Disease Named Entity Recognition (D-NER) Evaluation

Xie-Yuan Xie (xieyuanxienlp@gmail.com)

Short Report

Keywords: Named Entity Recognition, Biomedical NLP

Posted Date: September 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-911654/v1

Disease Named Entity Recognition (D-NER) Evaluation

Xie-Yuan Xie
Freelancer, Data Science
xieyuanxienlp@gmail.com

Abstract

Named Entity Recognition (NER) is a key task in Natural Language Processing (NLP). In medical domain, NER is very important phase in all end-to-end systems. In this paper, we investigate the performance of NER for disease (D-NER). TaggerOne was evaluated on 52 cardiovascular-related clinical case reports against hand annotation for diseases. Different training sets have been used to evaluate the performance of TaggerOne as a famous tool for NER in biomedical domain.

1 Introduction

Natural Language Processing (NLP) has been a research area that gained significant recent interest [1]. Expansions in the volume of unstructured free text has created a strong need for automated methods to identify, classify, normalize, and annotate these unstructured data into semantically meaningful structured data for knowledge generation. Traditional NLP methodologies have been rule-based or heuristic-based, encoding in the linguistic structure of English along with domain-specific semantic relations into algorithms to identify named entities. More recent machine-learning-based methods have attempted to broaden the generality regarding breadth of topics, with techniques that can apply to a wide variety of topics. These methods attempt to be agnostic to specific text types, with only the training set specific to a knowledge domain to evaluate performance across domains. Recent work [2] have demonstrated that domain-specific structural information can show significant improvement by combining these two approaches. By encoding semantic information specific to a

domain via a well-chosen training set, significant performance was observed. Thus, the role of domain-specific NLP models is a valuable but poorly-characterised area, particularly as it applies to biomedical texts and clinical case reports for cardiovascular diseases.

The structuring of biomedical texts has been a growing area of interest growing in parallel with the more general expansion of unstructured free text. PubMed, the central repository of biomedical texts, has been growing exponentially, and an increasingly major challenge is organizing the vast corpus of knowledge for easier access and knowledge generation. NLP research in biomedical data is unusually challenging in comparison to NLP research in other text areas due to a paucity of well-annotated gold standards – understanding biomedical or clinical texts require specific education and precludes crowdsourcing or large extant gold standard corpora. Thus, the investigation of NLP approaches for biomedical texts is a research area of specific interest.

Recent approaches to try to organize PubMed using named entity recognition and normalization, called PubTator [3, 4, 5, 6], have been applied to biomedical texts, but no in-depth analysis of performance exists in the literature. DNorm [7, 8, 9, 10], the technology behind PubTator, uses conditional random fields for normalization of disease names. While these methodologies have demonstrated high statistical performance metrics, the reasons and characteristics for errors have been less well-described. Even less known is the application of these methods in cardiovascular clinical texts such as clinical case reports. Investigating the performance of NER algorithms in the cardiovascular clinical texts will inform future research approaches on areas of improvement. This paper focuses on the specific types of errors that PubTator, disease name normalization, and conditional random fields generate specifically in the context of cardiovascular clinical case reports.

2 Methodology

The frame work of our overview is shown in Figure 1. PubMed Central was queried using the term "heart failure". The results were limited to full-text clinical case reports in English, with no restrictions on publication date or journal. 52 reports were randomly selected. On the 52 reports, automatic annotation was performed first using PubTator Central and hand-annotation of disease names (the "gold standard") by a single individual was also performed. PubTator annotations are stored in XML,

