Method Entity Extraction from Biomedical Texts

Waqar Bin Kalim
Department of Computer Science
The University of Western Ontario
London, Ontario, Canada
waqaar.199@gmail.com

Robert E. Mercer
Department of Computer Science
The University of Western Ontario
London, Ontario, Canada
mercer@csd.uwo.ca

Abstract

In the field of Natural Language Processing (NLP), extracting method entities from biomedical text has been a challenging task. Scientific research papers commonly consist of complex keywords and domain-specific terminologies, and new terminologies are continuously appearing. In this research, we find method terminologies in biomedical text using both rule-based and machine learning techniques. We first use linguistic features to extract method sentence candidates from a large corpus of biomedical text. Then, we construct a silver standard biomedical corpus composed of these sentences. With a rule-based method that makes use of the Stanza dependency parsing module, we label the method entities in these sentences. Using this silver standard corpus we train two machine learning algorithms to automatically extract method entities from biomedical text. Our results show that it is possible to develop machine learning models that can automatically extract method entities to a reasonable accuracy without the need for a gold standard dataset.

1 Introduction

Method Entity Extraction from unstructured biomedical text has been an important, yet challenging and somewhat under-examined, Natural Language Processing (NLP) task. Especially in the field of biomedicine, the automatic extraction of methodology names and terminologies is imperative. With thousands of research papers published each week, the biomedical research community is constantly creating new terminologies. As a result, it becomes difficult for researchers to find relevant information.

Wang et al. (2022) give a good discussion of method entities and provide the following definition: “named entities that represent specific methods”. More specifically, “method entities in the academic literature are nouns or noun phrases representing the specific ways, means, and channels used to solve tasks or problems proposed by the authors, including sub-categories such as discipline-specific methods, software, models, algorithms, and metrics”.

Biomedical method entities are a type of biomedical named entities. Biomedical named entity recognition (NER) is defined as the task of recognizing and categorizing entity names in the biomedical domain (Lee et al., 2004). As mentioned by Song et al. (2018), biomedical NER faces difficulties for various reasons. The first one is the increasing rate of newly created terminologies and keywords requiring new rules and patterns to be manually added to the rule-based methods, which can be a tedious and time-consuming task. Regarding biomedical method entity recognition, because it has been little studied (e.g., (Houngbo and Mercer, 2012)), a comprehensive definition of what constitutes a biomedical method entity beyond that given by Wang et al. (2022) still needs to be determined. Secondly, with information extraction tasks, the same words can have different meanings and significance in terms of the context.

In this study, our main contribution is the exploration of a rule-based approach to create a silver standard corpus annotated for method entities[1]. Regarding the rule-based approach, rules to extract method entities based on patterns of universal dependency relations (Universal Dependencies, 2014) between words will be designed. To evaluate the machine-made silver standard corpus, this corpus will be used as training data for two machine learning models, Conditional Random Fields (CRF) (Lafferty et al., 2001) and BiLSTM (Graves and Schmidhuber, 2005). CRF was used by Houngbo and Mercer (2012) and our CRF results are compared to their results as a baseline. The BiLSTM results indicate that a larger corpus will be needed

[1] The corpus and code to create it are available at https://github.com/waqarkalim/method-mention-extraction-from-biomedical-text

Proceedings of the 29th International Conference on Computational Linguistics, pages 2357–2362
October 12–17, 2022.

for neural learning models.

The structure of the remainder of this paper is as follows. In Section 2, we will review the background and related work. In Section 3, we will provide our research contributions. We will review the methodology of our research in Section 4. In Section 5, we will review the results. And in Section 6, we will conclude the paper and suggest directions for the continuation of this study.

2 Background and Related Work

Named Entity Recognition (NER) is an application of Natural Language Processing (NLP) where entities are tagged according to various semantic and syntactic rules. Surveys of research in the field of NER (Nadeau and Sekine, 2007; Wang et al., 2022) indicate that automatic NER extraction has received significant coverage in the past three decades. For the study reported in the current paper, we are interested in the extraction of a subcategory of named entities, method entities in particular, from biomedical literature. In the biomedical literature, extraction of named entities has tended to focus on biological entities (for example, Settles (2004) and Habibi et al. (2017)). Extraction of method entities, in particular, has received some attention, but the vast majority extract this type of named entity from non-biomedical literature (Wang et al., 2022). One exception is Houngbo and Mercer (2012) who deal with method entities in biomedical research articles. Lam et al. (2016) extract methodology terms as well as sleep disorder entities from biomedical literature that focusses on sleep disorders. Zhao et al. (2019) propose a new annotation scheme, manually annotate 3,088 resource citations (algorithms are the only methods included in these resources) found in biomedical and non-biomedical literature, and use BiLSTM as the machine learning method.

