preprocessed splits online at https://github.com/basaldella/psytarpreprocessor.

Acknowledgements

The authors would like to thank HealthUnlocked for providing the dataset used in this paper, Taher Pilhevar for his useful suggestions, and NVIDIA Corporation for the donation of the GPU cards used to train the models presented in this paper. The authors acknowledge support by the EPSRC grant EP/M005089/1.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. CADEC: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73 – 81.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc*, 22(3):671–681.

Serguei VS Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B Melton. 2016. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, 32(23):3635–3644.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12 – 20.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Nilay D. Shah, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu, Soo Kyung Park, Kelly Xu, and Hamideh Moayyed. 2019. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in Brief*, 24:103838.