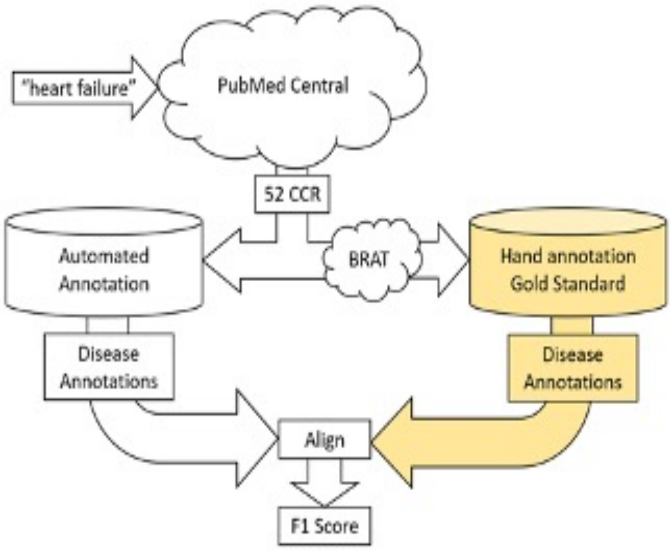

Figure 1: Framework to Evaluate the Performance of Disease NER

while the gold standard annotations are stored in BRAT format. Disease annotations were extracted from PubTator XML using regular expressions and aligned to the gold standard annotations. Missing annotations and false positive annotations were tabulated, and an F1 score was computed for each report. To provide context for the PubTator Central F1 score, two additional comparisons were run by training TaggerOne on the NCBI Disease Corpus and BioCreative V Chemical-Disease Relation (BC5CDR) corpus. F1 scores based on models trained on these datasets were also calculated.

A deep dive was also performed on the types of errors observed and to look for any patterns. Tabulated charts of missed and false positive disease annotations of each of the 52 reports were hand-scrutinized and categorized by type. Proposals for potential improvements were conceived based on these results.

3 Results

Below are the F1 score distributions of the three models – PubTator, NCBI Disease Corpus, CB5CDR – on the 52 reports.

The mean F1 score was 0.42 for PubTator Central, 0.28 for the NCBI Disease Corpus-trained model, and 0.36 for the BC5CDR-trained model. Seeing that the PubTator Central outperforms both trained models, it seems likely that the PubTator Central

