Creating Domain Specific Embeddings toWork with Imbalanced Datasets in Automatic Extraction of Medication Names in Tweets

Renzo Rivera Zavala[a], Paloma Martinez[a] and Jose Luis Martinez[b]
[a] Carlos III University of Madrid, Avda. Universidad, 30, 28911 Leganés, Madrid
[b] Konplik Health, http://www.konplik.health

***Abstract*—In this work, we introduce two Deep Learning architectures for the identification of drug mentions in tweets. We propose a deep neural approach based on two models: two Bidirectional Long Short-Term Memory (Bi-LSTM) networks and a Conditional Random Field (CRF) network using character and contextualized-word embeddings to deal with the extraction of semantic, syntactic, and morphological features; and a Bidirectional Encoder Representations from Transformers (BERT) using pre-trained contextualized word embeddings created from scratch with FastText with a collection of tweets. Both models have been evaluated on the BioCreative VII Track 3 dataset obtaining an F-measure of 64.2% and 67.7%, respectively.**

Keywords—Natural Language Processing, Deep Learning, Contextual Information

I. Introduction

Currently, social networks are a common way to make questions, opinions, or answers about different biomedical aspects. Specifically, medication mentions in tweets are an important source for pharmacoepidemiological research. Therefore, the efficient access to information on medical data described in tweet posts is of growing interest in the biomedical industry, research, and so forth. In this context, improved access to medical concepts mentioned in tweet texts is a crucial step prior to downstream tasks such as drug and protein interactions, chemical compounds, adverse drug reactions, among others.

Named Entity Recognition (NER) is an essential task in biomedical Information Extraction (IE), intending to automatically extract and identify mentions of concepts of interest in running text, typically through their mention offsets or by classifying individual tokens whether they belong to entity mentions or not. The NER task has been addressed using Dictionary-based methods, which are limited by the size of the dictionary, spelling errors, the use of synonyms, and the constant growth of vocabulary. Rule-based methods and Machine Learning methods usually require both syntactic and semantic features as well as specific language and domain features. More recently, state-of-the-art deep learning methods results in NER are based on pre-trained models (word embeddings) obtained from a considerable volume of unlabelled texts (scientific literature, social media texts, Wikipedia, among others). However, public available biomedical pre-trained models in social media texts are limited. To the best of our knowledge, only one public available work addresses the generation of biomedical word embeddings on tweets text (1).

In this paper, we propose two deep neural models: Bi-LSTM+CRF and BERT. To do this, we adapt the NeuroNER model proposed in (2) for NER offset and entity classification of the BioCreative VII Track 3 task [a] . Specifically, we have extended NeuroNER by adding contextualized-word information and information about overlapping or nested entities. Moreover, in this work, we use an existing pre-trained noncontextualized-word model as well as our trained from scratch contextualized-word model: i) a Glove 6B Embedding model (3), trained on Wikipedia and GoogleNews; ii) word2vec PubMed-and-PMC-w2v (4) trained on PubMed and PMC articles; iii) the FastText English Twitter 100d (5) trained on general tweets post; iv) our English medical word embeddings trained using the FastText model, and v) a sense-disambiguation embedding model (sense2vec) (6). Finally, we fine-tune BERT existing pre-trained contextualized-word model as well as our trained from scratch contextualized-word model: i) BioBERT-Large v1.1 (7) trained on PubMed and PMC articles and ii) BERTweet (8) trained on General and COVID tweets.

Experiment results on BioCreative VII track 3 showed that our features representation improved each separate representation, implying that LSTM-based compositions play different roles in capturing token-level features for NER tasks, thus improving their combination. Moreover, the use of specific domain contextualized word vector representations outperforms general domain word vector representations.

II. Materials and Methods

In this section, we first describe the corpora used to generate our train from the scratch word representation, the training procedure, and the pre-trained non-contextualized and contextualized word models used in our study. Then, we describe our system architecture for offset and entity

[a] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-3/

classification. Finally, the datasets used for training, validating, and evaluating our deep learning model performance are explained.

