PhD thesis proposal

Title :

Analysis of scientific papers for the automatic construction of states of the art

Context :

This thesis is a collaboration between the company Idexlab and La Rochelle University (L3i Lab) under a CIFRE contract. The physical location of the student is to be discussed with candidates, in Paris (Idexlab) or La Rochelle (L3i laboratory).

Based in Paris, Idexlab offers an innovation platform that allows the analysis of a large volume of data from scientific papers coming from different resources on the web and specialized databases in order to carry out a bibliography mapping. This step is a prior that will help to address scientific or technological issues on various subjects. The platform is designed to facilitate the construction of state of the arts by accessing the latest scientific knowledge, patents and peer-reviewed publications, thus ensuring the reliability of the information. Its availability means the PhD student will be able to focus on improving the state of the art within an existing pipeline.

The L3i laboratory, located at the University of La Rochelle, was founded in 1993 and has slightly over 100 members. Focused on the analysis of digital content created by and for humans, it is organised into 3 research teams. The thesis will be supervised within the "Images and Contents" team, whose activities were most recently evaluated by the HCERES as "highly visible at the international level" and "excellent" in terms of scientific production and publications. In this team, fifteen researchers work in natural language processing and artificial intelligence, focusing mainly on cross- and multi-lingual, and contrastive approaches. The main current NLP projects are the following: H2020 NewsEye 2018-2022, H2020 Embeddia 2019-2022, ANNA 2021-2024 and Termitrad 2022-2025). This work will be in close collaboration with the Embeddia and Termitrad projects, gu

Aim of the thesis: to identify emerging themes with knowledge extraction from scientific article databases in order to deploy artificial intelligence and natural

language processing solutions to improve their platform and provide it with new functionalities. This requires the analysis of numerous sources that can be written in different languages, and a representation of the knowledge extracted from these documents.

Keywords:

Natural language processing, information retrieval, artificial intelligence, text mining, knowledge management, knowledge extraction and visualization.

Use case :

In this section, we introduce a brief use case that helps better understand the Idexlab platform. Through this use case we highlight the need of Idexlab to automate their processes using artificial Intelligence and natural language processing. This is the main goal and scope of this thesis.

Idexlab's platform enables the extraction and presentation of the knowledge contained in various documentary resources. It also provides tools for the creation of state of the art structured in a visual form : tree structure, thus enabling better grasp of the complexity of the studied domains.

The following example illustrates a real-life case with a customer : the problem of fallen leaves on train tracks during autumn, causing delays in rail traffic. The customer wants to identify possible solutions to handle this problem.

The search engine allows to quickly explore different patents and papers related to the subject. Once the query is executed, the user can save the result in the form of abstracts. The abstracts are saved in the form of boxes (Figure 1), which title is the query itself.

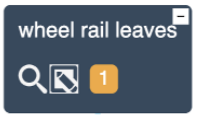

Figure 1 : knowledge box

Within the boxes, every abstract can be processed manually (highlighted) in order to identify the context, shortcomings, problem/objectives, means, results and prospects as shown in Figure 2. This makes it easier for the user to summarize the knowledge embedded in a given domain.

- **Context** - usually at the beginning
- **Shortcomings** - search for the words: however, but, although, previous, despite, ...
- **Objectives, problem** - introduced by: the objective, in this paper, this work, in this study, this approach, herein, in order to, ...
- **Means** - search for the words: through, using, we use, ...
- **Results** - search for the words: we provide, step(s) towards, we report, highlight, ...
- **Prospects** - usually at the end when present

example :

Marine plastic pollution is a global problem with considerable ecological and economic consequences. Quantifying the amount of plastic in the ocean has been facilitated by surveys of accumulated plastic on beaches, but existing monitoring programmes assume the proportion of plastic detected during beach surveys is constant across time and space. Here we use a multi-observer experiment to assess what proportion of small plastic fragments is missed routinely by observers, and what factors influence the detection probability of different types of plastic. Detection probability across the various types of plastic ranged from 60 to 100%, and varied considerably by observer, observer experience, and biological material present on the beach that could be confused with plastic. Blue fragments had the highest detection probability, while white fragments had the lowest. We recommend long-term monitoring programmes adopt survey designs accounting for imperfect detection or at least assess the proportion of fragments missed by observers.

