Semantic MediaWiki (SMW) for Scientific Literature Management

Bahar Sateli, René Witte

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

SMWCon Spring 2014

Outline

Motivation

- Abundance of publications leads to bottlenecks in curating literature
- Existing bibliography management systems have limited content analysis support
- We need an environment that can encompass various activities of a researcher
- We hypothesize that our tool can improve knowledge-intensive literature analysis tasks through a novel collaboration pattern between humans and AI assistants.

We envision a collaborative, wiki-based solution for the semantic management of research literature that integrates:

- a web-based interface
- semantic knowledge representation
- text mining for automatic content analysis

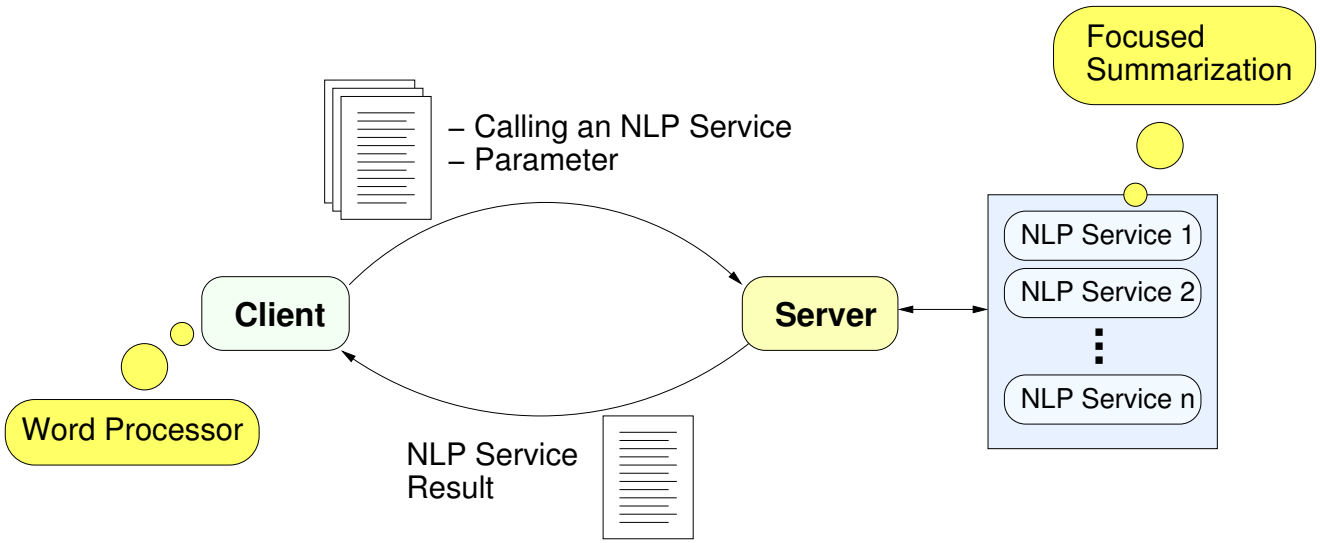
Related Work

- Post-publication semantic analysis
  - WikiPapers[1] uses Semantic Forms to collect literature focused on wiki research
  - AcaWiki[2], designed to collect summaries and literature reviews of peer-reviewed academic research
- Pre-publication semantic enrichment
  - The SALT (Groza et al., 2007) framework uses custom LaTeX commands with explicit semantics
- Our work is complementary to these efforts: we generate bibliographical and semantic entities using human-AI collaboration
- We transform papers into queryable artifacts, while remaining amenable to human-created semantic annotation

[1] WikiPapers, http://wikipapers.referata.com
[2] AcaWiki, http://www.acawiki.org

Natural Language Processing (NLP)

- A branch of AI that uses various techniques to process content written in natural language
- Multitude of NLP techniques exist, e.g.,
  - Named Entity Recognition (e.g., finding Persons, Organizations, etc.)
  - Quality Assessment
  - Summarization
- Various NLP APIs (e.g., OpenCalais, GATE, . . .)
- Semantic Assistants framework



