Extraction of Epistemic Items from Scientific Publications

Amanda Agyeiwaah
University of Passau
agyeiw01@ads.uni-passau.de

Bhaumi Panchal
University of Passau
pancha01@ads.uni-passau.de

Umme Rabab
University of Passau
rabab01@ads.uni-passau.de

KEYWORDS

Rhetorical structure Theory (RST), Bayesian model, Natural Language Processing (NLP), keyword extraction

1 INTRODUCTION

It is an undeniable fact that the internet is plagued with many scientific papers, journals, conferences and articles. This is because day in and day out there are new research areas into interesting topics ranging from medicine, IT, agriculture, literature, entertainment, etc. The thirst for knowledge and advancement of technologies call for research and publications of scientific findings. This accounts for the ever-growing number of scientific papers published in various fields. These days students and researchers find themselves overwhelmed by the huge and rapid growth of scientific publications on the world wide web especially during literature analysis and assessment to find relations between studies in diverse domains.
Anytime there is the need to do research, one cannot shy from sampling a few research papers that have a bearing or aligns with the topic under study as this allows for easy comparison and drawing of motivation, approaches, methodologies and conclusions from various research literature to make one's own assessment; something we call information gathering in a particular field. This knowledge gathering becomes difficult when it is done entirely manually. Extracting information and establishing their syntactic and semantic meaning in these scientific publications can be daunting hence the need to apply NLP techniques to achieve semi-automatic knowledge extraction and classification methods relevant for decreasing time and effort spent on extracting information from scientific papers thus increasing efficiency. With the help of NLP techniques, only essential concepts are captured. When this is achieved, it will go a long way to help students and researchers sample any kind of research papers they want and easily extract the right information for the right purposes.

2 TECHNIQUES

2.1 Rhetorical Structure Theory (RST):

RST analysis is achieved by the ability to construct a tree based on a text read to establish relation between each span. RST comprises of a *nucleus* - the most important part of text and a *satellite* - an additional information to the nucleus. Without the nucleus, it is almost impossible to understand the text whereas the vice versa is true. We identify rhetorical relations between texts by considering cue phrases as a useful indicator. According to [5] several studies make use of RST to investigate linguistic issues and thus successful use of RST this way validates this assumption.
Figure 1, is a title and summary that appears on top of a publication. The text is broken into units and numbered as shown in the figure.

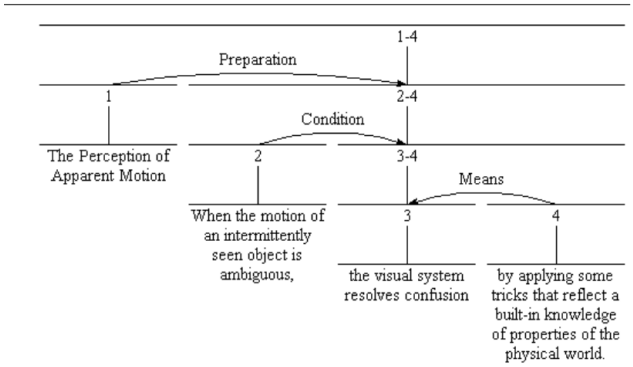

Figure 1: Diagram of RST Analysis.[4]

2.2 Bayesian Model:

The Bayesian model is very useful in natural language processing for prediction and classification purposes. Thus its relevance in our project is to

- Predict the probability of a claim in the papers.
- Predict if a claim belongs to the author of the paper or not.
- Predict the presence of anaphora in text. Another reason is to determine accurate predictions and compare non-nested models, which is not common practice in the frequent approaches.[2].

3 PROBLEM STATEMENT

Claims ought to be argumentative. Hence, a text or phrase qualifies to be a claim if you have to argue for a certain clarification or understanding of your subject matter. A good claim has to be specific. We are more interested in scientific claims which follow the natural language sentence structure by demonstrating the relationship that exist between two entities [3]Thus, how does one affect the other? In this paper, we propose an extension of [1] by:

- Identifying and extracting claims from scientific papers.
- Finding rhetorical relations that exist.
- Performing clustering analysis of scientific claims to find similar and dissimilar relations.
- Explore if an expression makes meaning by depending on another text called antecedent(presence of anaphora-a linguistic evidence). Example of Anaphora: Hanson dropped the ball. It bounced back quickly - The pronoun 'it' is an anaphor; it points to the left toward its antecedent 'the ball'.

4 OBJECTIVES

- We aim to model scientific argumentation to support information access
- We aim to provide basis for robustness in terms of extraction of information from scientific papers

Univeristät Passau, Text Mining Project, 2020
.

- We aim to determine the probability that these claims can lead to more scientific publications in future through classification
- We aim to find out if claims semantically contradict each other or not

REFERENCES

[1] Tudor Groza, Siegfried Handschuh, and Georgeta Bordea. 2010. Towards automatic extraction of epistemic items from scientific publications. In *Proceedings of the 2010 ACM Symposium on Applied Computing*. 1341–1348.

[2] Ioannis Ntzoufras, Athanassios Katsis, and Dimitris Karlis. 2005. Bayesian assessment of the distribution of insurance claim counts using reversible jump MCMC. *North American Actuarial Journal* 9, 3 (2005), 90–108.

[3] José María González Pinto, Janus Wawrzinek, and Wolf-Tilo Balke. 2019. What Drives Research Efforts? Find Scientific Claims that Count!. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 217–226.

[4] Maite Taboada and William C Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies* 8, 3 (2006), 423–459.

[5] Sandra A Thompson and William C Mann. 1987. Rhetorical structure theory: A theory of text organization. *The structure of discourse, Norwood NJ, Ablex* (1987).