One aspect of this current work is to create a silver standard corpus for method entities in biomedical text. The automatic method to generate this corpus will use Stanza's dependency parser (Qi et al., 2020) with the biomedical packages (Zhang et al., 2021). This parser produces graph structures whose edges are labelled with universal dependencies (Universal Dependencies, 2014).

To evaluate using the silver standard corpus, we will use Conditional Random Fields (Lafferty et al., 2001) and BiLSTM (Graves and Schmidhuber, 2005), the machine learning techniques used in other studies (Houngbo and Mercer, 2012; Chiu and Nichols, 2016).

3 Research Contributions

Few researchers have addressed the question of automatic method entity extraction from biomedical text. Generating a human-labelled corpus is time-consuming. This study aims to address these issues.

1. We have created rule-based methods to use Stanza's universal dependencies while extracting a wider variety of method entities compared to Houngbo and Mercer (2012) and have successfully created an accurate silver standard corpus (precision score of 97.59) prepared from full text biomedical articles selected from the PubMed Central dataset with method entities annotated automatically.

2. By training on this silver corpus, we have improved on the performance benchmarks provided by Houngbo and Mercer (2012).

4 Methodology

Because Houngbo and Mercer (2012) is the only previous study that investigates the same problem, we use that work as the baseline and compare our results with those reported there.

In the initial stage of the study, we prepare a collection of sentences that contain mentions of method names, or "method sentences", using the method proposed by Houngbo and Mercer (2012). By employing the properties of anaphoric relations between sentences, we can collect the "method sentences" in a convenient and feasible manner. We collect these sentences by scanning through research papers using the Unix command grep and selecting some number of sentences that precede any sentence containing the words "this method". This approach successfully generates a corpus containing solely "method sentences".

After the corpus creation has been completed, the next stage involves utilizing linguistic rules and patterns to automatically label method entities. In this study, for tagging the entities, we will be using the IOB tagging format. In the IOB tagging scheme, every token is labelled as B-label if the token is the beginning of a named entity, I-label if it is inside a named entity but not the first token within the named entity, or O otherwise (Lample et al., 2016). Leveraging rule-based methods allows for labelling the method mentions without the use of any pre-existing training data; additionally, a secondary benefit of this approach is the potential of introducing new rules and patterns based on the linguistic features of the terminologies. This step is an essential aspect of our research as it allows for implementing an accurate silver standard dataset.

After the rule-based methods have been applied, traditional and neural learning techniques can be explored in combination with the newly developed silver standard dataset. The primary benefit of utilizing a machine learning approach is the ability to generalize beyond the limits of the rules and patterns that are manually defined in the rule-based approach. In this stage of the research, we will explore two machine learning algorithms related to Natural Language Processing tasks: Conditional Random Field (CRF) models (Lafferty et al., 2001) and the Bidirectional Long Short Term Memory (BiLSTM) models (Graves and Schmidhuber, 2005). We opted for choosing these algorithms as CRF models are trained for sequence segmentation and labelling and BiLSTM models have been used in other named entity extraction research (Chiu and Nichols, 2016; Zhao et al., 2019).

5 Results

5.1 Silver Corpus Creation

In this study, we curated a collection of 2,839 biomedical research papers from which to derive our method entity silver corpus. Based on the findings reported by Torii and Vijay-Shanker (2005) that nearly all antecedents can be found within two sentences from the demonstrative anaphors, we used the technique suggested by Houngbo and Mercer (2012) and employed the anaphoric relations between sentences to find the "method sentence" candidates. So, to generate our corpus, we searched through our collection of papers for sentences that begin with the anaphor "This method" and then selected the three sentences that precede it for our corpus. By selecting three sentences rather than two, we achieve an extra layer of certainty that the selected sentences contain at least one method entity. As a result, we retrieved 10,974 potential "method sentences".

An example of a retrieval is:
Sentence 1: In tracheal samples, YCW increased concentrations of mucosal IgA compared to Control (P < 0.05).
Sentence 2: No significant differences were observed between Vaccine and Coccidiostat.
Sentence 3: The effect of different treatments on cell-mediated immune response was examined by the cutaneous basophilic hypersensitivity test.
Sentence 4: This method reveals the status of the T-cell response.
In this example, Sentence 4 contains the "this method" anaphor. The sentences found in the silver corpus would be Sentences 1, 2, and 3. Sentence 3 contains the antecedent "cutaneous basophilic hypersensitivity test" which is a method entity.

