Komp'juternaja Lingvistika i Intellektual'nye Tehnologii 2017 vol.1 N16, pages 155-166

Identifying disease-related expressions in reviews using conditional random fields

Miftahutdinov Z., Tutubalina E., Tropsha A.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

As the as the volume of user-generated content in social media expands so do the potential benefits of mining social media to learn about patient conditions, drug indications, and beneficial or adverse drug reactions. In this paper, we apply Conditional Random Fields (CRF) model for extracting expressions related to diseases from patient comments. Our method utilizes hand-crafted features including contextual features, dictionaries, clusterbased and distributed word representation generated from unlabeled user posts in social media. We compare our CRF-based approach with deep recurrent neural networks and a dictionary-based approach. We examine different word embeddings generated from unlabeled user posts in social media and scientific literature. We show that CRF outperformed other methods and achieved the F1-measures of 69.1% and 79.4% on recognition of disease-related expressions in the exact and partial matching exercises, respectively. Qualitative evaluation of disease-related expressions recognized by our feature-rich CRF-based approach demonstrates the variability of reactions from patients with different health conditions.

Keywords

Conditional Random Fields, CRF, Disease named entity recognition, Information extraction, Opinion expressions

References

[1] Benton A., Ungar L., Hill S., Hennessy S., Mao J., Chung A., Holmes J. H. (2011), Identifying potential adverse effects using the web: A new approach to medical hypothesis generation, Journal of biomedical informatics, Vol. 44(6), pp. 989-996.

[2] Brown P. F., Desouza P. V., Mercer R. L., Pietra V. J. D., Lai J. C. (1992), Classbased n-gram models of natural language, Computational linguistics, 18(4), pp. 4 67-479.

[3] Chernyshevich M. (2014), IHS R&D Belarus: Cross-domain extraction of product features using conditional random fields, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 309-313.

[4] Cho K., Van Merriënboer B., Bahdanau D., Bengio Y. (2014), On the properties of neural machine translation: Encoder-decoder approaches, arXiv preprint a rX iv:14 0 9.1259.

[5] Choi Y., Cardie C. (2010), Hierarchical sequential learning for extracting opinions and their attributes, Proceedings of the ACL 2010 conference short papers, pp. 269-274.

[6] Freifeld C. C., Brownstein J. S., Menone C. M., Bao W., Filice R., Kass-Hout T., Dasgupta N. (2014), Digital drug safety surveillance: monitoring pharmaceutical products in twitter, Drug safety, 37(5), pp. 343-350.

[7] Deftereos S. N., Andronis C., Friedla E. J., Persidis A., Persidis A. (2011), Drug repurposing and adverse event prediction using high-throughput literature analysis, Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 3(3), pp. 323-334.

[8] Gareev R., Tkachenko M., Solovyev V., Simanovsky A., Ivanov V. (2013), Introducing baselines for Russian named entity recognition, Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, pp. 329-342.

[9] Hochreiter S., & Schmidhuber, J. (1997), Long short-term memory. Neural computation, Vol. 9(8), pp. 1735-1780.

[10] Huang C. C., Lu Z. (2016), Community challenges in biomedical text mining over 10 years: success, failure and the future, Briefings in bioinformatics, 17(1), p p. 132-14 4.

[11] Irsoy O., Cardie C. (2014), Opinion Mining with Deep Recurrent Neural Networks. EMNLP, pp. 720-728.

[12] Jagannatha A. N., Yu H. (2016). Bidirectional RNN for Medical Event Detection in Electronic Health Records, Proceedings of NAACL-HLT, pp. 473-482.

[13] Jakob N., Gurevych I. (2010), Extracting opinion targets in a single- and cross-domain setting with conditional random fields. Proceedings of the 2010 conference on empirical methods in natural language processing, pp. 1035-1045.

[14] Karimi S., Metke-Jimenez A., Kemp M., Wang C. (2015), Cadec: A corpus of adverse drug event annotations, Journal of biomedical informatics, Vol. 55, pp. 73-81.

[15] Kinga D., Adam J. B. (2015), A method for stochastic optimization, International Conference on Learning Representations (ICLR).

[16] Lafferty J., McCallum A., Pereira F. (2001), Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Proceedings of the eighteenth international conference on machine learning, ICML, Vol. 1, pp. 282-289.

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

[18] Lee H. C., Hsu Y. Y., Kao H. Y. (2015), An enhanced CRF-based system for disease name entity recognition and normalization on BioCreative V DNER Task, Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, pp. 226-233.

[19] Li D., Afzal N., Mojarad M. R., Elayavilli R. K., Liu S., Wang Y., Liu H. (2015), Resolution of chemical disease relations with diverse features and rules, Proceedings of the fifth BioCreative challenge evaluation workshop, pp. 280-285.

[20] Li L., Jin L., Huang D. (2015), Exploring recurrent neural networks to detect named entities from biomedical text, Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 279-290.

[21] Liu P., Joty S. R., Meng H. M. (2015), Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings, Proceedings of EMNLP, pp. 1433-14 4 3.

[22] Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. (2015), SentiRuEval: testing object-oriented sentiment analysis systems in Russian, Proceedings of International Conference Dialog, Vol. 2, pp. 3-13.

[23] Lu Y., Ji D., Yao X., Wei X., Liang X. (2015), CHEMDNER system with mixed conditional random fields and multi-scale word clustering, Journal of cheminformatics, Vol. 7 (1).

[24] Metke-Jimenez A., Karimi S. (2015), Concept extraction to identify adverse drug reactions in medical forums: A comparison of algorithms, arXiv preprint arXiv:1504.06936.

[25] Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, pp. 3111-3119.

[26] Moghaddam S., Ester M. (2011), ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews, Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 665-674.

[27] Nikfarjam A., Sarker A., O'Connor K., Ginn R., Gonzalez G. (2015), Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, Journal of the American Medical Informatics Association, pp. 1-11.

[28] Pyysalo S., Ginter F., Moen H., Salakoski T., Ananiadou S. (2013), Distributional semantics resources for biomedical text processing, Proceedings of Languages in Biology and Medicine.

[29] Qu L. et al. (2016) Named Entity Recognition for Novel Types by Transfer Learning//arXiv preprint arXiv:1610.09914.

[30] Stanovsky G., Gruhl D., Mendes P. N. (2017) Recognizing Mentions of Adverse Drug Reaction in Social Media Using Knowledge-Infused Recurrent Models.

[31] Titov I., McDonald R. T. (2008), A Joint Model of Text and Aspect Ratings for Sentiment Summarization, ACL, Vol. 8, pp. 308-316.

[32] Tjong K., Sang E. F., De Meulder F. (2003), Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Vol. 4, pp. 142-147.

[33] Wang W. (2016), Mining adverse drug reaction mentions in twitter with word embeddings. Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing.

[34] Wei C. H., Peng Y., Leaman R., Davis A. P., Mattingly C. J., Li J., Lu Z. (2015), Overview of the BioCreative V chemical disease relation (CDR) task, Proceedings of the fifth BioCreative challenge evaluation workshop, pp. 154-166.

[35] Wei Q., Chen T., Xu R., He Y., Gui L. (2016), Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks, Journal of Biological Databases and Curation.

[36] Wong T. L., Bing L., Lam W. (2011), Normalizing web product attributes and discovering domain ontology with minimal effort, Proceedings of the fourth ACM international conference on Web search and data mining, pp. 805-814.