Supporting Wiki Users with Natural Language Processing

Bahar Sateli and René Witte
Semantic Software Lab
Department of Computer Science and Software Engineering
Concordia University, Montréal, QC, Canada
[sateli,witte]@semanticsoftware.info

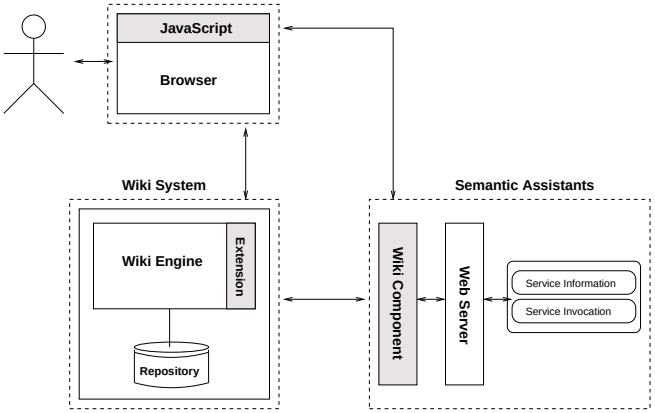


Figure 1: High-level Design of our Wiki-NLP Integration

WikiSym '12, Aug 27–29, 2012, Linz, Austria.
ACM 978-1-4503-1605-7/12/08.

Abstract

We present a "self-aware" wiki system, based on the MediaWiki engine, that can develop and organize its content using state-of-art techniques from the Natural Language Processing (NLP) and Semantic Computing domains. This is achieved with an architecture that integrates novel NLP solutions within the MediaWiki environment to allow wiki users to benefit from modern text mining techniques. As concrete applications, we present how the enhanced MediaWiki engine can be used for biomedical literature curation, cultural heritage data management, and software requirements engineering.

Author Keywords

Natural Language Processing; Semantic Assistants; Wiki Systems; Human-AI Collaboration Patterns

ACM Classification Keywords

H.3.1 [Content Analysis and Indexing]: Abstracting methods, Indexing methods, Linguistic processing; H.5.2 [User Interfaces]: Natural language, User-centered design; H.5.4 [Hypertext/Hypermedia]: Architectures, Navigation, User issues; I.2.1 [Applications and Expert Systems]: Natural language interfaces; I.2.7 [Natural Language Processing]: Text analysis

Wiki-NLP Integration Architecture

The Wiki-NLP integration is a collaborative approach that combines the power of a lightweight MediaWiki extension with a server-side wiki component. While the extension is responsible for the wiki-specific tasks, such as patrolling content changes, the wiki component plays the role of an intermediator between the user's browser, the wiki engine and the Semantic Assistants framework [5] – an open source project that brokers NLP pipelines as context-sensitive web services or *assistants*.

The Wiki Component shown in Figure 1 is essentially an HTTP proxy server that dynamically creates a wiki-independent interface for the Wiki-NLP integration and injects it to the user's browser, thus, giving users the impression that they are still working with the wiki's native interface.

Our solution is designed from the ground up for scalability, robustness, and is based on fully open source software.

Introduction

Natural Language Processing is a branch of computer science that employs various Artificial Intelligence (AI) techniques to process content written in natural language. The presented work is based on our previous idea that NLP-enhanced wikis can support users in finding, developing and organizing knowledge contained inside the wiki repository [4]. We realized this idea by developing a comprehensive architecture that offers novel NLP solutions within a wiki environment through a user-friendly and dynamically-generated user interface [3].

Motivation

By demonstrating our Wiki-NLP architecture, we want to exhibit how a seamless integration of NLP techniques into wiki systems helps to increase their acceptability and usability as a powerful, yet easy-to-use collaborative platform. The feedback we will gather will help us to identify new human-computer interaction patterns, allowing us to further enhance the Wiki-NLP integration architecture, in particular its user interface and identify new NLP services useful to the wiki context.

Demonstration

The presented work is essentially a general architecture for *enhancing* the MediaWiki engine with NLP techniques, rather than a new wiki system. This means that the architecture can be applied to any MediaWiki instance. Also, since the Wiki-NLP is a service-oriented architecture, we will demonstrate how the same architecture can deliver a multitude of NLP solutions to wiki systems. Therefore, during our demonstration we will first describe our Wiki-NLP integration architecture [1] and then present three different wikis, albeit with the same underlying MediaWiki engine.

Scenario 1: Biomedical Literature Curation

In Scenario 1, we demonstrate *GenWiki* [2] – a wiki for collaborative biomedical literature curation. Literature curation is a labour-intensive and time-consuming task, during which researchers extract relevant knowledge from a massive amount of literature available in multiple repositories. Recently, efforts have been made to automate the curation task by using advanced techniques from the NLP domain. However, employing these techniques usually requires the curators to have expertise in NLP or use specialized applications. In GenWiki, on the other hand, the motivation is to hide the complexity of applying NLP techniques on the wiki content from the point of view of the users, by bringing the NLP services directly into the wiki environment – thereby eliminating the need for an external application.