A. Corpora

The BioCreative VII track 3 dataset available in this task to train learning models is highly unbalanced, as it happens in the real world. From more than 89,000 tweets available in the training dataset, only 212 of them contain a valid drug mention. One of the approaches we have followed in our research has been to increase the number of tweets containing drug mentions. In order to do this, we have grown the training dataset with other available collections, as the SMM4H'18 shared task tweets, a balanced dataset containing 10,000 tweets. Obviously, this is not enough to obtain a useful training dataset so we have also gathered additional data from Twitter. A crawling component to gather Twitter posts using the public Twitter API has been developed. Only tweets containing drug mentions are interesting, so we have used Konplik's text analytics technology (supported by MeaningCloud [b]) to filter out irrelevant posts. For this purpose, the Topics Extraction API has been customized with dictionaries containing drug names. These dictionaries have been built integrating different taxonomies, such as UMLS (9). Besides, drug mentions identified in the CADEC dataset have also been included. CADEC (CSIRO Adverse Drug Event Corpus) (10) is a corpus containing Adverse Drug Effects reported by patients in medical forums provided by AskAPatient [c]. The TwiMed collection (11), containing around 1,500 drug mentions found in Twitter messages and PubMed sentences extracted from abstracts, has also been integrated. The post's contents can be assimilated into Twitter messages. All the datasets have been converted to BRAT notation so they can be used for training.

Combining these collections has helped us build a dataset containing 160,000 tweets containing a mention of a drug. This collection has been used to train the Bi-LSTM + CRF and BERT-based models used in this research work. Moreover, we used tweet texts to build our own corpora and our train from the scratch word embedding model.

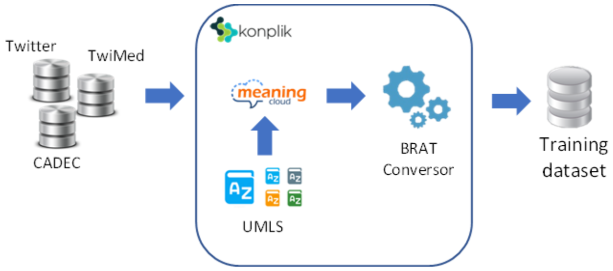

Fig. 1. Konplik Twitter crawler.

All the corpora are in TXT format files. TXT files were not processed. Raw texts from all files were compiled in a single TXT file. Texts were processed, setting all to lower, removing punctuation marks, trailing spaces and stop words and used as input to generate our word embeddings. Sentences preprocessing (split and tokenization) were made using Spacy [d], an open-source python library for advanced multi-language natural language processing.

B. Word embedding models

The use of word representations from pre-trained unsupervised methods is a common practice and a crucial step in NER pipelines. Previous word embedding models such as Word2Vec (12), Glove (3), and FastText (13) focused on non-contextualized word representations. However, in the last few years models are focused on learning contextualized word representations, such as ELM (14), CoVe (15), and the state-of-the-art BERT model (16).

In this work, we used various English pre-trained embedding models. The Glove 6B (G6B) is a pre-trained word embeddings model trained on different general domain text corpora written in Spanish (Wikipedia 2014 + Gigaword) using the Glove implementation. The PubMed and PMC (PubW2V) model use texts and their combination with the word2vec implementation. The FastText English Twitter (FastTwitter) pre-trained word embeddings model was trained on the general domain tweets posts using the FastText implementation. We also integrate the sense2vec model, which provides multiple dense vector representations for each word based on the sense of the word. This model is able to analyze the context of a word based on the lexical and grammatical properties of words and then assigns its more adequate vector. Each word in this model is paired with its corresponding Part-of-Speech (PoS) tag. Sense2vec uses the Polyglot Part-of-Speech tagger from AlRfou (more details in (6)). We used the Reddit Vector, a pre-trained model of sense-disambiguation representation vectors presented by (6). This model was trained on a collection of general domain comments published on Reddit (corresponding to the year 2015) written in Spanish and English. Table 1 shows word embeddings details.