Requirements

- **Centralized Repository of Knowledge (R1)**
  The system must provide users with the ability to store raw data as well as any information generated by users and analysis tools

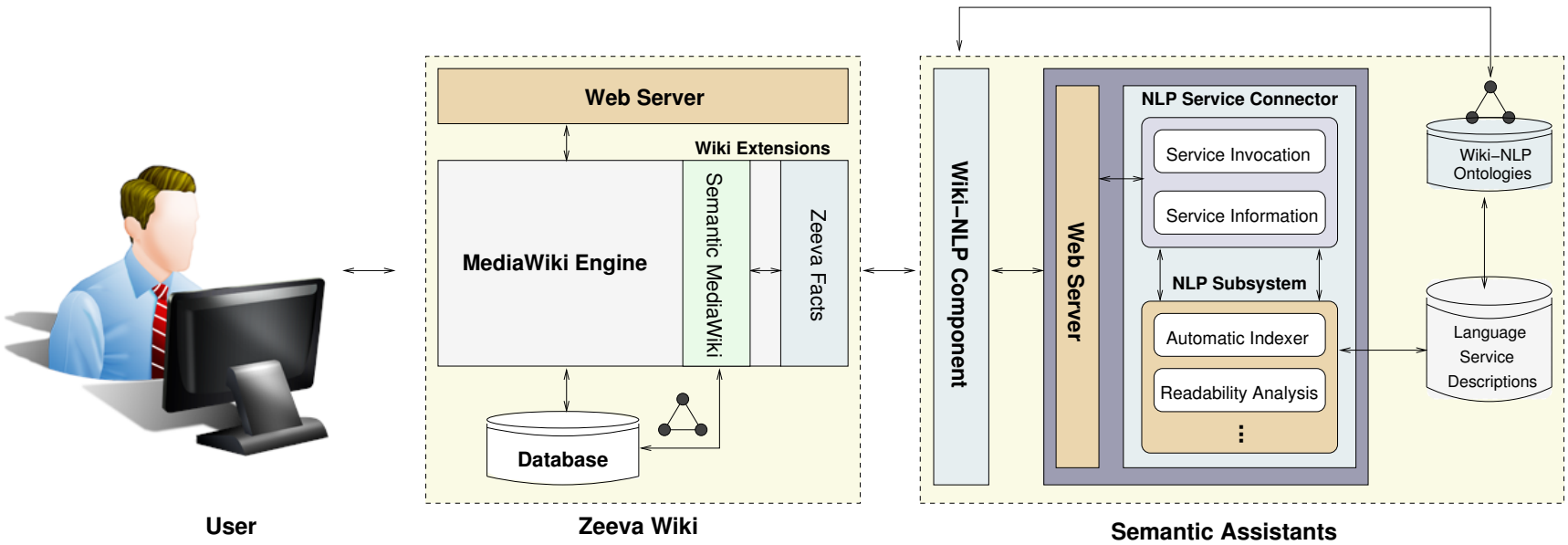
- **Automatic Text Analysis Support (R2)**
  The proposed system must provide access to various NLP pipelines in a unified manner.

- **Collaborative Analysis Environment (R3)**
  The proposed system shall provide an environment where all researchers have access to the most up-to-date information and can keep track of content modifications.

System Architecture

Design Decisions:

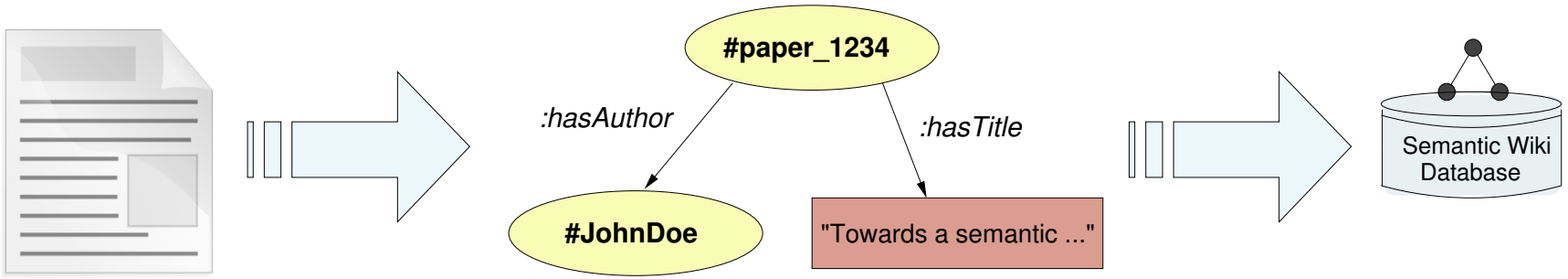
- Wiki-based Collaborative Web Interface (R3)
- Semantic MediaWiki as a Knowledge Base (R1)
- Text Mining Pipelines for Literature Analysis (R2)



