Integrating Tools for Automated Annotation of Biomedical Texts

JSPS project

at
DBCLS

under the supervision of
Dr. Jin-Dong Kim

presented by **Nicola Colic**
Zurich, Switzerland
submitted on the 4^{th} of October, 2016

Abstract

In this report, the grant beneficiary describes his activities during a JSPS-funded project, which lasted 6 months and took place at DBCLS in Tokyo. The beneficiary was engaged in three activities mainly: Firstly, he developed several web services that can be used by the text mining community to obtain annotations from different parsers. Secondly, he was contributed to the LODQA program, a question answering service for medical practitioners. Thirdly, the grant beneficiary participated and was involved in the organisation of the BioHackathon 2016 in Tsuruoka.

Furthermore, the beneficiary's supervisors in Zurich and Tokyo and the beneficiary collaborated intensely on a proposal for JSPS to fund two three-year doctoral positions for a joint project. Through these activities, the grant beneficiary was allowed to advance as a researcher, and strengthened ties between the home and host institution.

Contents

In this report, I, the grant beneficiary, explain the context and contents of my work of the JSPS-funded project.

1 Research Background

In this section the general research field and its importance are described, as well as the more specific research of the my supervisors. The work this report summarises is then situated within the context of their research as well as my own previous research.

1.1 General Research Field

The work presented in this report is situated in the domain of biomedical text mining, which is concerned with aiding the extraction of relevant information from the ever-growing wealth of biomedical publications. In recent year especially, the rate of such biomedical publications has increased tremendously, making it increasingly difficult for experts to keep track of advancements in their own field of research. Currently, curators maintain databases that summarise the most relevant findings of publications in order to alleviate the problem. However, manual curation is costly and slow, creating a delay between the time of publication and the inclusion of the relevant information into the respective database. Because of this, researchers turn to text mining to aid curation of these databases.

Currently, the biggest corpus of biomedical publications is the PubMed database, which contains over 26 million article abstracts. Researchers around the world try to provide various annotations for this database, such as automatically identifying named entities such as diseases or genes or extracting relations between proteins.

1.2 Supervisors Research

Both the my supervisors at his home institution and host institution, respectively, have vast experience in this field, and are both involved with initiatives to strengthen the communication and collaboration of the biomedical text mining community.

My supervisor in Zurich, Dr. Fabio Rinaldi, is leader of the OntoGene[1]

[1] http://www.ontogene.org/

project, which organises challenges for the biomedical text mining community and maintains a pipeline aimed at automatically extracting named entities such as drug or protein names from unstructured text as it is found in the PubMed database.

The supervisor in Tokyo, Dr. Jin-Dong Kim, is also involved with several similar initiatives such as BLAH[2] and BioHackathon[3]. Such initiatives aim at improving communication and collaboration in the field of biomedical text mining, and at reducing duplicated efforts. In fact, many groups around the world develop similar pipelines and programs which provide annotations to the same articles found in PubMed. Their annotations and formats, however, are largely incompatible. PubAnnotation[4] is a project developed by Dr. Kim which aims to alleviate this duplication of efforts by allowing researchers to upload their annotations, and have them automatically be mapped to the corresponding articles. Like this, different annotations obtained by different programs and in possibly different formats for the same publications can be compared.

Furthermore, Dr. Kim is also developing the LODQA[5] service. LODQA is a question answering service for the medical domain. It allows biomedical practitioners without technical training to phrase natural language queries such as *Which genes are related to Alzheimer's disease?*. This program relies on several technologies, such as the appropriate parsing of input questions in order to convert the natural language query into the query language, which is then sent to a knowledge database (such as the ones described in 1.1).

1.3 Own Previous Research

As part of his master's dissertation at the University of Zurich, I explored novel ways to extract relations from unstructured biomedical texts. In particular, a novel scheme to combine different tools for preprocessing the documents has been proposed, and a parser not previously subjected to scientific testing been evaluated. The annotations obtained by this approached could be used to improve on the existing OntoGene pipeline maintained at UZH. However, the format was proprietary, and not readily available for other research communities.

