Towards Automatic Extraction of Research Findings from the Literature

Dimitar Hristovski, PhD¹; Debra Revere, MLIS, MA²; Paul Bugni²; Sherrilynne Fuller, PhD²; Carol Friedman, PhD³; Thomas C. Rindflesch, PhD⁴
¹University of Ljubljana, Slovenia; ²University of WA, Seattle, WA, USA; ³Columbia University, NY, NY, USA; ⁴National Library of Medicine, Bethesda, MD, USA

Abstract. Most information extraction systems only process an article’s title and abstract. However, a major source of research findings is an article’s tables and figures. The aim of this study is to: (1) explore the efficacy of applying a hybrid information extraction system to the problem of identifying research findings in the scientific literature and (2) improve results by processing article title, abstract and table/figure text.

Introduction. Efforts to overcome information overload are being explored with literature-based discovery systems, much of this work is focused on an article's title and abstract. The Telemakus Knowledgebase System [1] provides a solution by analyzing, displaying and summarizing articles across domains with a focus on tables and figures which are a major source of research findings. Data represented in a document's tables and figures is represented as concept relationships. However, identification of concepts requires intensive human curation. Speeding this up so Telemakus can effectively scale for comprehensive treatment of domains has been a pressing problem.

Methods. Our approach is to use a combination of two existing information extraction systems, SemRep [2] and BioMedLee [3]. First, articles are prepared in a special text format that contains the article’s title, abstract, figure/table legends (captions) and text from the body of the article that describes the figures and tables. Our previous research has shown that figures and tables are the major source of research findings [1]. Next, prepared files are submitted to both SemRep and BioMedLee for processing. Output files are produced in the so-called common format, first mentioned in [4] and contain the extracted concepts and relations for each input text line. These results are post-processed as follows: (1) concepts and relations are scored using a weighted linear combination of several parameters and (2) results are transformed into a format for loading into the Telemakus database. For scoring, the most important parameters are the number of times a concept or relation is found in the article’s title, abstract or figure/table text lines and whether it appears in the Telemakus thesaurus and the assigned MeSH terms. In addition, extracted concepts and relations are expanded by including the more general concepts in the MeSH hierarchy.

Results. Using a human-curated corpus of documents as the standard, we processed 3500 lines of text from 98 articles, which included 520 figure/table legends. SemRep extracted 1836 and BioMedLee 1432 candidate relations respectively. A web-based tool was built for users to view, evaluate and select candidate concepts and relations. Further work is planned to evaluate candidate extractions against gold standard human curation to inform improvement of scoring and to apply machine learning to improve system weight-setting.

Conclusions. Our preliminary experience with this hybrid methodology is positive. We plan to continue development of these tools to support semi-automatic extraction and, eventually, use this method to automate processing so Telemakus can scale up to more rapidly build knowledgebases in new domains.

Acknowledgements. Carol Friedman was sponsored by grants LM007659 and LM008635 from the NLM.

References