Furthermore, we used the FastText (13) implementation to train our own word embeddings using the Twitter corpora described in section Corpora (MedTwitter).

TABLE 1 NON-CONTEXTUALIZED-WORD EMBEDDING MODELS DETAILS.

Detail	**G6B**	**PubW2V**	**FastTwitter**	**MedTwitter**
Domain	General	Biomedical	General	General
Corpus size	6 billion	5.7 billion	2.5 million	2 million
Vocab size	200k	120k	31k	45k
Algorithm	Glove	Word2vec	FastText	FastText

BERT is a context-dependent word representation model based on a masked language model and trained using the transformer architecture (16). Even though BERT learns a lot about language through pre-training, it is possible to adapt the

[b] https://www.meaningcloud.com
[c] https://www.askapatient.com
[d] https://spacy.io/

model by adding a customized layer on top of BERT outputs, and then new training is done with specific data (this phase is called fine-tuning). We refer readers to (16) for a more detailed description of BERT.

Due to the benefits of the BERT model, we adopted the English Biomedical BioBERT and the English Twitter BERTweet (8) pre-trained BERT models. BERT pre-trained models are shown in Table 2

TABLE 2 CONTEXTUALIZED-WORD EMBEDDING MODELS DETAILS.

Detail	**BioBERT-Large v1.1**	**BERTweet**
Domain	Biomedical	General
Corpus size	21 billion	16.5 billion
Vocab size	64k	64k
Algorithm	BERT	BERT

C. System Description

Our approach involves the adaption of a state-of-art NER model named NeuroNER as proposed in (2), based on a deep learning network with a preprocess step, learning transfer from pre-trained models, two recurrent neural network layers, and the last layer for CRF (see Figure 3). The input for the first Bi-LSTM layer is character embeddings. In the second layer, we concatenate character embeddings from the first layer with contextualized word representations for the second Bi-LSTM layer. Finally, the last CRF layer obtains the most suitable labels for each token using a tag encoding format. For more details about NeuroNER, please refer to (2).

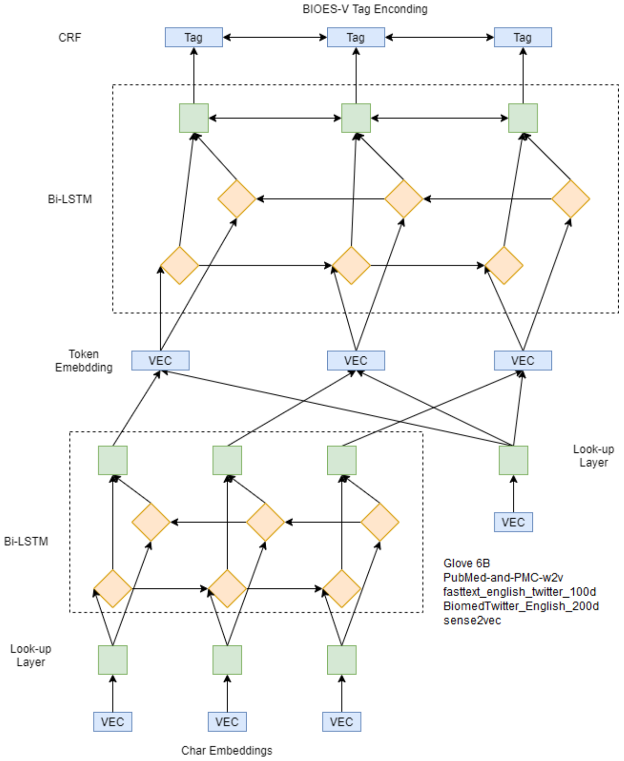

Fig. 2. The architecture of the Bi-LSTM CRF model medication names identification.