[2] http://3.linkedannotation.org/home

[3] http://www.biohackathon.org/

[4] http://pubannotation.org/

[5] http://lodqa.org/

1.4 JSPS Project

I, Nicola Colic, worked from March to September 2016 at the host institution, DBCLS in Tokyo, under the supervision of Dr. Kim. His supervisor at his home institution, the University of Zurich (UZH), was contacted frequently and informed about the progression of the project.

2 Research Methodology

I worked at the office at DBCLS as a junior researcher, working on various tasks and projects assigned by Dr. Kim, as well as participating in the general office life. In developing programs which utilise the PubAnnotation platform, various problems and bugs could be discovered, which were reported in frequent discussions with the supervisor, and subsequently corrected.

Furthermore, I participated in BioHackathon in Tsuruoka, Japan, a one-week programming event for the biomedical community organised by Dr. Kim and his colleagues at DBCLS.

Apart from the work in the office, I was enrolled in Japanese language classes and cultural experiences, allowing him to form deeper ties with his co-workers.

The following section lists the tangible outcomes of this project.

3 Research Implementation and Results

Over the project period, three web services were implemented by me, as well as improvements made to a question answering service developed by Dr. Kim. Furthermore, I participated in an intense programming session organised by DBCLS. Finally, together with my supervisors, we submitted a proposal for another grant that would allow the home and the host institution to collaborate on a joint project.

3.1 Web Services

The first project was to make the previous research accessible to non-expert users and for PubAnnotation. More precisely, as part of my master's dissertation, a novel combination of pre-processing tools was evaluated and found to be superior in some applications than currently commonly used tools. A program was developed that utilises the approach described in the master's dissertation, and runs as a REST-ful server (henceforth called *spaCy*, owing to the new parser used), allowing users to submit their own texts to the server to obtain an annotated version of their text without having to worry about the interaction of the different tools. This program was installed on DBCLS servers and also maintains a web-interface which can be accessed at `http://spacy.dbcls.jp/spacy_rest`.

While this web service works independently, it was developed particularly in order to be used by the PubAnnotation platform. Conceptionally, the PubAnnotation allows using other annotation service through a REST-ful API. PubAnnotation will send the article to be annotated to the respective REST-ful service, obtains the annotation from the service and stores it in its own system.

This feature, however, has not been thoroughly tested previously. Through the development of the web service, numerous errors and bugs in PubAnnotation could be discovered, making PubAnnotation easier to use for future developers of similar web services. The interaction of PubAnnotation and the web service has been thoroughly tested using large quantities of sample text[6]. The annotations provided by the web service and obtained through PubAnnotation were then used in the BioNLP 2016 challenges[7].

Through this work, I was able to familiarise himself with several new technologies, such as web server development, and the infrastructure and projects at DBCLS, especially PubAnnotation.

Building on the success of this project, and in reaction to recent advancements in the field, two more similar implementations were developed. One one hand, the supervisor became aware of a recent version of the *Turku Event Extraction System* (TEES)[8], for which a similar web service was developed[9]. Again, annotations obtained using this service were used in the BioNLP 2016 challenges[10].

While similar in its implementation as a web service, the TEES service provides a different kind of annotations than the spaCy service. This required again adaptations to the PubAnnotation infrastructure. These developments are described more thoroughly in [2], for which I is listed as co-author.

[6] `http://pubannotation.org/projects/spacy-test`

[7] `http://2016.bionlp-st.org/`

[8] `http://jbjorne.github.io/TEES/`

[9] `https://github.com/Aequivinius/tees`

[10] `http://pubannotation.org/projects/bionlp-st-ge-2016-test-tees`

In May 2016, Google announced the release of their own parsing tool called *Parsey McParseface*[11], creating a big stir in the natural language processing and text mining communities. Given previous success at turning parsers into web services, the supervisor and I decided to try to implement a further web service that would make this much anticipated parser directly useable for the larger community as a web service. While ultimately a functioning implementation could be presented, considerable difficulties were met, owing to the arcane coding style of the original parser and its byzantine architecture. Still, the final program created much interest within DBCLS.

