Detecting and Extracting of Adverse Drug Reaction Mentioning Tweets with Multi-Head Self-Attention

Suyu Ge[†], Tao Qi[†], Chuhan Wu, Yongfeng Huang
Department of Electronic Engineering, Tsinghua University Beijing 100084, China
{gesy17,qit16,wuch15,yfhuang}@mails.tsinghua.edu.cn

Abstract

This paper describes our system for the first and second shared tasks of the fourth Social Media Mining for Health Applications (SMM4H) workshop. We enhance tweet representation with a language model and distinguish the importance of different words with Multi-Head Self-Attention. In addition, transfer learning is exploited to make up for the data shortage. Our system achieved competitive results on both tasks with an F1-score of 0.5718 for task 1 and 0.653 (overlap) / 0.357 (strict) for task 2.

1 Introduction

Automatic adverse drug reaction (ADR) detection and extraction are of great social benefits to public health, with which pharmacovigilance (Sarker and Gonzalez, 2015) can be performed at a broader and more automatic level. Recent research focus their attention on online public sources such as tweets due to their availability and authenticity (Onishi et al., 2018; Adrover et al., 2015; Salathé and Khandelwal, 2011).

The SMM4H shared task is proposed (Weissenbacher et al., 2019) to enhance ADR recognition. Task 1 is a binary classification task between ADR mentioned tweets and drug name only tweets, followed by task 2 to extract the particular position of ADR entities. Based on the work we did last year (Wu et al., 2018), we extend our previous model with hierarchical tweet representation and multi-head self-attention (HTR-MSA) to a model using both hierarchical tweet representation and attention (HTA) to jointly participate both tasks. Moreover, additional features and a language model are incorporated to enhance the semantic representations. In task 1, transfer learning on a smaller dataset is exploited. In task 2, we add a CRF layer for the named entity recognition task.

2 Our Approach

Our HTA model can be divided into the following three parts: hierarchical word representation, hierarchical tweet representation and tweet classification, which are introduced as follows.

2.1 Hierarchical Word Representation

In order to combat out-of-vocabulary medical terminology, misspellings and user created abbreviations, we propose a character modeling at a lower level before traditional word representation. We denote the character sequence of i_{th} word as $\mathbf{w}_i = [\mathbf{C}_{i,1}, \mathbf{C}_{i,2}, ..., \mathbf{C}_{i,N}]$, where N is the word length. A character embedding matrix $\mathbf{M}^c \in \mathcal{R}^{V \times D}$ is utilized to convert $\mathbf{w}_i$ into vector sequence $\mathbf{E}_i^c = [\mathbf{e}_{i,1}, \mathbf{e}_{i,2}, ..., \mathbf{e}_{i,N}]$, where V denotes the character vocabulary size and D denotes the dimension of character embedding.

After a character embedding is generalized, character-level convolutional neural network is employed to capture local combined character feature. Assuming the window size of CNN filters is $2w + 1$ and $\mathbf{U}_c$, b_c are kernel and bias parameters respectively, a convolutional representation $\mathbf{h}_{i,j}$ of character embedding vectors from position $j - w$ to $j + w$ is formed as follows:

$$\mathbf{h}_{i,j} = ReLU(\mathbf{U}_c \times \mathbf{e}_{i,(j-w):(j+w)} + b_c) \quad (1)$$

To remove unnecessary information, we apply the max pooling to pertain only the most salient feature of the i_{th} word.

Other features are added at a word level, such as word2vec-twitter (Godin et al., 2015) word embedding, pos-tag from NLTK library (Bird et al., 2009) and sentiment lexicon[1]. To strengthen the

[†]Equal contribution.

[1]http://sentiwordnet.isti.cnr.it/

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 96–98
Florence, Italy, August 2, 2019.

medical meaning of word representation, word appearance in SIDER 4.1 medical lexicon[2] is transformed to one-hot vector as additional feature. Besides, the language model ELMo embedding (Peters et al., 2018) is incorporated to overcome the shortage of limited data and get better semantic meaning. Since ELMo contains character level information in their model, it fits better to our task goal than other language model that utilizes a fixed word look-up dictionary.

The final output of our hierarchical word representation is the concatenation of character representation, word embedding, pos-tag, sentiment lexicon, medical lexicon feature and language model output.

2.2 Hierarchical Tweet Representation

We first send word representation obtained in the previous module to a Bi-LSTM layer to encode long-distance information. The Bi-LSTM output of a sentence of length M is denoted as $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_M]$.

