ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ARTIFICIAL INTELLIGENCE

MASTER THESIS

in

Natural Language Processing

Graph-Based Keyword Extraction from Scientific Paper Abstracts using Word Embeddings

Author:
Dinno Koluh

Supervisor:
Prof. Paolo Torroni

Co-Supervisor:
Dr. Federico Ruggeri

Bologna,
October 2023.

Abstract

In the era of information overload it became essential to efficiently extract concise, precise and quality information from large texts. One aspect of information extraction is keyword extraction where large texts are represented as sets of tokens and then the most important words i.e. keywords, represent that text. This prospect of keyword extraction is paramount to researchers as they deal with huge numbers of scientific papers, and having a good and concise representation of those papers is essential for them. This thesis paper addresses that problem in the realm of natural language processing (NLP).

Using core concepts of NLP and modeling texts as graphs, in this paper we are going to build a model for the automatic extraction of keywords. This is done in an unsupervised manner as the importance of a word is calculated through the position and weights associated with respective words in the graph. One of the sources of the word weights are word-embeddings as they became a crucial way of representing words as dense vectors.

The results of this paper were compared with keywords that were provided by authors of scientific papers in the area of computer science which act as the ground truth, but crucially are not a component in the model construction, but just serve as a verifier of the model's accuracy.

Keywords: NLP, keyword extraction, scientific papers, graphs, word-embeddings

Contents

List of Figures

1 Introduction and Motivation

Natural Language Processing (NLP) is a subfield or Artificial Intelligence (AI) and linguistics that has a focus on the interaction between computers and human languages. This is mostly restricted to written language as other fields (like Speech Processing) deal with spoken language (using audio instead of textual features). In the past decade NLP has seen huge attention with the introduction of some key concepts like word-embeddings [**?**] (which we are going to use) and transformers [**?**]. At the moment of writing of this paper, NLP is the subfield of AI that has the most resources invested into it, mostly in research and development, with new large language models (LLMs) coming out on a daily basis. Our focus will be shifted a bit from Deep Neural Network (DNN) models and more to the traditional Machine Learning (ML) models.
NLP has many subfields and application areas such as:

- Text classification
- Information retrieval
- Automatic translation
- Speech analysis
- Question answering
- Conversational agents
- Sentiment analysis

The field we are going to be working on will be information retrieval, more precisely *keyword extraction*. Keyword (keyphrase) extraction is the automatic selection of important and topical phrases from the body of a document [**?**]. Scientific papers are usually the area where keyword are most frequently used as researchers use them for a quick overview of the paper and also sorting paper into different categories. This is the topic we are going to be working on as well. The keywords are going to be extracted from the abstracts of the scientific papers. One detail we should address that will be important later is the distinction between *keywords* and *keyphrases*.

Keywords would be single words or at most MWEs (Multi-Word expressions) while keyphrases are a more complicated entity comprised of several words and they act as a single unit (e.g. the phrase "scientific paper" would be a keyphrase). When doing keyword extraction we might have as an output a combination of both, keywords and keyphrases. Usually keyphrases carry more information than keywords but a combination of both as an output is most representative. From now on, when referring to *keywords* it will also include *keyphrases*, if not explicitly stated otherwise.

The model to be used for keyword extraction is graph-based. The idea is to model words in an abstract as nodes of a graph. Not all the words in the abstract should be included, as keywords are usually composed of a combination of nouns and adjectives (e.g. scientific [ADJ] paper [N]). Then the question comes how to model the edges of the graph?

It starts from the assumption that words that occur in the same context tend to carry a similar meaning. So, the idea is to build a sliding window of some predefined size and words that are in that window are connected with and edge. In that way the final graph carries semantic (the meaning of words) information.

2 NLP Pipeline

3 Word Embeddings

4 Graph Construction

5 Implementation of Keyword Extraction

6 Testing, Results and Discussion

7 Conclusion

8 Literature

[1] Jaillet, Léonard, Juan Cortés, and Thierry Siméon. "*Transition-based RRT for path planning in continuous cost spaces.*" 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2008.
URL: https://hal.laas.fr/hal-01986342/document

[2] LaValle, Steven M. "*Rapidly-exploring random trees: A new tool for path planning*". Computer Science Department, Iowa State University (TR 98–11), October 1998.
URL: http://msl.cs.uiuc.edu/~lavalle/papers/Lav98c.pdf

[3] Kavraki L. E., Svestka P., Latombe J.C., Overmars M. H. "*Probabilistic roadmaps for path planning in high-dimensional configuration spaces.*". IEEE Transactions on Robotics and Automation, 1996.
URL: http://dspace.library.uu.nl/handle/1874/17328

[4] Kuffner, James J., and Steven M. LaValle. "*RRT-connect: An efficient approach to single-query path planning.*", Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065). Vol. 2. IEEE, 2000.
URL: http://kuffner.org/james/papers/kuffner_icra2000.pdf

[5] LaValle, Steven M. "*Planning Algorithms*". Cambridge University Press. ISBN 978-1-139-45517-6, May 2006.

[6] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo. "*Robotics: Modelling, Planning and Control*". Springer, London, UK, 2009.

[7] Kroese D. P., Brereton T., Taimre T., Botev Z. I. "*Why the Monte Carlo method is so important today*", 2014.

[8] Jurić Ž. "*Diskretna matematika za studente tehničkih nauka*". ETF Sarajevo, UNSA, 2017.

[9] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. "*Introduction to Algorithms*", Second Edition. MIT Press and McGraw-Hill. ISBN 0-262-03293-7. *Section 33.3: Finding the convex hull*, pp. 947–957, 2001.

[10] John D'Errico (2020). Efficient test for points inside a convex hull in n dimensions. MATLAB Central File Exchange. Retrieved December 11, 2020.
URL: https://www.mathworks.com/matlabcentral/fileexchange/10226-inhull

[11] Michael Yoshpe (2020). Distance from points to polyline or polygon. MATLAB Central File Exchange. Retrieved December 11, 2020.
URL: https://www.mathworks.com/matlabcentral/fileexchange/12744-distance-from-points-to-polyline-or-polygon

[12] Matt J. (2020). Analyze N-dimensional Polyhedra in terms of Vertices or (In)Equalities. MATLAB Central File Exchange. Retrieved December 10, 2020
URL: https://www.mathworks.com/matlabcentral/fileexchange/30892-analyze-n-dimensional-polyhedra-in-terms-of-vertices-or-in-equali

[13] `lsqlin` function description. Solve constrained linear least-squares problems.
URL: https://www.mathworks.com/help/optim/ug/lsqlin.html