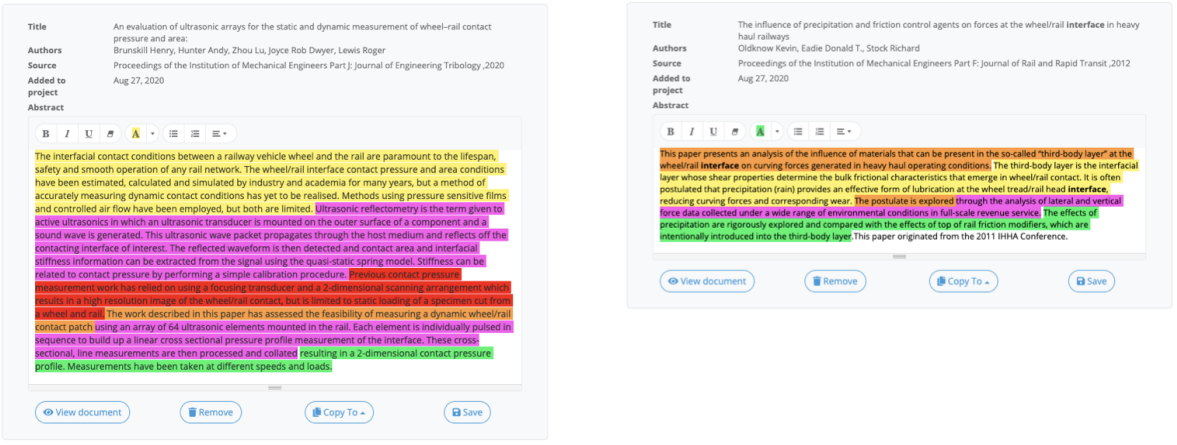

Figure 2 : Example of segmentation of an abstract

From Figure 2 we can see that the structure of abstracts varies from one paper to another, hence, it is sometimes difficult to classify a set of words to figure whether it is the context, objectives, etc.

During the process of exploration and creation of the boxes, the user can progressively order them to create a tree structure (e.g, mindmap) of accumulated knowledge. A common usage is to establish hierarchical links between the papers abstracts like : uses/ is used or studies/ is studied, as depicted in Figure 3.

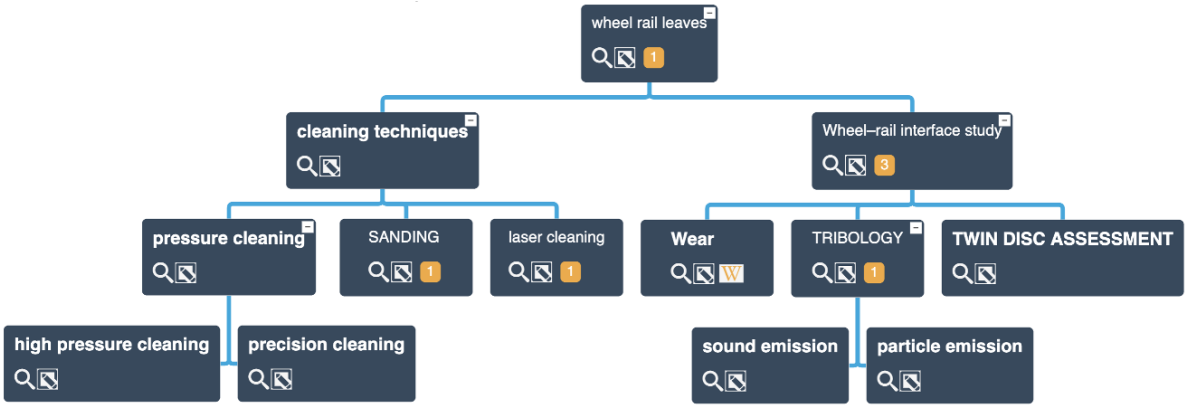

Figure 3 : Thematic tree organisation of abstracts

Combining the highlighting of abstracts and the tree visualisation is a powerful means for absorbing and sharing new knowledge.

In this respect, the main goal of this doctoral research is to explore different approaches allowing us to rely on machine learning in order to automate the processes. The automation can be achieved following two main steps : (i) knowledge extraction based on discursive analysis of abstracts for text segmentation; (ii) creation of taxonomies (tree structures of graphs) to organise and visualise the extracted knowledge.

Objectives of the research

The research will address the following questions :

1. How to automate the abstracts' segmentation and annotation to extract knowledge?
2. How to organise the extracted knowledge using a tree or a graph structure?

More specifically, the doctoral students will be in charge of the following tasks :

- Bibliographical analysis and summary;
- Proposal of a methodology to address the research question
- Abstracts' segmentation ;
- Definition of a learning model for the automatic segmentation of abstracts and classification of segments, trained with manually processed documents provided by Idexlab;
- Extraction of information from the discursive segmentation of the summaries by identifying their context;
- Creation of a taxonomy by ordering the extracted information and defining the rules that link them;

- Evaluation of the results and validation of the proposed methodology;
- Publication of the associated results in journals and conferences.

