# Method Mention Extraction from Biomedical Text

Final Progress Report

Waqar Kalim
COMPSCI 4490Z
Department of Computer Science
Western University

March 14, 2021

Project Supervisor: Dr Robert Mercer, Computer Science Department
Course Instructor name: Prof. Nazim Madhavji

## Glossary

**Method Mention**: Sequence of words that name a method in biomedical text  
**NLP**: Natural Language Processing is a branch of Artificial Intelligence that involves the understanding of human language  
**NER**: Named Entity Recognition is a subtask of Information Extraction that involves extracting named entities such as a name, a person, a company, etc. from unstructured text.  
**UD**: Universal Dependency is a framework for annotating grammar across various languages in a consistent manner.  
**ML**: Machine Learning is a branch of Artificial Intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.  
**RNN**: Recurrent Neural Networks (RNN) is a class of Artificial Neural Networks where connections between nodes form a directed graph along a temporal sequence.  
**CRF**: Conditional Random Fields is a class of statistical modelling method often applied in pattern recognition and machine learning and used for structured prediction.  
**LSTM**: Long Short-Term Memory model is an RNN architecture used in the field of deep learning.  
**Bi-LSTM**: Bidirectional Long Short-Term Memory model is a sequence processing model that consists of two LSTMs: one taking the input in a forward direction, and the other in a backward direction.  
**SVM**: Support Vector Machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis  
**POS tagging**: Part-of-Speech tagging is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context.  
**Stanza**: Stanza is a Python natural language analysis package that comprises a collection of accurate and efficient tools for many human languages in one place.  
**Gold Standard Corpus**: A Gold Standard Corpus is a human-made dataset that is created through human means.  
**Silver Standard Corpus**: A Silver Standard Corpus is a machine-made dataset that is created through automatic means.

*Definitions have been drawn from online sources.*

## Structured Abstract

In the field of Natural Language Processing (NLP), extracting method mentions 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 extract method terminologies from scientific text using both rule-based and machine learning techniques. We use linguistic features to extract method sentences and use the Stanza dependency parsing module for the rule-based methods to construct a Silver Standard biomedical corporus to be used with various machine learning algorithms. This thesis will provide a set of methodologies to automatically extract method mentions from scientific corpora. Our results show that it is possible to develop machine learning models that can automatically extract method mentions to a reasonable accuracy without the need for a gold standard dataset.

Keywords: method, information extraction, named entity recognition, Stanza

## 1) Introduction

Method Mention Extraction from unstructured text has been an important field in Natural Language Processing (NLP). Especially in the field of biomedicine, the automatic extraction of methodology names and terminologies has been 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.

Biomedical named entity recognition (NER) is defined as the task of recognizing and categorizing entity names in biomedical domains [1]. As mentioned in [2], biomedical NER faces difficulties for various reasons. The first one is the increasing rate of newly created terminologies and keywords, which requires new rules and patterns to be manually added to the rule-based method, which can be a tedious and time-consuming task. Secondly, with information extraction tasks, the same words can have different meanings and significance in terms of the context.

In this study, we will explore both a rule-based approach and a machine learning-based approach. With the rule-based approach, universal dependency relations between words will be used to create specific rules and patterns to extract method mentions. The use of rule-based methods allows for results without the use of any pre-existing training/testing data. With the machine learning approach, a machine-made silver standard corpus will be used as training/testing data and a variety of different machine learning algorithms will be implemented and compared.

The structure of the rest of the report is as follows. In Section 2, we will review the background and related works. In Section 3, we will review the research objectives. In Section 4, we will review the methodology of our research. In Section 5, we will review the results. In Section 6, we will discuss some of the threats to the validity, implications, limitations, etc. of our results. And in Sections 8 and 9, we will conclude the paper and review what work can be done in the future for the continuation of this study.

