Method Mention Extraction from Biomedical Text

Thesis Proposal

Waqar Kalim

COMPSCI 4490Z - Thesis

Department of Computer Science

Western University

12th November 2020

Project Supervisor: Dr Robert Mercer Computer Science Department

Course Instructor name: Prof. Nazim Madhavji

1. **Abstract**

* **Content and motivation**: Extraction of method mentions from biomedical text has been an important task in the field of Natural Language Processing (NLP).
* **Question/problem/Objectives:** Scientific research articles commonly consist of complex keywords and terminologies that are specific to the domain. We will implement different approaches and observe the results to determine an efficient solution for our goals.
* **Principal ideas:** In this research, we will extract method terminologies from scientific text using both rule-based and machine learning techniques.
* **Research approach (or Methodology):** We will use regular expressions for the rule-based methods, and linguistic parsing tools to construct the biomedical corpora to be used with various machine learning algorithms.
* **Anticipated contributions (or results):** This thesis will provide a set of methodologies to extract method mentions from scientific corpora. The findings of this thesis paper can help build resources that can be used in various Natural Language Processing applications in the biomedical field.
* **Anticipated impact of results:** This thesis will be improving on the results of (Houngbo & Mercer, 2012). By automating the extraction of method mentions, we can better index information and potentially improve information lookup.
* **Limitations:** We are aware that our research may have one limitation so far. The linguistic parsing tools that are being used for the corpora construction are prone to rare cases of error.

1. **Background**

Method Mention Extraction from unstructured text has been an important field in Natural Language Processing. Especially in the field of biomedicine, the automatic extraction of methodology names and terminologies has been imperative. Past studies such as (Settles, 2004), (Maynard & Ananiadou, 2000) have focused on automatic terminology extraction. And many other works such as (Yu, Bohnet, & Poesio, 2019) and (Houngbo & Mercer, 2012) have primarily focused on automatic extraction of method mentions. The latter research paper has been a huge influence on this study.

Similar to (Houngbo & Mercer, 2012), the goal of this study is to extract method terminologies from biomedical research papers and create a lexical resource derived from relevant information mentioned in the research articles. However, our approach in this study will be slightly different from the aforementioned paper, in terms of both rule-based techniques and our use of neural learning techniques. Overall, our objectives will be based on improving the results obtained in (Houngbo & Mercer, 2012).

1. **Detailed Proposal:**
   1. **Research Objectives:**
      1. **GOAL:** As stated in the Abstract, our main goal was to improve on the results of (Houngbo & Mercer, 2012) by using a variety of linguistic tools, such as dependency and constituency parsing, and various machine learning algorithms in hopes of creating more sophisticated methods that can be useful in the medical field.
         1. **OBJECTIVE 1:** One of the objectives is to successfully create a silver standard corpus using Stanford's CoreNLP toolkit which can then be used in combination with different machine learning 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 the data 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.
         2. **OBJECTIVE 2:** Another objective is to improve on the performance benchmarks produced in (Houngbo & Mercer, 2012). 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 the (Houngbo & Mercer, 2012) research paper.
      2. **SIGNIFICANCE:** The purpose of the goal mentioned in Section 3.1.1 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 in the world, conventional approaches may not be adequate and feasible anymore 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 test 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.
   2. **Research Approach and Methodology:**

* **Technical issues:** At this stage of the research, it is not possible to know all of the technical issues that we will face down the line, but there are still some potential issues and problems that we can expect. Firstly, the third-party linguistic tools we are using for dependency and constituency parsing (primarily Stanford's CoreNLP) does have potential issues. For instance, for some ambiguous sentences, the toolkit might not return the correct output. This issue potentially hinders to some extent the quality of the silver standard corpus that we are generating. We are unable to fix this problem and will take into account this technical issue in our study.
* **Technologies, components:** In our study, we are going to be employing the use of Stanford's CoreNLP toolkit, which is an open-source project, with the aim of generating the silver standard corpus for our research. Along with the linguistic toolkit, we will also be employing the use of various machine learning algorithms specifically related to Natural Language Processing tasks such as Conditional Random Fields (CRFs), Support Vector Machines (SVMs), Hidden Markov Models (HMMs), among other algorithms.
* **Identify key issues in the research methodology:** This study will employ both a rule-based approach and a machine learning-based approach. With the rule-based approach, regular expressions will be used to create specific rules and patterns to extract method mentions. Use of rule-based methods allows for results without the use of any pre-existing training/testing data; however, the main disadvantage is the need to constantly add new rules and new patterns depending on the linguistic patterns of the terminologies; this can promptly become a time-consuming task. With the machine learning approach, both a human-made gold standard corpus and 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 gold standard corpus will be constructed manually, whereas, the silver standard corpus will be constructed automatically using Stanford's CoreNLP toolkit.
* **Technology/Solution (results):** The conclusion of this study will be a comparison of the different approaches that can be utilized for method mention extraction alongside accurate benchmarks to determine which approach performs better for which situation.
* **Challenges anticipated:** This research study faces some anticipated challenges, especially since our research is aimed toward the field of biomedicine. As mentioned in (Song, Jo, Park, Kim & Kim, 2012), biomedical named entity extraction faces difficulties for various reasons. The first one being 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.
* **Mitigation of challenges:** Although the anticipated challenges will constrain the results of the study, these challenges can be mitigated. Creating new rules and patterns for the rule-based approach can be a tedious and time-consuming task, so to mitigate the cost of this approach, a machine-learning approach will also be implemented. And in addition to that, employing an implementation based on machine-learning will allow us to accomplish contextualized information extraction which will mitigate the issue of the same words having different words in terms of the context.
* **Validation strategy:** The results from both the rule-based approach and the machine learning-based approach would need to be validated. Our validation strategy will consist of two techniques: 1) splitting our corpora into train/test data and evaluating our trained models (trained by the training data) on the testing data, and 2) performing cross-validation on our trained models (trained by the entire corpora).

1. **Value of the results and industrial relevance:**

* In the field of research, our study will improve on the results found in (Houngbo & Mercer, 2012) and outperform the methods used for both the rule-based and the machine learning approaches.
* In terms of the practical application in the 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 the 'tags'. For instance, with a conventional search system, the algorithm would need to search through every word of every scientific text to find the search query, but with our research, the algorithm would only need to locate the search query in our index, which speeds up the search aspect of the algorithm significantly.
* 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.

1. **References**

Houngbo, H., & Mercer, R. E. (2012, December). Method mention extraction from scientific research papers. In *Proceedings of COLING 2012* (pp. 1211-1222).

Settles, B. (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 (NLPBA/BioNLP)* (pp. 107-110).

Maynard, D., & Ananiadou, S. (2001). TRUCKS: A model for automatic multi-word term recognition. *Journal of Natural Language Processing*, *8*(1), 101-125.

Song, H. J., Jo, B. C., Park, C. Y., Kim, J. D., & Kim, Y. S. (2018). Comparison of named entity recognition methodologies in biomedical documents. *Biomedical engineering online*, *17*(2), 158.

Yu, J., Bohnet, B., & Poesio, M. (2019). Neural Mention Detection. *arXiv preprint arXiv:1907.12524*.