UNIVERSITY OF BIRMINGHAM

Final Year Project

Extracting Key Phrases and Relations from Scientific Publications

Dissertation for B.Sc in Computer Science

School of Computer Science, University of Birmingham

Author
Thomas Clarke (1443652)

Supervisor
Dr Mark Lee

April 2018

Declaration

The material contained within this thesis has not previously been submitted for a degree at the University of Birmingham or any other university. The research reported within this thesis has been conducted by the author unless indicated otherwise.

Acknowledgements

I would like to give acknowledgement to those who helped me throughout the completion of this project.

Firstly, a thank you to Dr Mark Lee for being a supportive and informative supervisor, as well as an entertaining host during project meetings.

I also wish to thank my friends and family in supporting me during the year leading preceding this dissertation, ensuring I kept on track and in a good frame of mind.

Abstract

This project presents solutions developed to solve the SemEval 2017 ScienceIE task - analysis of scientific publications to extract key information. This includes three sub tasks: *(A)* key phrase extraction, *(B)* classification and *(C)* relation extraction.

To achieve subtask A, the text of a paper is parsed to find it's semantic tree. Then, each word in succession is tested in a Support Vector Machine (SVM), based around a words' semantic attributes to determine if it should be a, or part of a, key phrase. Each phrase generated is also sanitised to reduce excess information. Subtask B involved treating each key phrase as a Bag-Of-Words, and calculating the phrases' distance to each classification type using Word2Vec. Finally, subtask C experimented with using the Word2Vec representation of a phrase and the relative distances between phrases combined with an SVM to try too detect relations.

Scores of NLP go here

To explore how this system could be used, a website was created hosting the information. This used Spring Boot to create a Java based web project which supported not only an archive of processed papers, but also the means to search using query strings and automatic processing of submitted papers to the system (through using the most successful versions of systems described above). *Evaluation summary goes here...*

Keywords

Natural Language Processing, Key Phrase Extraction, Classification, Relation Extraction, Support Vector Machine, Word2Vec, Spring Boot

Contents

List of Figures

List of Tables

1 Introduction

When conducting scientific study, being able to search existing literature around a subject can be vitally important. A search system which can automatically sort scientific papers into order, returning the one likely to be most useful first, can speed up the process of gathering this information. A system which can go further and extract important pieces of information from the paper to help present answers to user queries has the potential to be even more effective.

At SemEval 2017[1], a task which heavily applied to the above was presented: ScienceIE[2]. This natural language processing (NLP) based task was to analyse scientific papers to extract key pieces of information, classify those pieces and attempt to draw relations between them. In short, this is a information extration problem, spefically for scientific papers. The idea behind it is to support faster research as systems will be presented with information to help better gather relevant research when querying databses of existing literature.

1.1 Aims and Objectives

This project shall initially target the main goals of ScienceIE, and one of the methods of evaluation shall be through processing of the sample data and execution of the marking tools supplied as part of the task. Explicity, the overall task is split into 3 subtasks:

- **A**: The identification of all the key phrases in a scientiific publication
- **B**: The classification of each key phrase into one of the following categories:
  - **Process** (scientiific models, algorithms, processes)
  - **Task** (an application, end goal, problem, task)
  - **Material** (resources, materials)
- **C**: The identification of relationships between identified key phrases, where the relation is either none, or one of the following:
  - **Hyponym-of** (where the semantic field of key phrase A is included in that of key phrase B's semantic field, but not vice versa)
  - **Synonym-of** (where the semantic field of key phrase A and B are the same)

Therefore, through research of the systems created during ScienceIE and other research in the field, the largest and most obvious goal of this project is to create a system where any scientiific paper can be input, some processing happens (with no time constraints) and the desired key phrase information is produced as an output, in the expected format specified for ScienceIE. This is the 'brat' annotations format, which houses all of information described above about a paper in a single text document, saved seperately from the original paper. An example annotations file can be seen in APPENDIX.

The above can be refered to as the *NLP system* part of the project. To evaluate the NLP system, not only will the marking tools be used, but further analysis of the information extracted shall also be conducted; for instance exploring the differences in key phrases automatically extracted compared to the expected results (seeing cases where a shorter or longer key phrase was extracted and what difference this might make when using the generated data).

There currently existing many search engines that specifically deal with research papers; Google Scholar[3] and ScienceDirect[4] are well known, popular choices currently. As an extension to the ScienceIE task, motivated by existing search engines publically available on the web, the secondary goal of this project is to create a *product* based on the NLP system. It should use information extracted by the NLP system to present useful information to the user, given suitable input through a graphical user interface (GUI).

[1] http://alt.qcri.org/semeval2017/
[2] https://scienceie.github.io/
[3] https://scholar.google.co.uk/
[4] https://www.sciencedirect.com/

It should maintain a collection of scientific papers that are prepared for user query to effectively help them navigate to the most useful piece of information relating to their query first. As a minimum requirement, it should host at least the test data supplied by ScienceIE. The papers should be able to be read in full, or simply have the extracted information presented (at least in the brat format described above) for the users convinience.