## 2) Background and Related Work

### 2.1) Named Entity Recognition in Biomedical text

- NER is an application of Natural Language Processing (NLP) where entities are tagged according to various semantic and syntactic rules. Numerous studies have been conducted on this subject, employing a variety of novel approaches and methodologies. Studies like [3], [4], and [5] indicate that automatic terminology extraction has received a lot of coverage in the past.
- Especially in the field of biomedicine, extraction of terminologies and methodologies has been an imperative as well as challenging task. Other studies, such as [6] and [7] are predominantly based on biomedical NER.
- This study delves deeper into the task of NER and specifically focuses on the automatic extraction of method mentions in the field of biomedicine.

### 2.2) Automatic method mention extraction from scientific research papers

- Over the last three decades, NER has received much attention [4], yet only a few studies, such as [7] and [8], have focused on the subtask of method mention extraction.

- Biomedical NER can be thought of as a sequence segmentation problem, as stated in [5]. Conditional Random Field (CRF) models have shown to be useful towards sequential labelling and part-of-speech tagging tasks [8].
- Similar to [8], the goal of this study is to extract method terminologies from biomedical research papers and to provide method mentions for a future research project that will populate a lexical resource

### 2.3) Analysis and Research Gap

_Analysis:_ *Lee et al.'s* approach to biomedical NER [1] achieved a precision score of 74.4 and a recall score of 75.2 using multi-class support vector machines (SVMs) created by combining several binary SVMs [9]. Instead of using SVMs, we used a CRF model which achieved a precision score of 83.58 and a recall score of 85.49 while being able to accurately generalize outside the limits of its training data. *Houngbo and Mercer's* approach to automatic method mention extraction [8] through rule-based methods achieved a precision score of 85.40 using rules based on simple part-of-speech (POS) tags. We modified the rule-based methods to use Stanza's universal dependencies and achieved a precision score of 97.59 while extracting a wider variety of method mentions.

_Research Gap:_ Few researchers have addressed the question of automatic method mention extraction from biomedical text. Previous work has failed to address a convenient and scalable approach to this problem. Techniques to generate a human-labelled corpus are time-consuming and unfeasible on a larger scale. This empirical study aims to analyze different approaches and attempts to improve on these issues.

## 3) Research Objectives

- O1: One of the objectives is to improve on the performance benchmarks produced in [8]. Using a human-made gold standard corpus and a machine-made silver standard corpus along with better linguistic filters and better feature selection, this study aims to improve the precision, recall, and F-score benchmarks by at least a few percentage points than the performance benchmarks calculated in [8].
- O2: Another objective is to successfully create an accurate silver standard corpus using Stanford's CoreNLP (Stanza) toolkit which can then be used in combination with different machine learning (ML) algorithms. We will be using both a gold standard corpus (human-made) and a silver standard corpus (machine-made). A gold standard corpus will have less data but the data is more accurate. Consequently, the silver standard corpus will have more data, but it can be theoretically less accurate than the gold standard corpus. Both of these corpora can be used alongside different machine learning algorithms to achieve our goal.
- _*Significance*_

  The purpose of the aforementioned objectives (O1, O2) is to contribute to the developing knowledge in the field of Natural Language Processing and provide better understanding and better approaches toward the problem of Information Extraction. As more and more information is being produced globally, conventional approaches may not be adequate and feasible to accurately and efficiently extract method mentions from unstructured text.

For future research purposes, successfully creating a silver standard corpus in the biomedical field through automated means while retaining sufficient recall, precision, and F-score will open up the possibility of constructing machine-made datasets that can effectively be used as training data and testing data for machine learning models. In addition to that, automatic method mention extraction in scientific articles can be extremely useful in generating an index for information lookup and provides users with the ability to know the contents of a scientific paper by only reading the extracted information.

## 4) Methodology

This research aims to improve on the results produced in [8] and analyze the different approaches and methodologies used towards extracting method mentions from biomedical text. Our methodology was partly based on the methodology defined in [8].

In this study, for tagging our results, 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 [11].