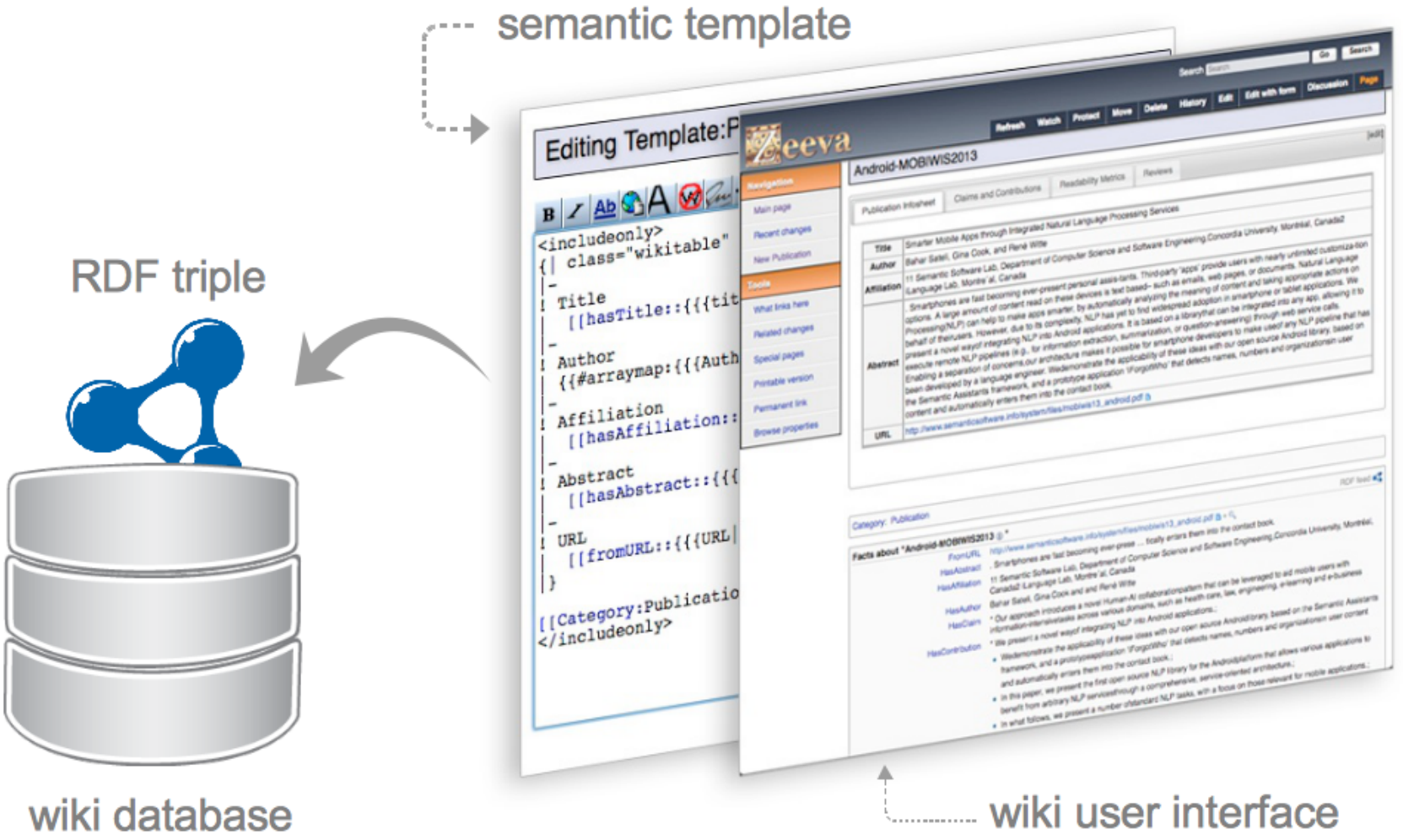
Semantic Metadata Extraction

Given a paper, we are interested in extracting:

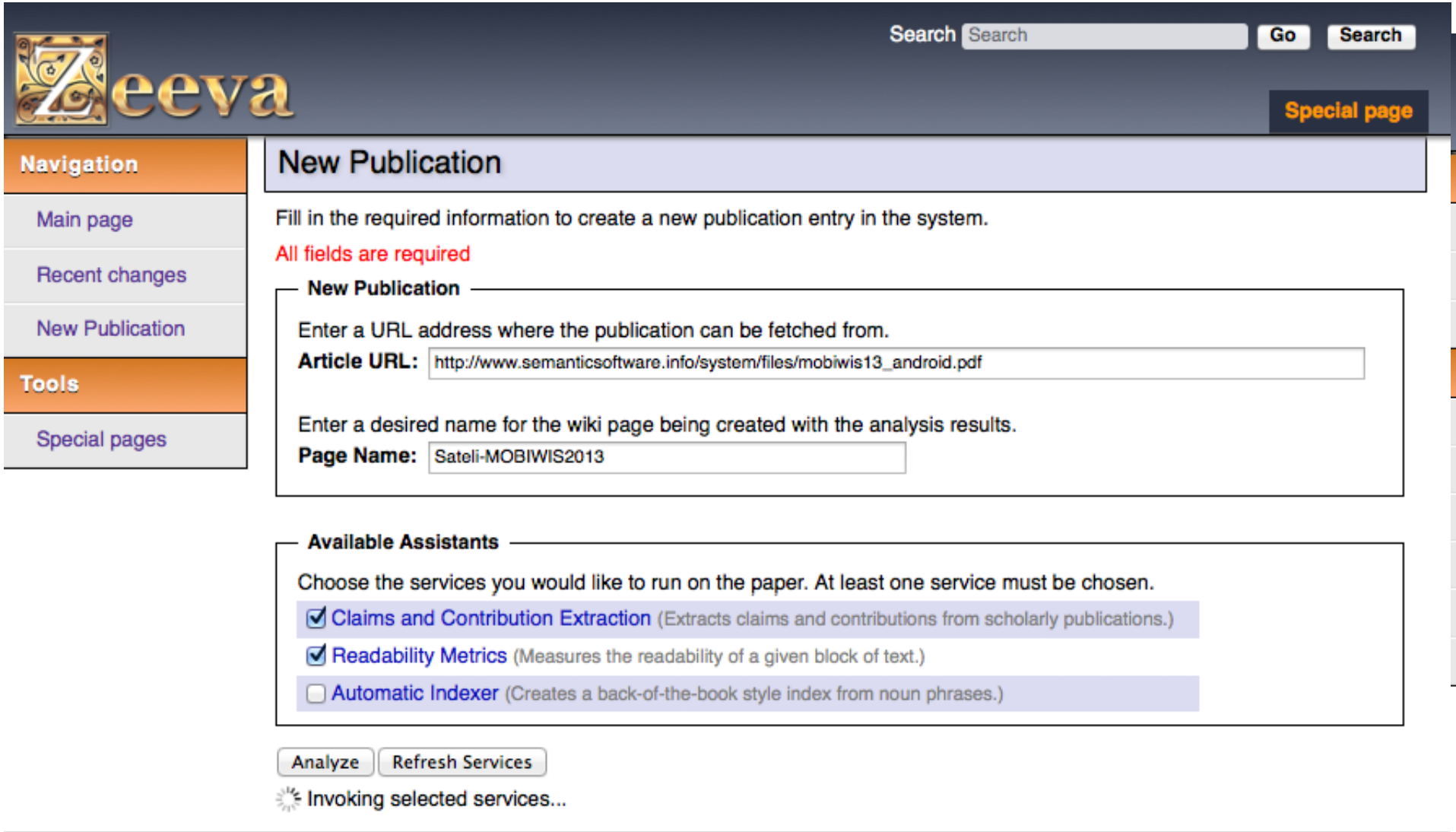
- *Structural Entities*, i.e., parts of text that uniquely identify a paper, like title or author
- *Semantic (Rhetorical) Entities*, i.e., parts of text that describe the contributions, claims, findings and conclusions postulated by the papers authors