Our contribution consists of extending the NeuroNER system with additional features. In particular, adding contextualized-word representations and the extended BMEWO-V encoding format has been added to the network.

Finally, we fine-tune BERT with pre-trained models described in Table 2 using the BMEWO-V encoding format. The BMEWO-V encoding format distinguishes the B tag for entity start, the M tag for entity continuity, the E tag for entity end, the W tag for a single entity, and the O tag for other tokens that do not belong to any entity. The V tag allows us to represent nested entities. BMEWO-V is similar to other previous encoding formats (17); however, it allows the representation of nested and discontinuous entities. Only the labels found in the annotations are used for training, validation, and evaluation steps. As a result, we obtain our sentences annotated in the CoNLL-2003 format (18).

III. EVALUATION

As it was described above, our system is based on two deep learning models, two Bi-LSTM layers, and the last layer for CRF and the BERT model. We evaluate our NER systems using the train, validation, and test subsets from the BioCreative VII track 3 dataset provided by the BioCreative task organizers. The training subset is composed of 89,000 tweets with 218 entities mentions, the valid subset is composed of 39,000 tweets with 93 entities mentions, and the test subset is composed of 54,000 tweets with no annotations. The BioCreative dataset is a manually annotated corpus of tweets posted by 212 Twitter users during their pregnancy written in English and annotated drug mentions on Twitter.

The F-measure is used as the main metric where true positives are entities that match with the gold standard entity boundaries and type. A detailed description of the evaluation can be found on the BioCreative VII Track 3 web [e].

The NER task is addressed as a sequence labeling task. For the NER track, we tested different configurations with various pre-trained non-contextualized and contextualized-word models. The pre-trained models and their parameters are summarized in Table 1 and Table 2.

In Table 3, we compare the different non-contextualized pre-trained models on the validation subset. As shown in Table 3, specific domain non-contextual-word models outperform general domain models by almost 3 points on strict evaluation.

TABLE 3 NON-CONTEXTUALIZED-WORD MODELS RESULTS FOR ENTITY CLASSIFICATION ON BIOCREATIVE VII TRACK 3 VALID SUBSET.

Dataset	**Pre-trained Model**	**Strict F**
Biocreative	G6B	55.99
Biocreative	PubW2V	54.17
Biocreative	FastTwitter	56.84
Biocreative	MedTwitter	58.14
Biocreative + SMMH4	MedTwitter	60.23
Biocreative + SMMH4 + Konplik	MedTwitter	64.36

[e] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-3/

In Table 4, we compare the different contextualized pre-trained models on the validation subset. As shown in Table 4, specific domain non-contextualized-word models outperform general domain models by almost 2 points on strict evaluation.

TABLE 4 NON-CONTEXTUALIZED-WORD MODELS RESULTS FOR ENTITY CLASSIFICATION ON BIOCREATIVE VII TRACK 3 VALID SUBSET.

Dataset	**Pre-trained Model**	**Strict F**
Biocreative	BioBERT-Large v1.1	65.42
Biocreative	BERTweet	67.73
Biocreative + SMMH4	BERTweet	68.22
Biocreative + SMMH4 + Konplik	BERTweet	70.98

For the test subset, we applied our best system configuration NeuroNER and Biocreative + SMMH4 + Konplik datasets and BiomedTwitter_English_200d model obtaining an f-score of 63.10% and BERT fine-tuning and Biocreative + SMMH4 + Konplik dataset and BERTweet model obtaining an f-score of 67.70% for offset detection and entity classification on strict evaluation. Results on the test subset can be found in Table 5.

TABLE 5 BEST UC3M-KONPLIK SYSTEMS RESULT FOR ENTITY CLASSIFICATION ON BIOCREATIVE VII TRACK 3 TEST SUBSET.

