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

1.1 Aims and Objectives

1.2 Report Outline

2 Background and Literature Review

Do the literature review here

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 Discussion

Talk about overall results

6.1 Improvements

7 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.

7.1 Further Research

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

7.2 Design and Implementation

How it was pulled off

7.3 Web Interface

Exactly what was achieved

7.4 Testing

(Get) user feedback

7.5 Conclusion

Overall impact of the GUI on the project

References

[1] "ScienceIE," 2016.

[2] C.-C. Chang and C.-J. Lin, "libsvm," 2016.

[3] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

[4] V. Hristidis, L. Gravano, and Y. Papakonstantinou, "Efficient IR-style keyword search over relational databases," *Vldb*, pp. 850–861, 2003.

[5] S. Agrawal, S. Chaudhuri, and G. Das, "DBXplorer: A system for keyword-based search over relational databases," *Proceedings - International Conference on Data Engineering*, pp. 5–16, 2002.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *World Wide Web Internet And Web Information Systems*, vol. 54, no. 1999-66, pp. 1–17, 1998.

[7] J. Goodman, "Parsing Algorithms and Metrics," *ReCALL*, no. June, pp. 177–183, 1996.

[8] N. Nayak, "Learning Hypernymy over Word Embeddings," *CS224N Projects*, pp. 1–8, 2015.

[9] C. Manning and D. Jurafsky, "Using Patterns to Extract Relations," 2012.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," pp. 1–12, 2013.

[11] Y. Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," no. 2, pp. 1–5, 2014.

[12] C. Manning and D. Jurafsky, "Using Patterns to Extract Relations," 2012.

[13] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, "Automatic Keyword Extraction from Documents Using Conditional Random Fields," *Journal of Computational Information*, vol. 43, pp. 1169–1180, 2008.

[14] Chih-Wei Hsu, Chih-Chung Chang and C.-J. Lin, "A Practical Guide to Support Vector Classification," *BJU international*, vol. 101, no. 1, pp. 1396–400, 2008.

[15] S. Winters-Hilt and S. Merat, "SVM clustering," 2007.

[16] R. Snow, D. Jurafsky, and A. Y. Ng, "Learning syntactic patterns for automatic hypernym discovery," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 1–11, 2013.

[17] Z. Liu, P. Li, Y. Zheng, and M. Sun, "Clustering to Find Exemplar Terms for Keyphrase Extraction," *Language*, vol. 1, pp. 257–266, 2009.

[18] I. Augenstein, M. Das, S. Riedel, L. Vikraman, and A. McCallum, "SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications," pp. 546–555, 2017.

[19] W. Ammar, M. Peters, C. Bhagavatula, and R. Power, "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)*, vol. 10, pp. 592–596, 2017.

[20] E. Marsi, U. K. Sikdar, C. Marco, B. Barik, and R. Sætre, "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, 2017.

[21] S. Gupta and C. Manning, "Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers," *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 1–9, 2011.

[22] K. S. Hasan and V. Ng, "Automatic Keyphrase Extraction: A Survey of the State of the Art," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1262–1273, 2014.

[23] L. D. Baker and A. K. McCallum, "Distributional clustering of words for text classification," *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, pp. 96–103, 1998.

[24] Y.-f. B. Wu, Q. Li, R. S. Bot, and X. Chen, "Domain-specific keyphrase extraction," *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05*, p. 283, 2005.