A manual investigation of these sentences suggests that most of the method entities in our corpus are sequences of words represented by the following examples:

- Tukey's biweight method
- naive KNN method
- 10-fold cross-validation test
- Roche Amplicor Cystic Fibrosis test
- bimolecular fluorescence complementation analysis
- Felsenstein's independent comparison method
- statistical total correlation spectroscopy analysis method
- MANOVA-based scoring method
- protein sequence Jukes-Cantor model
- utaneous basophilic hypersensitivity test

From these examples, we can observe how the rules and patterns to extract our method entities would look like. First, all of these mentions end with key suffixes (as observed by Houngbo and Mercer (2012)) that would correspond to most method entities: method, analysis, test, model, algorithm, etc. In addition, we have dependency-parsed the method mention candidate sentences with Stanza (Qi et al., 2020) using the biomedical and clinical model packages included in the Stanza toolkit (Zhang et al., 2021). Investigating these dependency parses, all of these method entities have at least one universal dependency (UD) compound relation, most of them have at least one amod UD relation, and some of them have at least one nmod:poss UD relation. A compound UD is a modifier that relates to a noun and itself is a noun, whereas an amod UD is an adjectival modifier that serves to modify a noun or pronoun but itself is an adjective. An nmod:poss UD is a modifier that serves to show possessives. After generating the corpus which contains 10,974 potential "method sentences", we used linguistic rules and patterns to programmatically extract the method entities depending on the dependency relationships between the words. Based on these observations, we created three rules based solely on these UD relationships. These three rules were able to find 1338 method entities in our corpus in total; 629 for Rule 1, 680 for Rule 2, and 29 for Rule 3. The rules work as follows:

Rule 1: In a sentence, if there is a subtree with at least one compound relation, retrieve all the words between the first compound word to the last word of that subtree plus the subtree root (i.e., one of the key suffixes mentioned above) as a method entity.

Rule 2: In a sentence, if there is a subtree with exactly one compound relation and at least one amod relation, retrieve all the words between the first amod/compound word to the last word of that subtree plus the subtree root as a method entity.

Rule 3: In a sentence, if there is a subtree with exactly one nmod:poss relation, retrieve all the words between the nmod:poss word to the last word of that subtree plus the subtree root as a method entity.

The rules stated above are different from the rule used by Houngbo and Mercer (2012), a regular expression composed of POS tags and key suffixes, to find the method entities; whereas the rules in this study use universal dependencies provided by Stanza pre-trained on biomedical text.

With these rules, the rule-based model was able to achieve a precision score of 97.59, which is better than expected. Unfortunately, due to the sheer amount of data in our corpus, we were unable to manually determine the recall score, and accordingly, an F-1 score for our rule-based approach.

Using this rule-based model, we tag the extracted labels using IOB tagging in order to create our silver standard dataset which can be used in Section 5.2.

5.2 Machine-learning Approach

When the rule-based method to label the silver standard corpus has completed, we are now ready to investigate the machine-learning techniques. This study will investigate two machine learning models: 1) a traditional Conditional Random Field (CRF) model, and 2) a neural Bidirectional Long Short Term Memory (BiLSTM) model. Both are supervised machine learning models which require training data that is labelled with the feature to be learned. As our training dataset, we will use the silver standard corpus from Section 5.1.

System	P	R	F1
Conditional Random Field	83.58	85.49	84.53
Houngbo and Mercer (2012) CRF	81.80	75.00	78.26
BiLSTM	68.42	39.39	50.00

Table 1: Precision (P), Recall (R), and F1-Score (F1) for the Machine Learning Methods

CRF models are a framework for developing probabilistic models for segmenting and labelling sequence data (Lafferty et al., 2001). BiLSTM models (Graves and Schmidhuber, 2005) are a form of recurrent neural networks that can understand the context of a sentence quite well. BiLSTM models work well for NLP tasks as they can contextually scan through text in both forward and backward directions. For the word embeddings, the BiLSTM model uses BioWordVec, an open set of biomedical word embeddings that combines subword information learned from unlabeled biomedical text with a widely-used biomedical controlled vocabulary (Zhang et al., 2019).

Table 1 shows the results for each of the machine learning models. We observe from Table 1 that the highest performing machine learning model in this study (CRF) outperforms the machine learning model results of Houngbo and Mercer (2012) by a precision score of 1.78 pp, a recall score of 10.49 pp, and an F-1 score of 6.27 pp.

Our findings are based on inaccurate metrics, so the results should thus be treated with some caution. However, because this inaccuracy is due to the true positives being thought of as false positives, the actual precision and recall should be higher than what is displayed in Table 1. As an example, the CRF model produced 20 predictions that were labelled as not method entities in the Houngbo and Mercer (2012) gold-standard testing data, however, a manual check shows that 17 out of those 20 predictions actually are method entities.