In the initial stage of the study, we prepare a collection of sentences that contain mentions of method names, or “method sentences,” as Houngbo [8] refers to them. 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 extract method mentions. Leveraging rule-based methods allows for results without the use of any pre-existing training/testing 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 various machine learning algorithms related to Natural Language Processing (NLP) tasks, such as Conditional Random Field (CRF) models and the Bidirectional Long Short Term Memory (Bi-LSTM) models. We opted for choosing these algorithms as CRF models are discriminatively trained for sequence segmentation and labelling [7] and Bi-LSTM models have proven highly successful at language tasks, as they can consider both the left and right contexts of a word [10].

## 5) Results

### 5.1) Corpus Creation

In this study, we scanned through a collection of 2839 research papers. Similar to [8], this study also employs using the anaphoric relations between sentences to successfully find the "method sentences". According to the findings reported in [12], nearly all antecedents can be found within two sentences from the demonstrative anaphors. So, to generate our corpus, we searched through our research papers for sentences that contain the anaphor "this method" and then selected the three sentences that precede the "this method" sentence 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 mention. As a result, we retrieved 10974 potential "method sentences".

> **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.*

*Example 1: Showing how the corpus creation process works*

For instance, in Example 1, Sentence 4 contains the "this method" anaphor, therefore the corpus would contain Sentence 1, 2, and 3. And it is to be noted that Sentence 3 contains the antecedent "*cutaneous basophilic hypersensitivity test*" which is a method mention.

### 5.2) Rule-based Approach

After generating the corpus which contained 10974 potential "method sentences", we used linguistic rules and patterns to programmatically extract the method mentions depending on the dependency relationships between the words. To accurately evaluate the dependency relationships between the words, we used Stanza (The Stanford NLP Group's official NLP Python library) [13], the biomedical and clinical model packages included in the Stanza toolkit [14], and Genia Tagger, a part-of-speech tagger that is specifically suitable for processing biomedical text [15].

Most of the method mentions in our corpus can be represented by the following examples:

1. Tukey's biweight method
2. naive KNN method
3. 10-fold cross-validation test
4. Roche Amplicor Cystic Fibrosis test
5. bimolecular fluorescence complementation analysis
6. Felsenstein's independent comparison method
7. statistical total correlation spectroscopy analysis method
8. MANOVA-based scoring method
9. protein sequence Jukes-Cantor model
10. cutaneous basophilic hypersensitivity test

From these examples, we can observe how the rules and patterns to extract our method mentions would look like. For instance, it can be noted that all of these mentions end with key suffixes that would correspond to most method mentions, key suffixes such as *method*, *analysis*, *test*, *model*, *etc.* And in addition to that, it can be noted that all of these method mentions have at least one *compound* universal dependency (UD) relation, some of them have at least one *nmod:poss* universal dependency relation, and most of them have at least one *amod* universal dependency relation. A *compound* UD is a modifier that relates to a noun in the sentence 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.

By looking at these observations, we created five rules based solely on the universal dependency relationships between the words. Unfortunately, two of these rules did not meet the requirements for our study as one of them was not performing well enough, and the other one was modifying the extracted method mentions such that they could not be tagged by the IOB tagging format. However, the remaining three rules were performing successfully for our rule-based model. These three rules were able to extract 1338 method mentions from 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 as a method mention,
- 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 mention,
- 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 mention,

The rules stated above are different from the rule used in [8], which simply used POS tags to extract the method mentions; whereas the rules in this study use universal dependencies using

Stanza, which is pre-trained on biomedical vocabulary.

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.

Table 1 below shows the results for the rule-based model.

| System | Precision | Recall | F-Score |
|---|---|---|---|
| Rule-based | 97.59 | N/A | N/A |

*Table 1: Precision, Recall, F-Score for Rule-Based Method*

By utilizing this rule-based model, we will tag the extracted labels using IOB tagging in order to create our silver standard dataset which can be used in Section 5.3.

