Conditional Random Fields for Biomedical Named Entity Recognition Revisited

Xie-Yuan Xie
Freelancer
xieyuanxienlp@gmail.com

Abstract

Named Entity Recognition (NER) is a key task which automatically extracts Named Entities (NE) from the text. Names of persons, places, date and time are examples of NEs. We are applying Conditional Random Fields (CRFs) for NER in biomedical domain. Examples of NEs in biomedical texts are gene, proteins. We used a minimal set of features to train CRF algorithm and obtained a good results for biomedical texts.

1 Introduction

Although, Biomedical NER (BioNER) has received attention by many researchers, the task remains very challenging due to ambiguity, synonymy, nested and multi-word BioNE [1]. To tackle these challenges, many efforts have been put by researchers to develop efficient BioNER systems.

Approaches for BioNER include dictionary-based approach, rule-based approach, Machine Learning (ML) approach and hybrid approach. Machine learning approach uses ML algorithms, such as Support Vector Machines (SVM) [2], Conditional Random Fields (CRF) [3], Maximum Entropy (ME) [4] and Hidden Markov Model (HMM) [5] to build a model to detect and classify BioNEs. ML approach for BioNER depends on large training data and feature engineering process. Feature engineering includes extraction and selection of the features which will be used for learning the parameters of ML algorithm. Hybrid approach combines two or more different approaches for BioNER [6].

Classical ML classification algorithms have been applied in different NLP tasks such as Author Profiling [7, 8] and for multi lingual [9, 10, 11]. Applying such algorithms requires hand features engineering and the performance of the model depends basically on the experience of feature designer.
2 Related Work

Nayel and Shashirekha [6], proposed an ensemble-based approach for BioNER using CRF and SVM classifiers with different segment representation schemes. A multi-task learning approach has been applied to BioNER using neural network architecture [12]. The best f-score achieved for BioNER using NCBI dataset is 80.73%. Leaman and Lu [13], used semi-markov models to build TaggerOne, a tool for BioNER. TaggerOne reported a high competitive f-score of 82.9% for NCBI dataset. Nayel and Shashirekha [14] integrated dictionary based features with deep model for disease NEs. Sunil and Ashish [15], designed a Recurrent Neural Network model for BioNER. They utilized Convolution Neural Network (CNN) for representing character embeddings and bidirectional LSTM for word embeddings. They evaluated their system on NCBI dataset and achieved f-score of 79.13%. Nayel et al [16] proposed a deep learning based approach to improve multi-word BioNER extraction. Wei et al [17], developed an ensemble-based system for Disease-NER. At the base level, they built CRF with a rule-based post-processing system and Bi-RNN based system. At the top level, they used SVM classifier for combining the results of the base systems.

In this paper, we designed a CRFs-based model for BioNER. In this paper a minimal set of features has been used to train the classification algorithm. The rest of the paper organized as follows: section 2 shows the related work. In section 3, the methodology details are given and in section 4 experiments details are given. Conclusion is given in section 5.

3 Methodology

3.1 Conditional Random Fields

CRF [18] are undirected graphical models, a special case of which corresponds to conditionally trained probabilistic finite state automata. Being conditionally trained, CRF can easily incorporate a large number of arbitrary, non-independent features while having efficient procedures for non-greedy finite-state inference and training. CRF is used to calculate the conditional probability of values on designated output nodes given values on the designated input nodes.

When implementing CRF for the NER task, the sequence of the words in a sentence is assumed as an observation sequence and the state sequence as its corresponding tag sequence. The context window plays an important role in building the classification model, because CRF assumes that the observations (words) are dependent. This assumption makes CRF a suitable classification model for NER.
3.2 Performance Evaluation

The performance of our system is reported in terms of f-score [6]. F-score is a harmonic mean of Precision (P) and Recall (R) and is calculated as follow:

\[ f\text{-}score = \frac{2 \times P \times R}{P + R} \]

4 Experiments and Results

4.1 Datasets

We applied CRF algorithm for JNLPBA [19], NCBI [20], BC2GM [21], BC5CDR [22] and i2b2 [23] datasets.

4.2 Training

To train the model the following set of features has been extracted:

- Word length
- PoS tags
- Word affixes
- Context window
- Word Normalization
- Orthographic Features
- Stop Words

4.3 Results and Discussion

The detailed results in terms of Precision (P), Recall (R) and f1-measure for all datasets are given in Table 1. These results illustrate that CRF classifiers perform better with minimal features. The main point is the resources for CRF training are minimized in terms of time complexity, space and computations required. The results are still needs improvements such as applying CRF with embeddings as input space.

