NTTMUNSW SYSTEM FOR ADVERSE DRUG REACTIONS EXTRACTION IN TWITTER DATA

CHEN-KAI WANG[*1], ONKAR SINGH[*2], HONG-JIE DAI[†]
[1]*Department of Computer Science and Information Engineering, National Taitung University, 369, Sec. 2, University Rd.*
Taitung, 95092, Taiwan
[2]*Graduate Institute of Biomedical Informatics, Taipei Medical University, 250 Wu-Xin Street*
Taipei, 110, Taiwan
xd150612@yahoo.com.tw, onkarnims2009@gmail.com, hjdai@nttu.edu.tw

JITENDRA JONNAGADDALA
School of Public Health and Community Medicine,University of New South Wales, Samuels Ave, Kensington
NSW, 2033, Australia
jitendra.jonnagaddala@unsw.edu.au

TONI ROSE JUE
Prince of Wales Clinical School, University of New South Wales
NSW 2052, Australia
t.jue@unsw.edu.au

USMAN IQBAL, EMILY CHIA-YU SU, SHABBIR SYED ABDUL, JACK YU-CHUAN LI
Graduate Institute of Biomedical Informatics, Taipei Medical University, 250 Wu-Xin Street
Taipei, 110, Taiwan
usman.iqbal85@gmail.com, {emilysu, drshabbir, jack}@tmu.edu.tw

This work developed methods to recognize adverse drug reaction (ADR)-related information from Twitter data based on two machine learning algorithms, conditional random fields (CRFs) and recurrent CRFs. Different word representation methods were developed for the ADR recognition task, including the token normalization, and two state-of-the-art word embedding methods, namely word2vec and global vectors. Three runs were submitted to the Pacific Symposium on Biocomputing 2016 Social Media Mining shared task. The best run achieved an F-score of 0.540, which was ranked second in the shared task on social media mining (Task 2) of Pacific Symposium on Biocomputing.

---

[*] C.-K. Wang and O. Singh contributed equally to this work.

[†] Corresponding author

1. Introduction

Adverse drug reaction (ADR) is defined as an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product due to an injury caused by taking a medication (Edwards and Aronson, 2000). Many studies reported that ADRs is a major public health problem with deaths and hospitalizations in millions, and associated costs of about seventy-five billion dollars annually (Ahmad, 2003; Lazarou, et al., 1998). This work proposes method for recognizing information related to ADRs from Twitter data by using several off-the-shelf NLP tools. Two state-of-the-art supervised sequential labelling algorithms were used to implement the ADR-named entity recognition (NER) system. The first is the conditional random field model (Lafferty, et al., 2001), which has been successfully applied in several sequential labeling tasks and demonstrated an outstanding performance (Dai, et al., 2015; Eltyeb and Salim, 2014). The second is based on recurrent neural networks (RNNs) (Tomá, et al., 2010). Recently, Yao, et al. (2014) showed that combining CRFs with RNNs (R-CRFs) can improve the language understanding performance. This work studies their effect for the ADR-NER task.

2. Methods

The ADR-NER task is formulated as a sequential labelling task by using the IOBES scheme, which has been demonstrated as the best tag scheme for most NER tasks (Dai, et al., 2015; Liu, et al., 2015). Several computation linguistics techniques were applied on a given tweet. First, Tokenizer (Owoputi, et al., 2013) was used to tokenize a tweet into tokens and generate the part-of-speech (PoS) information for each of them. Each token was then processed by the spell check, Hunspell, to correct spelling errors and analyzed its morphemes.

2.1. *Context Features*

For a given token, its surrounding tokens were extracted as features. For a target token, its context is defined as the token itself with the three preceding tokens, the current token, and its three following tokens. The following three context representation methods were explored in this work.

2.1.1. *Normalized Token*

A numerical normalization preprocessing method is used to convert numeral parts in each token to one representative numeral (Tsai, et al., 2006). The main benefits of this representation includes the reduction of the number of features, the possibility in transforming unseen features into seen features and the improvement of the accuracy of feature weight estimation. In addition, special symbols, such as "@" and "#", are normalized by removing them.

2.1.2. *Word vector representation*

This work employed word2vec (Mikolov, et al., 2013) and global vectors (Pennington, et al., 2014) to generate vector representations of all unique tokens in a corpus of around one hundred thousand tweets. The corpus was compiled by searching Twitter website for a set of predefined query terms to collect 7-days of tweets. The query terms were collected by analyzing the ADR-NER training

and development sets to list the described ADRs, their related drugs as well as the ADR terms listed in the lexicon compiled by Nikfarjam, et al. (2015). In word2vec, the continuous bag of words scheme was used. After generating the token vectors, the simple K-means method was performed to group tokens into 200 different clusters. With the above word clusters, the value of the context feature is defined as the cluster number associated for the token. If the current token does not have an associated clusters, the normalized token will be used.