Model	**Precision (%)**	**Recall (%)**	**F-Score (%)**
NeuroNER and Biocreative + SMMH4 + KonPlik datasets + MedTwitter	91.00	48.30	63.10
Biocreative + SMMH4 + Konplik dataset + BERTweet	79.10	59.20	67.70

IV. CONCLUSIONS

Precision values obtained by the configuration based on NeuroNER and the combination of all available datasets go to 91%, but recall values are low (48%). This means that a lot of valid drug mentions are not properly identified. The HULAT-KONPLIK team is currently analyzing labeling errors to better understand the reasons. This requires analyzing the labeling guidelines to understand which words must be considered relevant for the task. On the other hand, it is very likely that out of vocabulary mentions also play a relevant role in the low recall produced by the model. We will include our final analysis in the final version of this paper for the BioCreative VII conference.

ACKNOWLEDGMENT

This work has been partially supported by the Madrid Government (Comunidad de Madrid- Spain) under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M17), and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation).

REFERENCES

1. Müller M, Salathé M, Kummervold PE. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. 2020 May 15 [cited 2021 Oct 12]; Available from: http://arxiv.org/abs/2005.07503

2. Dernoncourt F, Lee JY, Szolovits P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. 2017 [cited 2018 May 30]; Available from: http://arxiv.org/abs/1705.05487

3. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. Proc 2014 Conf Empir Methods Nat Lang Process [Internet]. 2014 [cited 2018 May 30];1532–43. Available from: http://aclweb.org/anthology/D14-1162

4. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional Semantics Resources for Biomedical Text Processing [Internet]. Vol. 5, Aistats. 2013 [cited 2018 Aug 18]. Available from: https://github.com/spyysalo/nxml2txt

5. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A. Advances in pre-training distributed word representations. Lr 2018 - 11th Int Conf Lang Resour Eval [Internet]. 2019 Dec 26 [cited 2021 Oct 12];52–5. Available from: https://arxiv.org/abs/1712.09405v1

6. Trask A, Michalak P, Liu J. sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings. 2015 Nov 19 [cited 2018 May 30]; Available from: http://arxiv.org/abs/1511.06388

7. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019 Sep 10;

8. Nguyen DQ, Vu T, Nguyen AT. BERTweet: A pre-trained language model for English Tweets. 2020 May 20 [cited 2021 Oct 12];9–14. Available from: https://arxiv.org/abs/2005.10200v2

9. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res [Internet]. 2004 Jan 1 [cited 2018 Jul 24];32(Database issue):D267-70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14681409

10. S K, A M-J, M K, C W. Cadec: A corpus of adverse drug event annotations. J Biomed Inform [Internet]. 2015 Jun 1 [cited 2021 Oct 12];55:73–81. Available from: https://pubmed.ncbi.nlm.nih.gov/25817970/

11. Alvaro N, Miyao Y, Collier N. TwiMed: Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms, and Their Relations. JMIR Public Heal Surveill [Internet]. 2017 Apr 1 [cited 2021 Oct 12];3(2). Available from: /pmc/articles/PMC5438461/

12. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality [Internet]. 2013 [cited 2018 Dec 3]. Available from: http://arxiv.org/abs/1310.4546

13. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. 2016 [cited 2018 May 31]; Available from: http://arxiv.org/abs/1607.04606

14. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. Association for Computational Linguistics (ACL); 2018. p. 2227–37.

15. McCann B, Bradbury J, Xiong C, Socher R. Learned in translation: Contextualized word vectors. Vols. 2017-Decem, Advances in Neural Information Processing Systems. 2017.

16. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding [Internet]. Vol. 1, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. 2019 [cited 2020 Jan 23]. Available from: http://arxiv.org/abs/1810.04805

17. Borthwick A, Sterling J, Agichtein E, Grishman R. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition [Internet]. Proceedings of the 6th Workshop on Very Large Corpora. 1998 [cited 2018 Aug 11]. Available from: http://acl.ldc.upenn.edu/W/W98/W98-1118.pdf

18. Sang EFTK, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition [Internet]. 2003 [cited 2018 Aug 15]. Available from: http://arxiv.org/abs/cs/0306050