### 5.3) Machine-learning Approach

When the rule-based approach has been completed, we are now ready to investigate 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 (Bi-LSTM) 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 5.2.

CRF models are a framework for developing probabilistic models for segmenting and labelling sequence data [7]. CRF models outperform other models used for tagging, such as Hidden Markov Models (HMM) and Maximum-entropy Markov Models (MEMM) since CRF models overcome the label bias problem and are more flexible in terms of feature selection [7].

Bi-LSTM models are a form of recurrent neural networks (RNN) that can understand the context of a sentence quite well. RNNs are a type of network which form memory through recurrent connections [16]. Bi-LSTM models work well for NLP tasks as they can contextually scan through text in both forward and backward directions. Neural nets require words to be encoded as vectors of floating point numbers, called word embeddings. For the word embeddings, the Bi-LSTM 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 [17].

Table 2 below shows the results for each of the machine learning models.

| System | Precision | Recall | F-Score |
|---|---|---|---|
| Conditional Random Field | 83.58 | 85.49 | 84.53 |
| Bidirectional LSTM | 68.42 | 39.39 | 50.00 |
| *Houngbo & Mercer's* CRF [8] | 81.80 | 75.00 | 78.26 |

*Table 2: Precision, Recall, F-Score for Machine Learning Methods*

We observe from Table 2 that the highest performing machine learning model in this study (CRF) outperforms the machine learning model results of [8] 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 2. As an example, the CRF model produced 20 predictions that were labelled as *not* method mentions in the testing data, however, a manual check shows that 17 out of those 20 predictions actually are method mentions but were labelled wrong in the testing data. This result provides insight into the fact that the CRF model managed to predict the method mentions that we as researchers had not defined in Section 5.2.

Novelty of Results

As far as we are aware, there is no work carried out on utilizing the Stanza dependency parser and the universal dependencies relationship between words to create rules and patterns tailored towards automatically extracting method mentions from biomedical text. In addition to that, as far as we know, using a Bi-LSTM model to find method mentions has not been attempted before either.

**6) Discussion**

- Threats to the validity of the results
    - Our definition of what a "method mention" is has been quite vague in this study as there is no correct answer to that. One individual's view of what a method mention is might be different from another individual's. This vagueness of definition has prevented us from retrieving accurate performance metrics.
    - To contain this threat to some extent, we generated multiple rules and patterns based on our definition of method mention, and we were quite lenient with that definition in order to extract data with more variance.
- Implications of the research results
    - In terms of the practical application in industry, our findings can result in the construction of a querying/search algorithm that can perform information lookup faster than conventional search algorithms by extracting relevant information from unstructured biomedical text and constructing an index of method mentions.
    - In addition to that, the results of this research paper can be used to create a recommendation system tailored to the biomedical field. For instance, medical professionals can find scientific papers and articles relating to what they are searching for. For example, if an individual was interested in the effect of

    - dopamine on the heart, and they find a research paper about this topic, the recommendation system will recommend other research papers that the individual will find relevant as well. Such a recommendation system will greatly speed up the research process for people in the biomedical field.
- Limitations of the results
    - We are aware that the results of our research may be limited due to the inaccuracy of our performance metrics. In Section 5.2, due to the large amount of data, the recall score for the rule-based mode could not be calculated. And in Section 5.3, due to the testing data not having the correct labels for the method mentions that were outside of the scope of the rules defined in Section 5.2, our performance metrics are lower than what they actually should be.
- Generalisability of the results
    - The results in this empirical study should be able to generalise in the scope of biomedical text. As the tools and resources used in this research are primarily pre-trained on biomedical vocabulary, it should be able to generalise within the field of biomedicine.

## 7) Conclusions

In this paper, we explored various methodologies to automatically extract method mentions from biomedical text. In the initial step, we created a corpus containing method sentences using anaphoric relations. Afterwards, we investigated rule-based methods as well as machine learning methods to automatically extract method mentions 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 mentions within the scope of the rules as shown in Section 5.2 as well as in Table 1. In addition to that, the results from Section 5.3 and Table 2 show how the CRF model outperforms the results from [8] and show the potential of machine learning models to accurately generalize outside the scope of the rules defined in Section 5.2.