State of the art :

As regards the first question, the objective of the work will first be to draw up the state of the art of artificial intelligence approaches for the segmentation and classification of textual content, applied to scientific summaries. It requires a manual segmentation of a set of abstracts, which will then be used to train a learning model. This model will aim at the automatic segmentation of the discourse of paper abstracts. The annotations made to these summaries will be saved in the corresponding boxes. The aim here is to store and represent the knowledge extracted from the summaries, in order to share it with users who can query, visualise and cross-check it.

To achieve the first part, we can rely on methods related to word embedding [Mikolov et al., 2013], or even sequence embedding [Devlin et al., 2019], language models [Lavrenko and Croft, 2017], and language-independent differential approaches [Gross et al., 2016].

Combined with vectorial word presentation methods, known as embeddings, contextualized embedding representation models have shown a much higher performance to conventional methods for information retrieval tasks such as named entity recognition [Moreno et al., 2019], referential binding, or relationship classification, in monolingual or multilingual configurations.

These representations have raised the emergence of new models that need very few annocations to give good results. However, it has been shown that, when combined with supervised models, they perform better [Glavas et al., 2019].

These models have as main characteristics :

1. A dynamic vocabulary through the use of word pieces,
2. A sequence-to-sequence model (*sequence-to-sequence auto-encoders* or *neural language models*),
3. Training on large collections of data (auto-supervised learning),
4. They are very efficient when presented with specialised examples (fine tuning).

In particular, we will focus on these approaches applied in the context of terminology and keyword extraction [Hasan et Ng, 2014; Repar et al., 2019; Martinc et al., 2020[1]]. Keywords and terminology extraction are essential for this work and are

[1] with source code: https://github.com/EMBEDDIA/tnt_kid

involved in two ways : first, as potential markers to assist in the segmentation of abstracts and the classification of the resulting segments, but also to compare the abstracts and form the knowledge boxes (Figure 1) used to build the tree structure (Figure 3). This work will benefit from the regional project of Nouvelle-aquitaine “Termitrad” led by Antoine Doucet from L3i lab and dedicated to cross-domain and cross-lingual terminology extraction. The candidate will collaborate with 2 post-doctoral students and another doctoral student recruited over 2021-2024.

This step requires a manual segmentation and classification of a set of abstracts that will be used to train a learning model. The learning model is already implemented in the Idexlab Platform and will be made available for the doctoral student. This model will allow the automatic segmentation and thus the recognition of speech patterns. The annotations made to these summaries will be saved in the corresponding boxes.

There is a need to identify the context of the keywords in order to build a taxonomy that highlights the rules and links between the information. Then it is necessary to propose a methodology and its implementation, in order to automate this process.

We plan to combine two approaches : statistical based on machine learning, and symbolic based on the creation of ordered taxonomies.

The first step is the manual identification of sentences that represent the context, constraints, objectives, methods/techniques, results and perspectives (see Figure 2). The documents are taken from the databases connected to the Idexlab platform. An important underlying question will be the added value of using full papers rather than simple abstracts, and the related cost-benefit ratio.

In a second step, we will proceed to the extraction of relevant keywords from the segments defined in the previous step to represent them in the form of a graph. Knowledge graphs are a formal way to describe the knowledge of the studied domain [Yanze et al.; 2020; Buscaldi et al.; 2019].

These keywords, considered in their context, will be organised within a taxonomy using for example using ontologies. If so, it shall rely on the work of [Wątróbski; 2020] and [Konys; 2019] who present an overview of approaches for the automatic design of ontologies from texts. This automated construction is based on the concept of Ontology Learning (OL) which is part of ontology engineering and more broadly knowledge engineering. It is a process that includes two steps: (i) the realisation of a semantic model using the concepts extracted from the abstracts and (ii) the development of expressive rich languages that would allow, through reasoning, the automation of the ontology learning throughout its life cycle. The resulting graph would actually represent a Knowledge Base (KB) that would allow not only to gather the knowledge extracted from a set of articles dealing with the same domain, but also to cross-reference it with other information from another domain.

Thus, it would be possible for the company to integrate new knowledge through the OL and also to measure this in comparison with what has already been analysed. This can be achieved by relying on contrastive analysis methods, which we believe,

allows for a much more accurate analysis as suggested by studies in related fields such as named entity disambiguation [Gross et al. 2013] and multilingual summarisation [Gross et al. 2016]. This will allow the representation of the boxes in a tree form or a graph as suggested by [Qureshi et al.;2009].

Through this doctoral thesis, we aim to improve the extraction and management of knowledge within the Idexlab platform, by performing a document analysis that allows knowledge to be structured and presented in a global framework. However, the evolution of the work and parallel advances in the state of the art may lead the thesis work to focus on one or other of the topics listed.

