Semantic Management of Scholarly Literature: A Wiki-based Approach

Bahar Sateli

Semantic Software Lab
Department of Computer Science and Software Engineering
Concordia University, Montréal, Canada

1 Introduction

The abundance of available literature in online repositories poses more challenges rather than expediting the task of retrieving content pertaining to a knowledge worker's information need. The rapid growth of the number of scientific publications has encouraged researchers from various domains to look for automatic approaches that can extract knowledge from the vast amount of available literature. Recently, a number of desktop and web-based applications have been developed to aid researchers in retrieving documents or enhancing them with semantic annotations [1, 2]; yet, an integrated, collaborative environment that can encompass various activities of a researcher from assessing the writing quality of a paper to finding complementary work of a subject is not readily available. The hypothesis behind the proposed research work is that knowledge-intensive literature analysis tasks can be improved with semantic technologies. In this paper, we present the *Zeeva* system, as an empirical evaluation platform with integrated intelligent *assistants* that collaboratively work with humans on textual documents and use various techniques from the Natural Language Processing (NLP) and Semantic Web domains to manage and analyze scholarly publications.

2 Research Question

The motivation behind our research is to improve the management and analysis of scholarly literature and aid knowledge workers in tasks like literature surveys and peer reviews, where a deep semantic processing of literature on a large scale is needed. To this end, we aim to develop an extensible platform to support the mentioned processes based on customizable literature analysis workflows. Provided with an intuitive user interface, various user groups, such as graduate students, peer reviewers or business analysts can create custom analysis workflows by combining sequences of system services (e.g., document clustering, entity extraction) to help them with their task at hand. In addition to generic bibliographic management services, the Zeeva system will provide its users with state-of-the-art NLP services that enrich the literature with semantic metadata suitable for both human and machine processing techniques.

We hypothesize that our tool can improve knowledge-intensive literature analysis tasks like reviewing papers with automatic detection of semantic entities in scholarly publications through a novel collaboration pattern between humans and AI *assistants*. To corroborate our hypothesis, the proposed research will be performed in four phases in an iterative and evolutionary fashion:

During the first phase, an extensive requirements elicitation will be performed in order to detect distinct and overlapping requirements of various user groups. The gathered requirements will help us to accommodate as many task-specific needs as possible in the design of our system with future extensibility and adaptation in mind. So far, more than 50 web-based and desktop bibliographic management systems, online digital libraries, open access repositories, scientific indexing engines and academic social networking websites have been studied to extract researchers' patterns and analysis workflows, as well as popular system features. An initial set of requirements have been gathered and a functional prototype has been built targeting researchers as the user group under study. The second phase encompasses the tasks of developing an extensible platform where various user groups can define custom analysis tasks by combining reusable processing components. For example, a graduate student can combine domain-specific *Information Retrieval* and *Extraction* services to obtain an overview of contributions from a set of papers, a journal reviewer can semantically compare the text of a submission against existing literature for plagiarism, and a business analyst can analyze trending topics in research by extracting affiliations and their corresponding contributions from literature of a specific domain. Although some generic NLP services such as *Named Entity Recognition* can be readily offered to Zeeva users through existing third-party libraries, other value-added, research-specific services identified from the elicited user requirements will be designed, developed and intrinsically evaluated for integration in the Zeeva system during the third phase. Finally, several user studies will be conducted with representative samples of each target user group to measure both the time needed to analyze a paper with and without the Zeeva text mining capabilities, as well as the quality of the generated results.

3 Background

In this section we provide a brief introduction of the tools and techniques used in the Zeeva prototype system, namely, the natural language processing domain and a framework for remote NLP service execution.

3.1 Natural Language Processing

Natural Language Processing (NLP) is an active domain of research that uses various techniques from the Artificial Intelligence and Computational Linguistics areas to process text written in a natural language. One popular technique from the NLP domain is *text mining* which aims at extracting high-quality structured information from free form text and representing them in a (semi-)structured

format based on specific heuristics. As an essential part of the literature curation process, text mining techniques prove to be effective in terms of the time needed to extract and formalize the knowledge contained within a document [3]. Motivated by similar goals, the Zeeva system aims to provide its users with a unified access point to a variety of text mining techniques that can help them with reading and analyzing scholarly publications.

3.2 Semantic Assistants

As the use of NLP techniques in software applications is being gradually adopted by developers, various tools have emerged to allow developers to include NLP capabilities in their applications through using third-party tools, such as OpenCalais,[1] without the need for a concrete knowledge of the underlying language processing techniques. In addition, a number of NLP frameworks, such as GATE [4], have also emerged to allow linguists to easily develop both generic and sophisticated language processing pipelines without a profound knowledge in software development. However, a seamless integration of these NLP techniques within external applications is still a major hindering issue that is addressed by the Semantic Assistants project [5]. The core idea behind the Semantic Assistants framework is to create a wrapper around NLP pipelines and publish them as W3C[2] standard Web services, thereby allowing a vast range of software clients to consume them within their context. Since different tasks in semantic literature support will need diverse analysis services, the service-oriented architecture of the Semantic Assistants framework facilitates our experiments by allowing us to easily add or remove services without the need to modify our system's concrete implementation.

