Method Mention Extraction from Biomedical Text

Progress Report 1

Waqar Kalim

COMPSCI 4490Z, Department of Computer Science

Western University

23rd December 2020

Project Supervisor: Dr Robert Mercer, Computer Science Department

Course Instructor name: Prof. Nazim Madhavji

* **Structured 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 report, we will extract method terminologies from scientific text-primarily using a rule-based approach to construct our silver standard corpus.
  + **Research approach (or Methodology):** We will use regular expressions, linguistic parsing tools, and Natural Language Processing toolkits for the rule-based methods in order to construct the biomedical corpora.
  + **Anticipated contributions (or results):** This report 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. **Summary of work done to date**

(a) *Read and understood Hospice Houngbo's code and datasets*: We contacted Hospice Houngbo and inquired about the code and the datasets he used for his (Houngbo & Mercer, 2012) research paper. After retrieving these resources, we went through all of the code and the datasets to understand exactly how the code was working as well as how the datasets were being used.

(b) *Extraction of method mentioning lines from research papers*: From a set of around 5000 research papers that are related to the field of biomedicine, sentences that potentially mention method names were extracted using specialized regular expressions and Linux commands tailored towards searching text.

(c) *Rule-Based Extraction on sentences*: On these extracted sentences that potentially contain method mentions, we performed narrower rules for rule-based extraction using Python's NLTK library for PoS Tagging alongside with narrower regular expressions to accurately determine whether the sentence contained a method mention and whether it did not.

(d) *Creation of Silver Standard Dataset*: We are currently in the process of creating a silver standard dataset. From the set of research papers, we will update our regular expressions depending on which sentences from (b) did not appear to contain any method mentions and try to extract different sentences as we did in (a). By repeating this process multiple times, a silver standard dataset containing our method mention data will be generated.

1. **Technologies, systems, programming languages, etc., used to date:**

|  |  |
| --- | --- |
| **Name of technology, system, etc.** | **Brief description of its use in the project.** |
| Regular Expressions | Used for defining a search pattern in a sequence of characters |
| Genia Tagger | A PoS (Part of Speech) Tagger trained on a biomedical corpus |
| Python’s NLTK Library | A Natural Language toolkit used for PoS tagging like the Genia Tagger |

1. **Literature reviewed to date**

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*.

1. **Any challenges or problem areas, if any.**

There have not been any serious issues in the project so far that could significantly slow down or thwart progress.