The goal of this *GUI* section of the project is to be able to explore the potential effectiveness of the extracted information in relation to a researcher trying to find relevant research and how effectively it can be presented to aid in understanding it. Evaluating this shall be from allowing academic peers to use the system to find out their experiences in its effectiveness when navigating the presented data, understanding what the data means as presented in from of them, and comparisons they can draw between this system and other products, such as Google Scholar.

1.2 Report Outline

This document begins with a background to the field, which feeds into spefification of what is to be explored annd implemented. This will cover the NLP system in detail and outline the GUI requirements, as this shall be discussed more towards the latter parts of the paper. Following that, the NLP system implementation is reported on, concluded in the next section which evaluates the strengths and weaknesses of the NLP system. One the NLP system has been discussed, the GUI concepts shall be explained in full, the implementation discussed and evaluation completed. To sum up, a final discussion section shall review the project as a whole, and reiterate the strongest positives and note some of the points for improvement or expansion.

2 Background and Literature Review

Evaluating the outcome of ScienceIE at SemEval indicates potential paths for future systems and document very recent activity in the key phrase extraction area. Three papers were published from the event regarding this task.

One team's results [**?**] included a summary of all participating teams attempts, which detailed the highest F1 score measured at 0.43 for all three sub-systems combined, which can be seen as a potential target for this project (at least to get close to matching and potentially surpassing it). Furthermore, when deciding the algorithms to be used in the system created, only 31% of all keywords found had been seen in training, meaning any system created must be generic to extract key phrases from future unseen papers. For feature extraction, it was evaluated that while many high scores were achieved with recurrent neural networks (NN), the highest scoring system was a support vector machine (SVM) using a well-engineered lexical feature set. SVMs and NNs were also popular choices for key phrase classification. For relation extraction, many methods were attempted and while a convolutional NN was the most effective, various other methods (including SVM, multinomial naïve Bayes and more) all achieved very similar and reasonably accurate scores (up to F1 0.54).

The best end-to-end ScienceIE team used a long short-term memory (LSTM) approach for phrase extraction, with labelling completed by a conditional random field (CRF) based sequence tagging model [**?**]. Their sequence tagging model employed gazetteers built from scientific words extracted from the web. Another team [**?**] also had similar ideas, using CRFs to complete some of the task using WordNet[5] as a data source for the classifier they created. Both teams here also used sensible rules to help improve their score, such as intuitively marking all instances of a key phrase as a key phrase upon finding one instance (so if 'carbon' is extracted and labelled as a 'material', then all other instances are labelled to match) and exploiting hypernym relationship's bidirectional property (so if word 1 is a hypernym of word 2, the reverse is also true and therefore recorded).

Unfortunately, while extraction and classification were generally well handled, relation extraction has very low accuracy across all teams taking part with the average F1 score only being 0.15, with the highest score being 0.28. [**?**] achieved this using gazetteer built from Wikipedia[6] and freebase. This seems more appropriate than hand written rules, which may seem appealing as they can be somewhat tailored and provide high accuracy, but require much more effort from the developer and may not work well on unseen conditions providing low accuracy [**?**]. As mentioned earlier, WordNet is also a potential source of information for building a classifier for relation extraction, and a study by [**?**] compared building a classifier off of Wordnet and Wikipedia for hypernym-only extraction. The result of this shows that Wikipedia may be more suited to creating this type of classifier as it achieved an F1 score higher than using WordNet (the Wikipedia based classifier got 0.36 while the WordNet based classifier got 0.27). While an improvement, it is not ultimately a huge increase and there is no evidence either Wikipedia is better for the specific area of scientific papers (as the 2013 study was completed on a generic set of data).

The results of ScienceIE demonstrate there are several potential systems that could be implemented to answer this problem, with the best system potentially being a combination of algorithms and a voting system to select and label key phrases. The product would likely involve supervised learning and previous knowledge for some algorithms, along with unsupervised learning sections as well.

On the note of combining systems, several teams chose to back up key phrase extraction with a CRF. CRFs can be used for key phrase extraction alone as well, as documented by Zhang et al. [**?**], and while this study implies that CRFs (with an F1 of 0.51) are more accurate than an SVM (the most accurate at ScienceIE) this paper is slightly older than papers produced at SemEval 2017 and so even if the SVM information used then was the best that was available at the time (the SVM F1 score was 0.46), the SVM implemented at ScienceIE beat both of these scores considerably achieving an F1 of 0.56 [**?**].

A method not attempted at ScienceIE was unsupervised learning by clustering key phrases, a method which has potentially very accurate results that also could not only be robust again new unseen data but even different languages. The idea is that candidate key phrases are selected by some heuristic and other phrases are clustered about them. With the simplest approach, the center of a cluster is the key phrase. Various clustering methods were attempted by [**?**] on top of a candidate selection process built on semantic term relatedness. They ran tests on relatively short articles and while at maximum they only achieved an

[5] https://wordnet.princeton.edu/

[6] https://en.wikipedia.org

F1 of 0.45, there was several improvements suggested which apply to the task at hand concerning scientific papers. Firstly, an achievable improvement for this project would be to cluster directly on noun groups as they found most clusters consisted of groups of nouns anyway, which is backed up by Augenstein et al. [**?**], who reports 93% of all key phrases are noun phrases. Furthermore, improving their initial filtering to extend it further than stop words may help reduce errors as well; improving this may be possible by employing a words TF-IDF score with some threshold. Finally, they suggested a similar algorithm be applied to longer scientific papers. ScienceIE's test data consists of extracts of scientific texts (i.e. short paragraphs), however, any unsupervised system created for this task could be ran again entire papers and then only those sections compared for evaluation later – allowing this suggestion to be evaluated.