5 Conclusion

While deep learning models performs better in many cases, although the use of classical ML approaches still interesting for many tasks specially which needs more resources such as pre-trained word representation, huge memory and time consuming. In this paper, we used a classical ML approach with minimal feature set.
| Dataset  | R     | P     | f1-measure |
|---------|-------|-------|-----------|
| JNLPBA  | 70.52%| 81.76%| 75.73%    |
| NCBI    | 71.77%| 83.31%| 77.11%    |
| BC2GM   | 72.08%| 83.17%| 77.23%    |
| BC5CDR  | 71.35%| 82.73%| 76.62%    |
| i2b2    | 71.77%| 83.41%| 77.16%    |

Table 1: Detailed results of all datasets

References

[1] H. L. Shashirekha and H. A. Nayel. A comparative study of segment representation for biomedical named entity recognition. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages 1046–1052, Sept 2016.

[2] Ki-Joong Lee, Young-Sook Hwang, Seonho Kim, and Hae-Chang Rim. Biomedical named entity recognition using two-phase model based on \{SVMs\}. *Journal of Biomedical Informatics*, 37(6):436 – 447, 2004. Named Entity Recognition in Biomedicine.

[3] U Kanimozhi and D Manjula. A crf based machine learning approach for biomedical named entity recognition. In Recent Trends and Challenges in Computational Models (ICRTCCM), 2017 Second International Conference on, pages 335–342. IEEE, 2017.

[4] Sujan Kumar Saha, Sudeshna Sarkar, and Pabitra Mitra. Feature selection techniques for maximum entropy based biomedical named entity recognition. *Journal of Biomedical Informatics*, 42(5):905 – 911, 2009. Biomedical Natural Language Processing.

[5] Shaojun Zhao. Named entity recognition in biomedical texts using an hmm model. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications, JNLPBA ’04, pages 84–87, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[6] Hamada Nayel and H L Shashirekha. Improving ner for clinical texts by ensemble approach using segment representations. In Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), pages 197–204, Kolkata, India, December 2017. NLP Association of India.

[7] Hamada A. Nayel and H. L. Shashirekha. Mangalore University INLI@FIRE2018: Artificial Neural Network and Ensemble based Models for INLI. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, Working Notes of FIRE 2018 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 6-9, 2018., volume 2266 of *CEUR Workshop Proceedings*, pages 110–118. CEUR-WS.org, 2018.
[8] Hamada A. Nayel and H. L. Shashirekha. Mangalore-University@INLI-FIRE-2017: Indian Native Language Identification using Support Vector Machines and Ensemble Approach. In Prasenjit Majumder, Mandar Mitra, Parth Mehta, and Jainisha Sankhavara, editors, *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017.*, volume 2036 of *CEUR Workshop Proceedings*, pages 106–109. CEUR-WS.org, 2017.

[9] Hamada A. Nayel, Walaa Medhat, and Metwally Rashad. BENHA@IDAT: Improving Irony Detection in Arabic Tweets using Ensemble Approach. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 401–408. CEUR-WS.org, December 2019.

[10] Hamada A. Nayel. NAYEL@APDA: Machine Learning Approach for Author Profiling and Deception Detection in Arabic Texts. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 92–99. CEUR-WS.org, December 2019.

[11] Hamada A. Nayel and Shashirekha H. L. DEEP at HASOC2019: A machine learning framework for hate speech and offensive language detection. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 336–343. CEUR-WS.org, 2019.

[12] Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18(1):368, Aug 2017.

[13] Robert Leaman and Zhiyong Lu. Taggerone: Joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 2016.

[14] Hamada A. Nayel and Shashirekha H. L. Integrating Dictionary Feature into A Deep Learning Model for Disease Named Entity Recognition. *CoRR*, abs/1911.01600, 2019.

[15] Sunil Sahu and Ashish Anand. Recurrent neural network models for disease name recognition using domain invariant features. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2216–2225, Berlin, Germany, August 2016. Association for Computational Linguistics.

[16] Hamada A. Nayel, H L Shashirekha, Hiroyuki Shindo, and Yuji Matsumoto. Improving Multi-Word Entity Recognition for Biomedical Texts. *International Journal of Pure and Applied Mathematics*, 118(16):301–3019, 2017.
[17] Qikang Wei, Tao Chen, Ruifeng Xu, Yulan He, and Lin Gui. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *Database*, 2016.

[18] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML ’01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[19] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. Introduction to the bio-entity recognition task at jnlpba. In Nigel Collier, Patrick Ruch, and Adeline Nazarenko, editors, *Proceedings of the International Joint Workshop on NLP in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 70–75, Geneva, Switzerland, August 2004. COLING.

[20] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47(Supplement C):1 – 10, 2014.

[21] Larry Smith, Lorraine K. Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W. John Wilbur. Overview of biocreative ii gene mention recognition. *Genome Biology*, 9(2):S2, Sep 2008.

[22] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068, 2016.

[23] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.