4 Zeeva Wiki

One major user group of the Zeeva system are knowledge workers aiming at extracting recent trends, advancements, claims and contributions from available publications in a domain of interest. Researchers and curators usually deal with a large amount of information available on multiple online repositories and are in need of a centralized approach to manage and analyze them. One of the prominent characteristics of Zeeva's system design is to not only provide a collaborative environment for scholars to store scientific literature, but also to aid them in the literature analysis process using state-of-the-art techniques from the NLP and Semantic Computing domains.

Previously, in the Wiki-NLP [6] project, we investigated the feasibility and impact of integrating NLP capabilities within a wiki environment. The Zeeva prototype has been developed based on a wiki platform and makes extensive use of the Wiki-NLP integration in its fundamental core. The Zeeva wiki uses MediaWiki[3] as its front-end and expands its core functionality through the

[1] OpenCalais, http://www.opencalais.com
[2] World Wide Web Consortium, http://www.w3.org
[3] MediaWiki, http://www.mediawiki.org

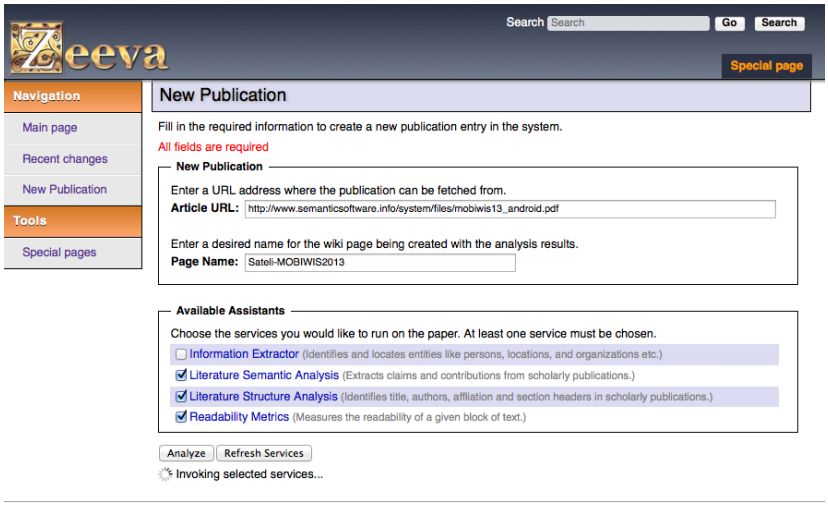

Fig. 1. Analyzing a sample publications with multiple assistants

installation of the *Semantic MediaWiki*[4] (SMW) extension. The underlying MediaWiki engine provides the basic wiki functionalities, such as user management and content revisioning, while the SMW extension allows us to augment the wiki content with so-called *semantic markup* and later query them directly within wiki pages.

The three aforementioned components, namely, the Zeeva Wiki, the Wiki-NLP integration and the Semantic Assistants server communicate with each other over the HTTP protocol. Users interact directly with the Zeeva wiki interface to create publication entries and verify the NLP results extracted from each document. In order for the NLP pipelines to have access to the text of articles, Zeeva provides users with a *special* page in the wiki, shown in Fig. 1, through which users can provide a URL and a desired name for the paper to be analyzed, as well as selecting one or multiple NLP *assistants* for the analysis task. Provided that the Wiki-NLP integration has adequate permissions to retrieve the article (e.g., from an open access repository or through an institutional license), the article is then passed on to all of the NLP pipelines chosen by the user. Each successful NLP service execution on the wiki generates metadata in form of SMW markup that is made persistent in the wiki database as semantic triples, hence, enriching the paper with a formal representation of the automatically extracted knowledge. Once all the pipelines are executed, the user is automatically redirected to the newly created page with the analysis results, transformed into user-friendly wiki templates like lists or graphs. Fig. 2 shows a list of claims and contributions extracted by Zeeva text mining pipelines from a sample paper. As the Zeeva's underlying wiki engine revisions all the changes to the wiki pages, users can review the services' output and modify the content in case of erroneous results.