Templating Mechanism



Automatic Processing of Publications



Querying the Zeeva Knowledge Base

- Transform Zeeva from a collaborative analysis environment to a knowledge base
- The generated semantic metadata can be used within wiki or exported to external applications
- Semantic MediaWiki provides a simple inline query syntax, e.g.,

NL Question: "Give me all the contributions of Bahar Sateli."

Corresponding SMW query:

```
{{#ask: [[Category: Publication]] [[hasAuthor:: Bahar Sateli]]
      | ?hasTitle = Title
      | ?hasContribution = Contribution}}
```

Advantages:

1. The results queried by the system are always up-to-date
2. Lets users *discover* knowledge created by other users of the wiki

Querying the Zeeva Knowledge Base

All Contributions of Bahar Sateli

Paper ID	Title	Contribution
ReqWiki-MUD2012	Can Text Mining Assistants Help to Improve Requirements Specifications?	We integrated a number of text mining assistants into a wiki-based requirements engineering platform to investigate two key questions: Can software engineers without prior training in NLP effectively leverage these techniques? In order to understand how software engineers would interact with these new text mining assistants during the development of the specification,we performed a case study within two university courses in requirements engineering (one at the undergraduate level, one at the graduate level). We conducted a number of experiments to evaluate the NLP integration for our ReqWiki system along two dimensions,namely the usability of the text mining service for non-experts and their effectiveness for quality improvements.
Sateli-MOBIWIS2013	Smarter Mobile Apps through Integrated Natural Language Processing Services	We present a novel way of integrating NLP into Android applications. We demonstrate the applicability of these ideas with our open source Android library, based on the Semantic Assistants framework, and a prototype application 'iForgotWho' that detects names, numbers and organizations in user content and automatically enters them into the contact book. In this paper, we present the first open source NLP library for the Android platform that allows various applications to benefit from arbitrary NLP services through a comprehensive, service-oriented architecture. In what follows, we present a number of standard NLP tasks, with a focus on those relevant for mobile applications. As a part of our contribution and in order to demonstrate a general-purpose app offering arbitrary NLP services to Android mobile users, we have implemented an Android app, called the Semantic Assistants App, that offers a unique user interface to inquire and invoke NLP services on a user-provided content. To better demonstrate this use case, we implemented the iForgotWho (iFW)Android app and used its NLP capability on an example email message. Dates, locations and people can be automatically detected using named entity recognition and integrated in the creation of new events in a user's agenda and entries in the contact book as we demonstrated with the iFW app.

Summary and Future Works

- The main question in this research is to evaluate whether concrete literature analysis tasks can be improved using state-of-the-art semantic technologies.
- We introduced Zeeva, an empirical wiki-based evaluation platform with an extensible architecture
- Identify literature analysis tasks that can improved with semantic technologies
- Develop more NLP services relevant to the context of literature analysis
- Perform an extrinsic evaluation of our hypothesis to assess the usability and efficiency of the proposed approach

http://www.semanticsoftware.info/zeeva