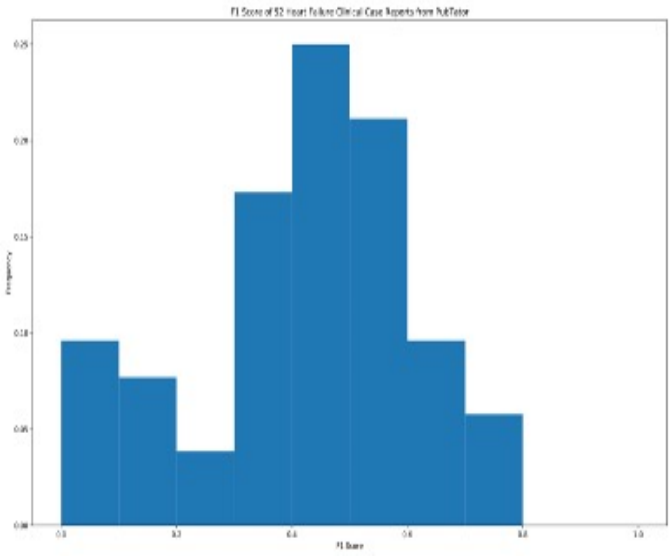

Figure 2: F1 Distribution of PubTator Central

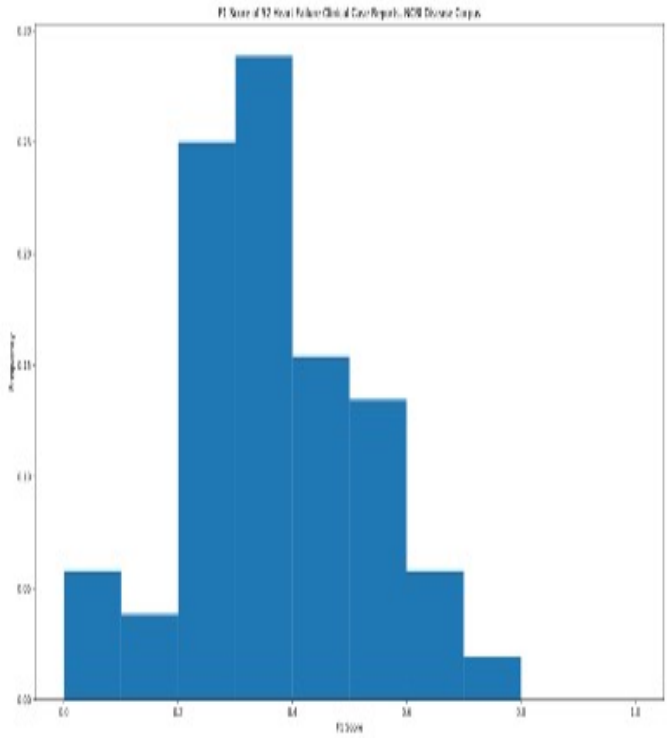

Figure 3: F1 Distribution of NCBI Disease Corpus-trained model

is trained on additional datasets or an expanded corpus beyond either just the NCBI Disease Corpus or the BC5CDR corpus. Some of the left-side outliers occurred due to the very short length of some of the reports, which would magnify errors in propor-

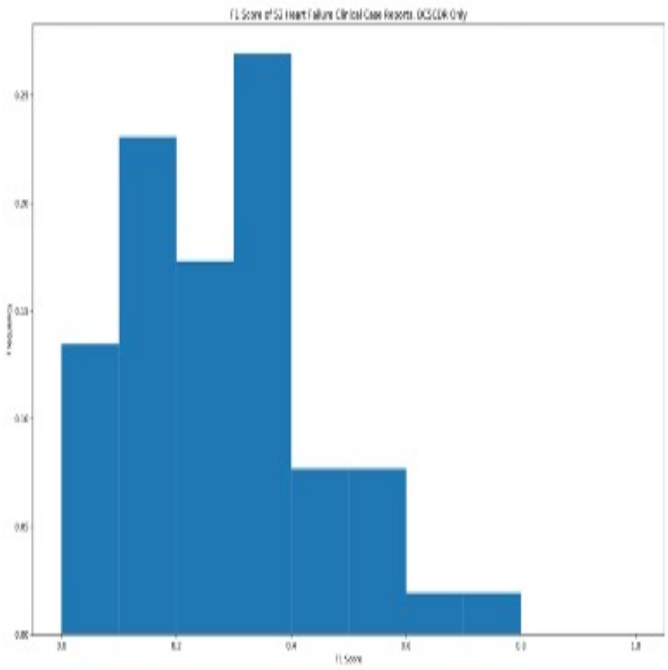

Figure 4: F1 Distribution of BC5CDR-trained model

tion to successes.

Here is a distribution of the types of errors found:

The errors were divided into misses and false positives. They were also categorized by type. The major types of errors were:

1. Named entities that the algorithm believe are diseases but do not qualify medically as a disease, merely as a symptom. For example: ""pain", "dyspnea", "jaundice", "death".

2. Incorrect acronym resolution. This included either not identifying disease acronyms or misidentifying acronyms of entities that were not diseases as diseases. Examples: "GBS" (Guillain–Barré syndrome), "AMI" (acute myocardial infarction), "AKI" (acute kidney injury), "DIC" (disseminated intravascular coagulation), "PDA" (patent ductus arteriosus), "STEMI" (ST-elevated myocardial infarction).

3. Span errors, where the NER algorithm does not capture the entire length of the term correctly. For example, incomplete terms such as "kidney injury", "Barre syndrome", "coagulation", "ST-elevation".