3.2 BioHackathon

BioHackathon is a one-week intense programming session organised by DBCLS. It was held in Tsuruoka in June 2016, and was attended by over 100 participants from both Japan and overseas. During this time, I commenced his work on the LODQA project described below, and took the opportunity to meet many of the leading figures in the field. Furthermore, I also helped the organisers with the planning and execution of some of the evening activities, which were greatly appreciated by the participants.

3.3 LODQA

The remaining time of the project was used to work on Dr. Kim's LODQA service. So far, the LODQA system relied on a single parsing tool to parse the queries provided by the user. The goal was to open this restriction and allow LODQA to utilise different parsers in order to evaluate whether performance could be improved. However, this necessitated a major refactoring of the existing LODQA architecture as it had not been developed with this in mind.

Following this refactoring, LODQA was changed to also utilise the aforementioned spaCy and Parsey services to obtain parses, and allow the user to select which parser to use. The different parses where then evaluated manually against the TREC genomic track question set[12]. In the process of evaluation of the parsers and their differences, several minor bugs in the existing LODQA system could be identified and fixed, especially as

[11] https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

[12] http://trec.nist.gov/tracks.html

they pertained the interpretation of the parses, making the system overall slightly more robust. The code of these developments can be found at `https://github.com/Aequivinius/lodqa`.

During this time [1] was also published, which I co-authored. This paper focuses predominately on research preceding the project described in this report. However, some of the software used in the paper was improved as part of the project.

4 International Exchange Achieved through the Research

Exchange was achieved in several ways.

Firstly, I am is now well-versed in the infrastructure and tools used at DBCLS, especially PubAnnotation. With its goal to integrate research communities and their efforts, PubAnnotation benefits from being used by different research groups. Upon my return I intend to use his expertise to make research efforts in Zurich compatible with PubAnnotation, and thus add to the usefulness of PubAnnotation.

Secondly, previous research efforts at UZH were made accessible and useful not only for DBCLS, but also to the biomedical text mining community at large. DBCLS has a tradition of developing tools for the community, and the aforementioned web services developed by myself stand very much in line with this tradition.

Thirdly, and most importantly, is the considerable effort that was put into writing a new proposal for a JSPS grant for two three-year doctorate positions at both DBCLS and UZH dedicated to a joint project. Both my supervisors and I collaborated intensely to write a proposal that would allow UZH and DBCLS to work closely together on a joint project over three years, continuing to join efforts and engage the wider biomedical text mining community.

Finally, I was immersed deeply in Japanese culture and is continuing his studies of Japanese language. Owing to this immersion, I am more sensitised to cultural differences and similarities between Switzerland and Japan.

5 List of Published Papers

[1] M. Basaldella, L. Furrer, N. Colic, T. R. Ellendorff, C. Tasso, and F. Rinaldi. Using a hybrid approach for entity recognition in the biomedical domain. *Proceedings of the 7th International Symposium on Semantic Mining in Biomedicine*, 2016.

[2] J.-D. Kim, Y. Wang, N. Colic, S. H. Baek, Y. H. Kim, and M. Song. Refactoring the genia event extraction shared task toward a general framework for ie-driven kb development. *ACL 2016*, page 23, 2016.

6 Comments to JSPS

I would like to take this opportunity to express my deepest gratitude for this opportunity to engage with the Japanese research community in my field of research, but also experience the Japanese culture. Through this generosity, I have been able to propel my career as a researcher, and to grow personally through my deepened understanding of Japanese culture. I am convinced that the experiences and friendships made in Japan will accompany me in the future and continue to inspire me.