[4]Semantic MediaWiki, http://semantic-mediawiki.org

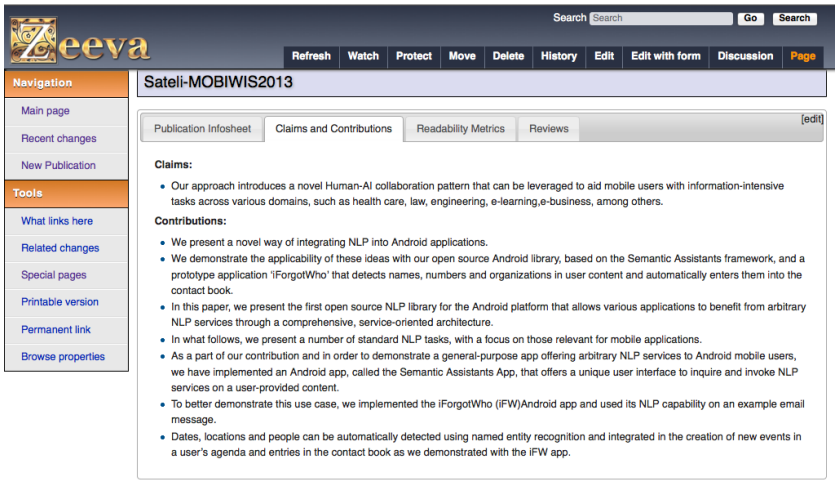

Fig. 2. Extracted semantic metadata from a sample publication

4.1 Literature Metadata Extraction

The integrated text mining capabilities of the Zeeva system is what distinguishes it from other literature management systems. In particular, Zeeva's literature analysis pipelines can extract two types of entities:

- *Structural Entities* refer to parts of the text that uniquely identify a paper, e.g., title or authors, as well as the parts that represent the structure of the paper, such as the abstract, section headers and references. Although, identification of such entities is a relatively easy task for human curators that can be achieved with a fair amount of effort, they stand essential as a leverage for extraction and disambiguation of the semantic entities.
- *Semantic Entities* refer to parts of the text that describe the contributions, claims, findings and conclusions postulated by the paper's authors. This is the most time-consuming task during the analysis process, as it requires the curators to manually read through the full text of each article.

Zeeva's literature structural and semantic analysis pipelines are developed based on the General Architecture for Text Engineering (GATE) [4] and uses the ANNIE plug-in to pre-process the text and detect named entities, such as persons and organizations. Zeeva transducers then look for specific sequences of person, location and organization named entities to generate structural entities, e.g., affiliations. Subsequently, the Zeeva pipelines try to extract semantic entities, in particular, claims and contributions from the paper by finding specific sequences of word tokens based on three gazetteer lists: *(i)* a segmentation trigger list that looks for variation of tokens such as "*Results*" and "*Discussion*", *(ii)* a list of tokens to extract authors' contributions such as "*our framework*" or "*the proposed approach*", and *(iii)* a list of claim triggers to extract sentences with comparative voice, e.g., "Our approach is faster than X", or claims of novelty.

An intrinsic evaluation of the Zeeva pipeline has been performed on a corpus of 5 manually-annotated open access Computer Science papers written by the author. The corpus documents are of various lengths (4 to 12 pages) and formatting styles (IEEE, ACM and LNCS) to ensure the applicability of the text mining pipeline. On average, the Zeeva literature analysis pipelines yield 0.85 and 0.77 F-measures on the tasks of finding structural and semantic entities, respectively.

5 Conclusions and Future Work

The main question in this research work is to evaluate whether concrete literature analysis tasks, like finding contributions, related work or writing peer reviews can be improved using state-of-the-art semantic technologies. In order to investigate the support needed for such knowledge-driven, literature-based tasks, we introduced Zeeva, an empirical wiki-based evaluation platform with an extensible architecture that allows researchers to invoke various natural language processing pipelines on scientific literature and subsequently, enriches the documents with semantic metadata for further human and machine processing. As an example, we developed two literature analysis pipelines that can automatically extract structural and semantic entities from a given full-text paper and generate semantic metadata from the results. Throughout our research, we are planning to identify literature analysis tasks that can be improved with semantic technologies and develop more NLP services relevant to the context of literature analysis. We are also planning to perform an extrinsic evaluation of our hypothesis using the Zeeva platform to assess the usability and efficiency of the proposed approach.

References

1. Zeni, N., Kiyavitskaya, N., Mich, L., Mylopoulos, J., Cordy, J.R.: A lightweight approach to semantic annotation of research papers. In: Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems. NLDB'07, Berlin, Heidelberg, Springer-Verlag (2007) 61–72
2. Yang, Y., Akers, L., Klose, T., Yang, C.B.: Text mining and visualization tools Impressions of emerging capabilities. World Patent Information **30**(4) (2008) 280 – 293
3. Sateli, B., Meurs, M.J., Butler, G., Powlowski, J., Tsang, A., Witte, R.: IntelliGen-Wiki: An Intelligent Semantic Wiki for Life Sciences. In: NETTAB 2012. Volume 18 (Supplement B)., Como, Italy, EMBnet.journal (2012) 50–52
4. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y., Peters, W.: Text Processing with GATE (Version 6). University of Sheffield, Department of Computer Science (2011)
5. Witte, R., Gitzinger, T.: Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients. In: 3rd Asian Semantic Web Conference (ASWC 2008). Volume 5367 of LNCS., Bangkok, Thailand, Springer (2008) 360–374
6. Sateli, B., Witte, R.: Natural Language Processing for MediaWiki: The Semantic Assistants Approach. In: The 8th International Symposium on Wikis and Open Collaboration (WikiSym 2012), Linz, Austria, ACM (2012)