2.2. *Gazette Features*

Four gazette features were implemented in this work. The first gazette feature was implemented as a binary feature that indicates whether or not the current normalized token partially matches with the gazette entry in a given gazette file. The second gazette feature encoded matched tokens using the IOB scheme. The ADR lexicon created by Leaman, et al. (2010), was employed as the gazette file for matching ADR terms. In addition to the ADR lexicon, the tokens annotated with Drug tags were collected from the training and development sets to form the drug lexicon. A similar gazette feature was developed based on the created drug lexicon.

2.3. *Linguistic Morphology Features*

For each normalized token, the lemma and stem generated by the employed spell checker and the snowball stemmer were selected as features. The PoS of each normalized token generated by Twokenizer and the prefix of the token were also encoded as features.

3. Results and Discussion

Table 1. The precision (P), recall (R) and F-measure (F) on the test set of the shared task.

CRF Model (Run 1)			**R-CRF (Run 2)**			**CRF Model' (Run 3)**		
P	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
0.782	0.412	**0.540**	0.718	**0.416**	0.526	0.778	0.414	**0.540**

Table 1 shows the overall recognition performance of the submitted three runs. The first and second runs were based on the CRF and the R-CRF models, respectively. The third run was also based on the CRF model but its word2vector information was replaced with the one created by Nikfarjam, et al. (2015), which was created based on one million unlabeled user sentences. Overall, our models achieved satisfactory precisions but low recalls. The best recall is 0.416 that was achieved by R-CRF. The results show that the inclusion of the larger corpus to generate the word vector (Run 3 vs. Run 1) does not illustrate significant advantages. Our error analysis reveals that all of the three models failed to recognize all of the mentioned drugs in the development set due to all of them being out-of-vocabulary terms. A further look at the false negative cases reveals that tweet users do not use technical terms recorded in any drugs or ADR lexicons. Instead, they tend to use abbreviations, or any other possible descriptive expressions. The above phenomenon leads to the low recall of all the trained models. The false positives mainly come from the recognition of the therapeutic side effect consequences of the use of a drug. Surprisingly, the R-CRF model doesn't show an improvement of F-measure over the CRF model based on the same feature sets. An error analysis will be conducted in the future.

Acknowledgments

The authors would like to thank the organizers of PSB 2016 Social Media Mining Shared Task and anonymous reviewers for their valuable feedback and comments. This research was supported by the Ministry of Science and Technology of Taiwan grant MOST-104-2221-E-143-005-.

Reference

1. Ahmad, S.R. Adverse drug event monitoring at the Food and Drug Administration. *Journal of general internal medicine* 2003;18(1):57-60.
2. Dai, H.-J.*, et al.* Recognition and Evaluation of Clinical Section Headings in Clinical Documents Using Token-Based Formulation with Conditional Random Fields. *BioMed Research International* 2015;2015.
3. Dai, H.J.*, et al.* Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *J Cheminform* 2015;7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S14.
4. Edwards, I.R. and Aronson, J.K. Adverse drug reactions: definitions, diagnosis, and management. *Lancet* 2000;356(9237):1255-1259.
5. Eltyeb, S. and Salim, N. Chemical named entities recognition: a review on approaches and applications. *Journal of Cheminformatics* 2014;6(1):17.
6. Lafferty, J., McCallum, A. and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In, *Proceedings of the 18th International Conference on Machine Learning (ICML)*. 2001. p. 282–289.
7. Lazarou, J., Pomeranz, B.H. and Corey, P.N. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *JAMA* 1998;279(15):1200-1205.
8. Liu, S.*, et al.* Feature engineering for drug name recognition in biomedical texts: feature conjunction and feature selection. *Comput Math Methods Med* 2015;2015:913489.
9. Mikolov, T.*, et al.* Distributed representations of words and phrases and their compositionality. In, *Advances in neural information processing systems*. 2013. p. 3111-3119.
10. Nikfarjam, A.*, et al.* Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc* 2015;22(3):671-681.
11. Owoputi, O.*, et al.* Improved part-of-speech tagging for online conversational text with word clusters. In.: Association for Computational Linguistics; 2013.
12. Pennington, J., Socher, R. and Manning, C.D. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* 2014;12:1532-1543.
13. Tomá, M.*, et al.* Recurrent neural network based language model. In, *Proceedings of the 11th Annual Conference of the International Speech Communication Association*. Makuhari, Chiba, Japan; 2010. p. 26-30.
14. Tsai, R.T.-H.*, et al.* NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinformatics* 2006;7(Suppl 5):S11.
15. Yao, K.*, et al.* Recurrent Conditional Random Field for Language Understanding. In, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Florence, Italy; 2014.

16. R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez, "Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks," *in Proceedings of the 2010 workshop on biomedical natural language processing*, 2010, pp. 117-125