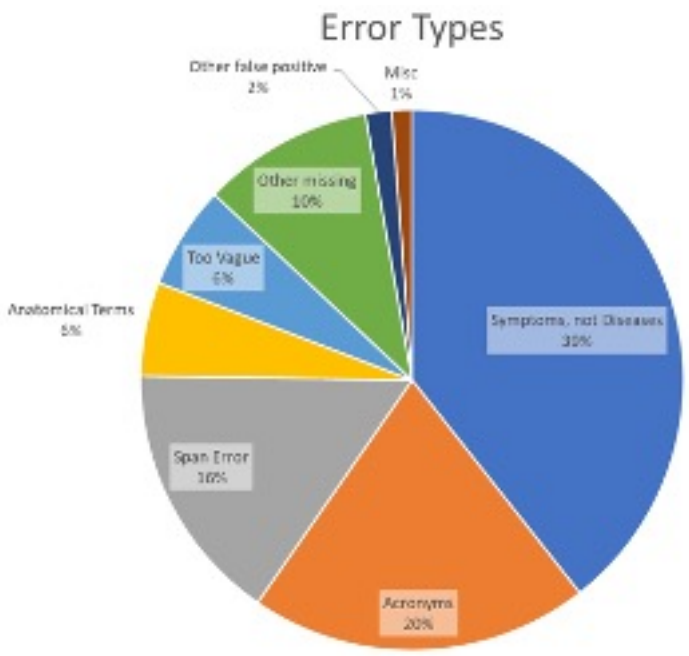

Figure 5: Proportion of Types of Errors Found

4. Anatomical terms that the algorithm incorrectly identify as diseases, such as "anastomosis", "patent ductus arteriosus".

5. Terms too vague to qualify as a disease, such as "deficiency", "malformation", "tumor".

4 Discussion

The aim of this paper is to examine the performance of NER algorithms on cardiovascular clinical case reports. The performance evaluating TaggerOne trained on various models against a hand-annotated gold standard shows that there is still room for performance improvement for NER on clinical texts. The performance also demonstrates areas of improvement for acronym resolution and incorrect term spanning. In particular, it is important to note that the training sets used were not based on texts from clinical case reports nor are they focused on cardiovascular disease. For the NCBI Disease Corpus, disease names are collected from biomedical research articles and their abstracts. For the BC5CDR corpus, it is a selection of disease and chemical names pulled from basic science research articles [4, 11]. Neither corpus uses clinical case reports, and neither specializes in cardiovascular diseases either.

Thus, it is unsurprising that ambiguities in biomedical naming, such as acronym resolution, fail noticeably in the test set. The superior performance of PubTator Central, however, seems to indicate that by merely improving and expanding the training corpus, significant performance improvement can be achieved without overhauling the underlying algorithm. It is thus likely that a model trained specifically on cardiovascular clinical case reports will show significant improvement over baseline. Clinical case reports possess a particular structure and writing style, such as a heavy use of acronyms, and training the model on these more representative names should show improvements in span error and acronym resolution, two of the biggest contributors to error [12]. The impact of using a well-constructed training corpus is demonstrated in this project, as is also the need for hand annotation. Some of the errors – vagueness and misattribution of symptom to disease – rely on meta-clinical knowledge that is challenging to identify from text in an unsupervised way. From a clinical perspective, symptoms and disease form a continuum, with stereotyped syndromes with physiologically coherent etiologies such as infective endocarditis on the disease end, and generalized and ambiguous physiological responses to a disturbance such as tachycardia on the symptom end. Two approaches towards identifying where the line between symptom and disease are hypothesized [13, 14, 15]. The first approach is a supervised way. By constructing a consistent corpus of identified disease names by hand, TaggerOne might be able to recognize disease from symptoms. Expanding the NCBI Disease Corpus with cardiovascular clinical case reports should yield improvements. However, hand annotation is challenging and unable to scale, so it might be possible to develop a weakly supervised method that pre-processes the training set bootstrapped from a smaller hand-annotated corpus or another corpus such as the NCBI Disease Corpus onto a larger unannotated corpus. These methods have not been tested and are potential future directions. With the room for improvement notwithstanding, the potential for creating structure in otherwise free text is demonstrated in these models. As disease terms form a notable minority of words in a text creating a highly unbalanced training set, an F1 score of up to 0.5 shows a clear ability to distinguish named entities from nonentities. Named entity recognition is an important step in interpreting clinical case reports and other biomedical texts for generating a semantic structure for the texts [1, 16, 17, 18]. This proof of concept demonstrates the state of the art for NER technology as applied to what is currently available as the most representative of clinical texts. Using NER for semantically structuring text can also be applied on electronic medical records, which currently are large bodies of

unstructured text. Identifying structure from unstructured medical record texts can open many new opportunities in clinical informatics areas such as clinical decision support, patient cohort identification, and disease nosology [19, 20].

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 evaluated the implementation of deep learning models to disease named entity recognition.

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] Hamada A. Nayel. *Biomedical Named Entity Recognition*. PhD thesis, Mangalore University, 2018.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

[8] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, March 2002.

[9] Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

[10] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.