Detection of Adverse Drug Reaction in Tweets Using a Combination of Heterogeneous Word Embeddings

Segun Taofeek Aroyehun
CIC, Instituto Politécnico Nacional
Mexico City, Mexico
aroyehun.segun@gmail.com

Alexander Gelbukh
CIC, Instituto Politécnico Nacional
Mexico City, Mexico
www.gelbukh.com

Abstract

This paper details our approach to the task of detecting reportage of adverse drug reaction in tweets as part of the 2019 social media mining for healthcare applications shared task. We employed a combination of three types of word representations as input to a LSTM model. With this approach, we achieved an F1 score of 0.5209.

1 Introduction

The social media mining for health care applications shared task aims to provide a benchmark for validating and comparing methods for healthcare applications using social media data (Weissenbacher et al., 2019). The focus of task 1 is on identifying adverse drug reaction as a medication related outcome. Participants on this task are expected to differentiate tweets as reporting adverse drug reaction or not and the performance metric is F1. This task demands that adverse drug reaction be distinguished from a similar and mostly confounding expression of the indication of a drug. The former is usually associated with the usage of the drug while the latter is a specification of the reason to use a drug. In addition, the task of detecting mention of adverse drug reaction is an extremely imbalanced binary classification task. About 1% of the training set are positive examples and approximately 99% are negative examples. Our approach is based on the combination of three different types of word embedding representations viz: character (Lample et al., 2016), non-contextual(Glove pre-trained on Twitter data) (Pennington et al., 2014), and contextual(BERT) (Devlin et al., 2018). The following section gives details of our model and training set-up. Section 3 shows the results of our experiments while we conclude and speculate on future directions in Section 4.

2 Model and Experimental Set-Up

We hypothesize that the different types of embeddings capture different relationships and their combination could help in the identification of adverse drug reaction in tweets. In our experiments, the word representation differs in two dimensions: whether they are pre-trained (Glove and Bert) or not (character embedding) and if they are contextual (Bert) or otherwise (Glove and Character embeddings). We briefly describe each representation:

- Character embedding - is a 50 dimensional representation of the characters in a word (how are they combined to form an embedding for the word). This representation is trained together with the model. It is based on a bidirectional LSTM. The advantage of character-based representation for social media text is that it eliminates the out-of-vocabulary problem which results from noise in the form of misspellings and abbreviations in word-based representation such as Glove. Also, this representation is specific to the task and domain of the training set.
- Glove (twitter) - is a 100 dimensional representation pre-trained on a huge twitter corpus. We expect this to contribute by reflecting the language of twitter users. However, the embedding is not a contextual one.
- BERT (en, base-uncased) - is a general domain contextual word representation where the representation of a word is based on other words in its context (sentence). The BERT base model which is not cased gives a word embedding of dimension 768. It has enabled state-of-the-art results on several NLP tasks. However, to the best of our knowledge, its application to social media text is limited.

Proceedings of the 4^{th} Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 133–135
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

	No. of Examples
train	24202
dev	6051
test	4575

Table 1: Details of the Data

In order to leverage some of the benefits of the representations above, we concatenated these representations for a given word in a tweet. This combination is of dimension 918. A linear layer then project this representation into a dimension of 256. This projection is meant to serve as a distillation step and/or as a fine-tuning step. The resulting representation is fed into an LSTM layer with hidden size of 512 to sequentially model a tweet. Finally, a dense layer is used as the classifier.

The model was trained for 100 epochs with learning rate annealing factor of 0.5 using SGD as the optimizer and a batch size of 8. We used a train-dev split of 80:20. Table 1 shows the number of training, validation, and evaluation examples used in our experiment. Weissenbacher et al. (2019) provide details on the collection and annotation of the dataset. Based on the validation split, a model with the best F1 score is saved during training as the best model. With the best model, we made predictions on the unseen evaluation examples as our first submission (sub1 in Table 2). Our second submission (sub2 in Table 2) was based on the model at the 100th epoch or the last epoch as training is terminated if learning rate becomes too small. Our experiments were performed using the Flair framework (Akbik et al., 2018).

3 Results

Table 3 shows the results obtained on the test set. We achieved our best submission with the final model with an F1 of 0.5209. This result ranks above the average score of all participants in the task with average F1, precision, and recall of 0.5019, 0.5351, 0.5054 respectively (Weissenbacher et al., 2019). Table 3 shows the results obtained from our ablation experiments with respect to the contributions of the different embedding representations and the distillation/fine-tuning step. The F1 scores reported are based on the model that achieved the best F1 score on the validation set during training. We observed a minimal drop in performance (0.0045) when we removed the fine-tuning layer. This suggests that the fine-tuning layer either hurts performance or the dimension of the resulting fine-tuned representation is an important parameter to tune with our approach. We assessed the contribution of the three embedding representations to performance by removing one at a time from the model while keeping our fine-tuning strategy. When the character embedding word representation is absent, a performance drop of 0.0238 is observed. When the BERT representation is removed, the performance improved by 0.0025. Without the Glove embedding, the performance increased by 0.0005. This result is consistent with our perceived advantages and disadvantages of the three embedding representations. With the character embedding contributing the most to the model performance. Remarkably, the removal of BERT and Glove leads to improved performance. This can be attributed to the out-of-vocabulary problem with Glove and domain mismatch in the case of BERT.

	P	**R**	**F1**
sub1	0.6145	0.4457	0.5167
sub2	**0.6203**	**0.4489**	**0.5209**

Table 2: Performance on the Test Set (Scores as provided by the organizers)

Model	F1
emb comb w/ fine tuning	0.9015
emb comb w/o fine tuning	0.9060
emb comb w/ fine tuning w/o character	0.8777
emb comb w/ fine tuning w/o Glove	0.9020
emb comb w/ fine tuning w/o BERT	0.9040

Table 3: Performance of Model Variants on the Validation Split

4 Conclusion

This paper outlines our participation in the 2019 social media mining for healthcare application challenge on identifying the reportage of adverse drug reaction in tweets. Our approach is based on the combination of three different types of embedding representations and a fine-tuning strategy. With this approach, we made two submissions using a model that achieved the best F1 score on the validation data and with a model trained till the last epoch possible. The latter gave a better performance. Through ablation experiments, we observed that our fine-tuning strategy results in a small drop in performance contrary to our expectation. In addition, the different word representations contribute to different degrees. The character embedding representation makes the most significant contribution, without it the model performance drops while there is a marginal performance improvement when both Glove and BERT representation are removed from the model.

As a follow-up work, we would like to investigate other fine-tuning or distillation approaches as well as parameter tuning of the size of the fine-tuning layer. It is also interesting to examine the impact of normalizing tweets and identifying usage expressions as an auxiliary task.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

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.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Task at ACL 2019. In *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task.*