3 Analysis and Specification

Say what I'm going to do, but probably a bad idea to have a section for this. It may work better to just have a all of the 'what im doing' in each section when we get there.

4 The ScienceIE Task: Specification, Design and Implementation

A break down of each sub task follows...

4.1 Subtask A - Key Phrase Extraction

A section all about what I did for part 1

4.1.1 Method 1: Support Vector Machine

Go through making the SVM and what tests helped a lot. As part 2 has already been described, I think it makes sense here to mention I tried adapting this slightly for task 2 but that it went terribly.

4.1.2 Method 2: Clustering

Talk about the experimentation with clustering.

4.2 ScienceIE Subtask B - Key Phrase Classification

A section all about what I did for part 2

4.2.1 Word2Vec Classification

Talk about using word2vec to simply find a good way to quickly classify key phrases with decent results. *** Where do I fit the SVM for this, as not worth a whole section

4.3 ScienceIE Subtask C - Relation Extraction

A section all about what I did for part 3

4.3.1 Support Vector Machine

Discuss the SVM I tried to do this with (including Word2Vec)
and hopefully more to come...

5 The ScienceIE Task: Evaluation

How each section went, including test results and maybe some info on other experiments.

5.1 Subtask A - Key Phrase Extraction

5.1.1 Conclusion

5.2 ScienceIE Subtask B - Key Phrase Classification

5.2.1 Conclusion

5.3 ScienceIE Subtask C - Relation Extraction

5.3.1 Conclusion

6 Creating a Service

Write about the GUI and all that went into that (probably a similar length to NLP part 1, although less (academic) references). Should already been introduced.

6.1 Further Research

Discuss the resources used to design maybe? Make sure to include research on searching I did...

6.2 Design and Implementation

How it was pulled off

6.3 Web Interface

Exactly what was achieved

6.4 Testing

(Get) user feedback

6.5 Conclusion

Overall impact of the GUI on the project

7 Discussion

Talk about overall results

7.1 Improvements and Extensions

References

[Ammar et al.]Ammar, Peters, Bhagavatula Power2017Ammar2017 Ammar, W., Peters, M., Bhagavatula, C. Power, R. 2017, 'The AI2 system at SemEval-2017 Task 10 (ScienceIE): semi-supervised end-to-end entity and relation extraction', *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* **10**, 592–596.
http://www.aclweb.org/anthology/S17-2097 [Augenstein et al.]Augenstein, Das, Riedel, Vikraman McCallum2017Augenstein2017 Augenstein, I., Das, M., Riedel, S., Vikraman, L. McCallum, A. 2017, 'SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications', pp. 546–555.
http://arxiv.org/abs/1704.02853 [Liu et al.]Liu, Li, Zheng Sun2009Liu2009 Liu, Z., Li, P., Zheng, Y. Sun, M. 2009, 'Clustering to Find Exemplar Terms for Keyphrase Extraction', *Language* **1**, 257–266.
http://portal.acm.org/citation.cfm?doid=1699510.1699544 Manning Jurafsky2012Manning2012 Manning, C. Jurafsky, D. 2012, 'Using Patterns to Extract Relations'.
https://www.youtube.com/watch?v=VodeEgvxgtA [Marsi et al.]Marsi, Sikdar, Marco, Barik Sætre2017Marsi2017 Marsi, E., Sikdar, U. K., Marco, C., Barik, B. Sætre, R. 2017, 'NTNU-1$@$ScienceIE at SemEval-2017 Task 10: Identifying and Labelling Keyphrases with Conditional Random Fields', *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* pp. 937–940.
http://www.aclweb.org/anthology/S17-2162 [Snow et al.]Snow, Jurafsky Y. Ng2013Snow2013 Snow, R., Jurafsky, D. Y. Ng, A. 2013, 'Learning syntactic patterns for automatic hypernym discovery', *Journal of the American Medical Informatics Association* **20**(1), 1–11.
http://dx.doi.org/10.1186/s12859-015-0606-0%5Cnhttp://dx.doi.org/10.1016/j.jbi.2015.02.004%5Cnhttp://dx.doi.org/1
[Zhang et al.]Zhang, Wang, , Wu, Liao Wang2008Zhang2008 Zhang, C., Wang, H., , Y., Wu, D., Liao, Y. Wang, B. 2008, 'Automatic Keyword Extraction from Documents Using Conditional Random Fields', *Journal of Computational Information* **43**, 1169–1180.
http://www.jofci.org