## 8) Future Work and Lessons Learnt

- Our future work would include:
    - Improving the definition of a method mention and making it more inclusive of a wider variety of terminologies and methodologies, and in turn, create a wider variety of rules and patterns for our rule-based approach.
    - Implementing a BiLSTM-CRF model and a BiLSTM-CNN-CRF model to improve our performance in our machine learning approach section. 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.
- Significant lessons that could be useful to others
    - Establishing a clear goal as well as small milestones at the start of the study provides a useful roadmap to follow throughout the research.

**9) Acknowledgements**
I would like to acknowledge and thank my supervisor, Dr Robert Mercer, for his advice and guidance throughout the period of this empirical study, as well as, Xindi Wang for providing me with the BioWordVec_PubMed_MIMICIII_d200 vector file in text format.

## 10) References

[1] K.-J. Lee, Y.-S. Hwang, S. Kim, and H.-C. Rim, “Biomedical named entity recognition using two-phase model based on SVMs,” Journal of Biomedical Informatics, vol. 37, no. 6, pp. 436–447, 2004.

[2] H.-J. Song, B.-C. Jo, C.-Y. Park, J.-D. Kim, and Y.-S. Kim, “Comparison of named entity recognition methodologies in biomedical documents,” BioMedical Engineering OnLine, vol. 17, no. S2, 2018.

[3] J. P. Chiu and E. Nichols, “Named Entity Recognition with Bidirectional LSTM-CNNs,” Transactions of the Association for Computational Linguistics, vol. 4, pp. 357–370, 2016.

[4] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” Benjamins Current Topics Named Entities, pp. 3–28, 2009.

[5] B. Settles, “Biomedical named entity recognition using conditional random fields and rich feature sets,” Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications - JNLPBA '04, 2004.

[6] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, “Deep learning with word embeddings improves biomedical named entity recognition,” Bioinformatics, vol. 33, no. 14, pp. i37–i48, 2017.

[7] J. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” *ScholarlyCommons*. [Online]. Available: https://repository.upenn.edu/cis_papers/159/.

[8] H. Houngbo and R. Mercer, “Method mention extraction from scientific research papers,” Proceedings of COLING 2012: Technical Papers, 2012.

[9] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” IEEE Transactions on Neural Networks, vol. 13, no. 2, pp. 415–425, 2002.

[10] S. Gooding and E. Kochmar, “Complex Word Identification as a Sequence Labelling Task,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[11] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural Architectures for Named Entity Recognition,” Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[12] M. Torii and K. Vijay-Shankar, “Anaphora Resolution of Demonstrative Noun Phrases in Medline Abstracts,” *(PDF) Anaphora Resolution of Demonstrative Noun Phrases in Medline Abstracts*. [Online]. Available: https://www.researchgate.net/publication/228748316_Anaphora_Resolution_of_Demonstrative_Noun_Phrases_in_Medline_Abstracts.

[13] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020.

[14] Y. Zhang, Y. Zhang, P. Qi, C. D. Manning, and C. P. Langlotz, “Biomedical and Clinical English Model Packages in the Stanza Python NLP Library,” arXiv.org, 29-Jul-2020. [Online]. Available: https://arxiv.org/abs/2007.14640. [Accessed: 07-Apr-2021].

[15] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, “Developing a Robust Part-of-Speech Tagger for Biomedical Text,” *Advances in Informatics*, pp. 382–392, 2005.

[16] A. Aziz Sharfuddin, M. Nafis Tihami, and M. Saiful Islam, “A Deep Recurrent Neural Network with BiLSTM model for Sentiment Classification,” *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018.

[17] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, “BioWordVec, improving biomedical word embeddings with subword information and MeSH,” Scientific Data, vol. 6, no. 1, 2019.