The second layer takes advantage of multi-head self-attention (Vaswani et al., 2017) to mine internal relation between words in the same sentence. In our layout, the representation vector $\mathbf{m}_{i,j}$ of the j_{th} word learned by the i_{th} attention head is computed by weighted summation of $\mathbf{H}$:

$$\hat{\alpha}^i_{j,k} = \mathbf{h_j}^T \mathbf{U_i} \mathbf{h_k}, \tag{2}$$

$$\alpha^i_{j,k} = \frac{\exp(\hat{\alpha}^i_{j,k})}{\mathbf{\Sigma}^M_{m=1} \exp(\hat{\alpha}^i_{j,m})}, \tag{3}$$

$$\mathbf{m}_{i,j} = \mathbf{W}_i(\mathbf{\Sigma}^M_{m=1} \alpha^i_{j,m} \mathbf{h}_m), \tag{4}$$

$\mathbf{U}_i$ and $\mathbf{W}_i$ are the parameters of the i_{th} self-attention head, and $\alpha^i_{j,k}$ represents the related weight between j_{th} and k_{th} words. After concatenating outputs from h different self-attention heads, we get the representation $\mathbf{m}_j = [\mathbf{m}_{1,j}; \mathbf{m}_{2,j}; ...; \mathbf{m}_{h,j}]$ of the j_{th} word.

2.3 Tweet Classification

For task 1, we use an additive attention mechanism to selectively combine word representations. The model is trained with a cost-sensitive weighted loss function (Santos-Rodrguez et al., 2009). Sentence level binary labels are then generated for task 1. However, in task 2 word level labels are needed, so we use a CRF layer to predict word level entity tags after self-attention vectors produced in the lower level.

[2]http://sideeffects.embl.de/

3 Experiments

3.1 Experiment Settings

In our experiments,the word embedding we use is 400 dimension and Bi-LSTM network has 2×200 units. The CNN network has 400 filters with window size of 3. There are 16 heads in the multi-head self-attention network, and the output dimension of each head is 16. Adam is selected as the optimizer.

Transfer learning is conducted on the CADEC medical ADR dataset (Karimi et al., 2015) first in task 1. However, we do not adopt this method in task 2 due to the relative small training dataset of this task. For the word classification, we train for this task a marginal CRF with probabilities as output.

3.2 Experiment Results

Detailed evaluation score is illustrated in table 1, which illustrated the effectiveness of our approach. In task 1, our model outperforms the average score among all participants by 0.070. In task 2, the improvement on relax F1 is also significant, we improve 0.115 on relax F1 and 0.040 on strict F1. Besides, compared to the best model we submitted for task 1 last year (Wu et al., 2018), which reached a 0.522 F1 score, our method with the language model and transfer learning improves the original model by 0.050.

4 Conclusion

We design HTA, a hierarchical tweet representation and attention model for SMM4H shared task 1 and 2, our model attains high evaluation scores on both tasks and generates promising application value.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant number 2018YFC1604002, and the National Natural Science Foundation of China under Grant numbers U1836204, U1705261, U1636113, U1536201, and U1536207.

References

Cosme Adrover, Todd Bodnar, Zhuojie Huang, Amalio Telenti, and Marcel Salathé. 2015. Identifying adverse effects of hiv drug treatment and associated

	Task 1	**Task 2** (relax)	**Task 2** (strict)
Precision	0.467	0.612	0.329
Recall	0.738	0.698	0.390
F1 Score	0.572	0.653	0.357
Average F1 (mean)	0.502	0.538	0.317
F1 Range (mean)	0.3308	0.486	0.422

Table 1: Evaluation Results.

sentiments using twitter. *JMIR public health and surveillance*, 1(2):e7.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Frderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van De Walle. 2015. Multimedia lab @ acl w-nut ner shared task: Named entity recognition for twitter microposts using distributed word representations. In *Workshop on User-generated Text*.

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.

Takeshi Onishi, Davy Weissenbacher, Ari Klein, Karen O'Connor, and Graciela Gonzalez-Hernandez. 2018. Dealing with medication non-adherence expressions in twitter. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task*, pages 32–33, Brussels, Belgium. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Marcel Salathé and Shashank Khandelwal. 2011. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology*, 7(10):e1002199.

Ral Santos-Rodrguez, Dario Garca-Garca, and Jess Cid-Sueiro. 2009. Cost-sensitive classification based on bregman divergences for medical diagnosis.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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 and Shared Task*.

Chuhan Wu, Fangzhao Wu, Junxin Liu, Sixing Wu, Yongfeng Huang, and Xing Xie. 2018. Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 34–37.