The results for BiLSTM are lower than what would be predicted by other research that has used this neural architecture (Zhao et al., 2019). We believe that this is due to insufficient training samples. So, we did not investigate any other BiLSTM-based architectures, leaving this to future studies when we have built a larger silver corpus.

6 Conclusions and Future Work

In this paper, we explored various methodologies to automatically extract method entities from biomedical text. In the initial step, we created a corpus containing candidate method sentences using anaphoric relations. Next, we investigated a rule-based method using information provided by a dependency parse to IOB tag the method entities in the corpus. This silver standard corpus is the main contribution of our work and has been made publicly available. To evaluate the quality of this corpus, we trained two machine learning methods using this silver standard corpus to automatically extract method entities from biomedical text.

The evidence from this study shows that using a dependency parser that is pre-trained on biomedical vocabulary allows for precise extraction of method entities within the scope of the rules as shown in Section 5.1. Additionally, the results from Section 5.2 and Table 1 show how the CRF model outperforms the results from Houngbo and Mercer (2012) and show the potential of machine learning models to accurately generalize outside the scope of the rules defined in Section 5.1.

Our future work will include

1. improving on the anaphoric method to gather candidate method entity sentences.

2. creating a wider variety of rules and patterns for our rule-based approach to create a more comprehensive silver standard corpus. In this study we noticed phrases such as “a method that uses regular expressions to look for section headings” and *method phrase* followed by “developed by *one or more names*” (e.g., “Robust Multi-Array Analysis developed by Irizarry”) which is like the possessive form in our Rule 3 above. These more complex patterns would be amenable to the dependency parse methodology. However, these patterns may not be amenable to an IOB-type annotation. And they include but do not end with one of the key suffixes (i.e., method, analysis, test, model, algorithm). So, this move may not be a simple extension of what was presented here.

3. developing a larger silver standard corpus. The BiLSTM results strongly suggest that the corpus is too small for neural learning methods.

4. detailing aspects of the silver standard corpus. Understanding what makes this corpus better than previous ones will inform further development of a definition of the text span of a method entity.

5. investigating more sophisticated machine learning models, such as a BiLSTM-CRF model and a BiLSTM-CNN-CRF model, to better evaluate the silver standard corpus and to improve the method entity performance beyond what has been achieved in previous works. Adding a CRF layer on top of a BiLSTM model, as well as adding a CNN and CRF layer on top of a BiLSTM model have proven to improve performance in a few sequence labelling problems.

Acknowledgements

This research is partially funded by The Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant to Robert E. Mercer. We also acknowledge the helpful comments provided by the reviewers.

References

Jason P. C. Chiu and Eric Nichols. 2016. Named entity recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6):602–610.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Hospice Houngbo and Robert E. Mercer. 2012. Method mention extraction from scientific research papers. In *Proceedings of COLING 2012*, pages 1211–1222.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. 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.

Calvin Lam, Fu-Chi Lai, Chia-Hui Wang, Mei-Hsin Lai, Nanl Hsu, and Min-Huey Chung. 2016. Text mining of journal articles for sleep disorder terminologies. *PLoS One*, 11(5):e0156031.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Ki-Joong Lee, Young-Sook Hwang, Seonho Kim, and Hae-Chang Rim. 2004. Biomedical named entity recognition using two-phase model based on SVMs. *Journal of Biomedical Informatics*, 37(6):436–447.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *LingvisticæInvestigationes*, 30(1):3–26.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA '04)*, pages 104–107.

Hye-Jeong Song, Byeong-Cheol Jo, Chan-Younng Park, Jong-Dae Kim, and Yu-Seop Kim. 2018. Comparison of named entity recognition methodologies in biomedical documents. *BioMedical Engineering OnLine*, 17(158).

Manabu Torii and K. Vijay-Shanker. 2005. Anaphora resolution of demonstrative noun phrases in medline abstracts. In *Proceedings of the Conference Pacific Association for Computational Linguistics (PACLING)*, pages 332–339.

Universal Dependencies. 2014. Universal dependencies. https://universaldependencies.org. Accessed: 2021-04-30.

Yuzhuo Wang, Chengzhi Zhang, and Kai Li. 2022. A review on method entities in the academic literature: extraction, evaluation, and application. *Scientometrics*.

Yijia Zhang, Qingyu Chen, Zhihao Yang, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6, Article number: 52.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.

He Zhao, Zhunchen Luo, Chong Feng, and Yuming Ye. 2019. A context-based framework for resource citation classification in scientific literatures. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*, pages 1041–1044.