Scientific supervisors :

The research work will be supervised by :

- Antoine DOUCET, Full professor (L3i Laboratory, La Rochelle University)
- Esma TALHI (Associate professor L3i Laboratory , EIGSI engineering school La Rochelle)
- Jean-Louis LIEVIN, CEO(Idexlab)

The doctoral student may either be based in IDEXLAB in Paris, or in L3i in La Rochelle (to be discussed with candidates). Regular exchanges will take place by video-conference as well as at least one physical meeting per month, mostly in Paris.

These arrangements may be reviewed or adapted to take into account the constraints linked to the current health crisis.

Bibliography :

Davide Buscaldi, Dessi Danilo, Motta Enrico, Osborne Francesco, Recupero Diego Reforgiato. ESWC (Satellite Events) - Mining Scholarly publications for Scientific Knowledge Graph Construction. The Semantic Web: ESWC 2019 Satellite Events, 2019

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL, 2019.

Goran Glavaš, Robert Litschko, Sebastian Ruder, Ivan Vulić. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. ACL (2019).

Oskar Gross, Antoine Doucet, Hannu Toivonen, Language-Independent Multi-Document Text Summarization with Document-Specific Word Associations, in Proceedings of the ACM Symposium on Applied Computing (SAC 2016), Pisa, Italy, p. 853-860, 2016.

Oskar Gross, Antoine Doucet and Hannu Toivonen, Term Association Analysis for Named Entity Filtering in Proceedings of the Text REtrieval Conference (TREC 2012), Gaithersburg, Maryland, USA, November 6-9, 10 pages, 2012.

Oskar Gross, Antoine Doucet and Hannu Toivonen, Named Entity Filtering based on Concept Association Graphs, in 14th International Conference in Computational Linguistics and Intelligent Text Processing (CICLing 2013), Samos, Greece, March 24-30, 12 pages, 2013.

Victor Lavrenko and Bruce Croft, Relevance-Based Language Models, ACM SIGIR Forum - SIGIR Test-of-Time Awardees 1978-2001, ACM New York, Volume 51 Issue 2, July 2017, Pages 260-267.

Y. Liu, C. Chan, C. Zhao et C. Liu, «Unpacking knowledge management practices in China: do institution, national and organizational culture matter?,» Journal of Knowledge Management, pp. 619-643, 2019.

M Martinc, B Škrlj, S Pollak. arXiv preprint arXiv:2003.09166. TNT-KID: Transformer-based Neural Tagger for Keyword Identification. 2020

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Proceedings of 2013 Conference on Neural Information Processing Systems (NIPS), 2013.

José G. Moreno, Elvys Linhares Pontes, Mickaël Coustaty, Antoine Doucet. TLR at BSNLP 2019: A Multilingual Named Entity Recognition System. 7th Workshop on Balto-Slavic Natural Language Processing, Aug 2019, Florence, Italy. pp.83-88,

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL), 2018.

Ingrid Petrič, Bojan Cestnik, Nada Lavrač, Tanja Urbančič; Outlier Detection in Cross-Context Link Discovery for Creative Literature Mining, The Computer Journal, Volume 55, Issue 1, 1 January 2012, Pages 47–61,

Senja Pollak et al. “NLP workflow for on-line definition extraction from English and Slovene text corpora”. In: 11th Conference on Natural Language Processing, KONVENS 2012, Vienna, Austria, September 19-21,

2012. 2012, pp. 53–60.

S. Qureshi, M. Kamal et p. Keen, «Knowledge networking to overcome the digital divide.,» Knowledge management and organizational learning, Boston, MA, Springer, , 2009, pp. 215-234.

Konys, Agnieszka. 2019. « Knowledge Repository of Ontology Learning Tools from Text ». Procedia Computer Science 159: 1614-28. https://doi.org/10.1016/j.procs.2019.09.332.

Repar, A., Martinc, M. & Pollak, S. Reproduction, replication, analysis and adaptation of a term alignment approach. Language Resources & Evaluation 54, 767–800 (2020).

Kazi Saidul Hasan and Vincent Ng. “Automatic Keyphrase Extraction: A Survey of the State of the Art”. In: Proceedings of ACL. 2014, pp. 1262–1273

Rui Wang, Wei Liu, and Chris McDonald. “Featureless Domain-Specific Term Extraction with Minimal Labelled Data”. In: Proceedings of the Australasian Language Technology Association Workshop 2016. 2016, pp. 103–112.

Wątróbski, Jarosław. 2020. « Ontology Learning Methods from Text - an Extensive Knowledge-Based Approach ». Procedia Computer Science 176: 3356-68. https://doi.org/10.1016/j.procs.2020.09.061.

Zhao Yanze, Huang Weiyuan, Pu Haitao. Domain knowledge graph-based research progress of knowledge representation. Neural Computing and Applications, 2020