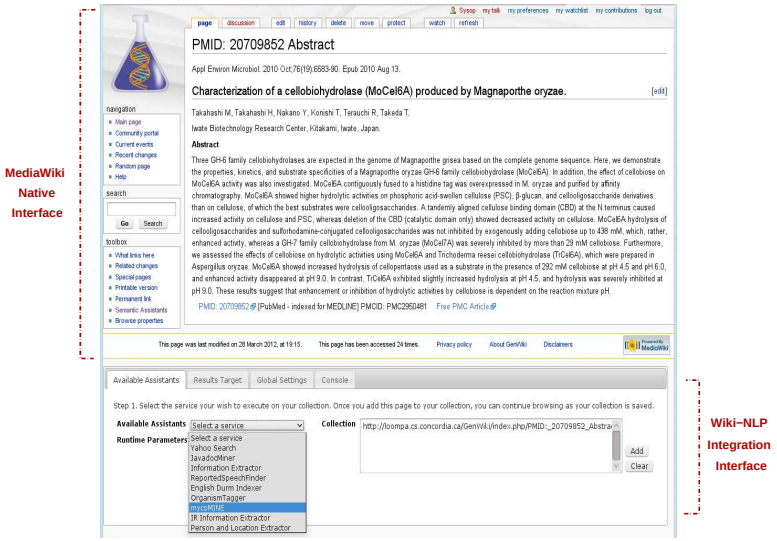


Figure 2: Wiki-NLP Integration Interface in GenWiki

Semantic Entity Retrieval

The Wiki-NLP integration adds a new discovery paradigm to an underlying wiki engine through providing semantic entity retrieval capabilities. By enriching the wiki content with NLP-derived metadata, wiki users are now able to retrieve various detected entities using their semantic properties, such as their types.

Figure 3 shows how, using a Semantic MediaWiki (SMW) inline query, GenWiki users can find wiki pages that contain entities of type *Enzyme*, as detected by an NLP service.

```
{{#ask: [[hasType::Enzyme]]
   | ?Enzyme = Enzyme Entities Found
   | format = table
   | headers = plain
   | default = No pages found!
   | mainlabel = Page Name
}}
```

Property:Enzyme

Page Name	Enzyme Entities Found
PMID: 20709852	Cellobiohydrolase
	Cellulases
	endoglucanases
	β-glucosidases
	Invitrogen
	DNA polymerase

Figure 3: Semantic Entity Retrieval in GenWiki

Figure 2 presents the GenWiki user interface, featuring a wiki page that contains the abstract of a paper. While the extraction of knowledge from a wiki page in a manual curation approach traditionally involves investigating the content and switching contexts in order to retrieve additional information, e.g., from a web search, in GenWiki users can achieve this goal by using the NLP services integrated into the wiki. A new menu item in the GenWiki toolbar allows users to request the Wiki-NLP user interface from any wiki page. Once the request is processed, the interface is injected into the GenWiki native interface to allow users to inquire about and invoke NLP services related to their task at hand – e.g., to automatically find entities such as *enzymes* or *organisms*.

We also present our experiments with GenWiki in a real-world project that highlights the impact of integrating automatic text mining pipelines within a wiki-based curation literature workflow, as we found it decreased the full paper curation time by 20% [1].

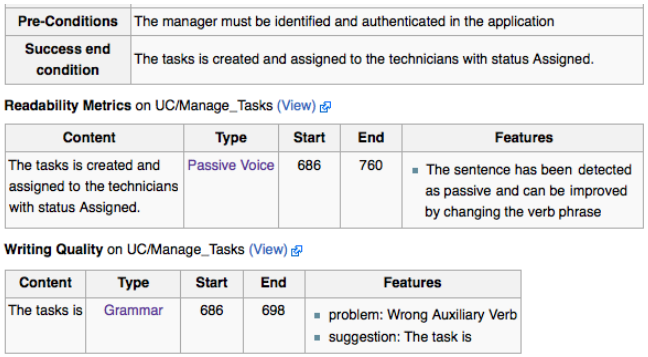
Scenario 2: Cultural Heritage Data Management
Cultural heritage data of a society, such as books, are often preserved in a digitized format and stored in distributed repositories. Such a body of content can be turned into a knowledge base accessible to both humans and machines using modern techniques from the Semantic Computing domain. In our second scenario, we present the *DurmWiki* [6] – a wiki containing a digitized version of a German historical encyclopedia of architecture. As browsing and keyword-based search are the only information retrieval means of a classical wiki system, discovering significant knowledge is a major challenge for users of the heritage data. This is further compounded by the fact that these texts contain outdated terminology no longer in use.

In DurmWiki, we demonstrate how an NLP service can perform automatic indexing of a wiki's content, storing it in the wiki itself, similar to classical back-of-the-book indexes. The generated index page, as shown in Figure 4, presents an alphabetically-ordered list of wiki terms and a direct link to their pages inside the wiki. In experiments with end users, we found that the presence of such an automatically maintained index page not only aggregates the wiki's embodied content on a high-level and enables users to find information at a glance, but also helps them to "discover" interesting concepts or entities that they did not know were present in the wiki [6].

Figure 4: Automatic Index Generation of DurmWiki Content

Scenario 3: Software Requirements Engineering
Software requirements engineering is the process of eliciting and documenting the needs of various stakeholders of a software project. Wikis, as an affordable, lightweight documentation and distributed collaboration

platform, have demonstrated their capabilities in requirements engineering processes. However, because of the lenient structure of wikis and the natural language that is used in software requirements specifications (SRS), the presence of semantic defects, such as ambiguity or vagueness, in SRS documents is inevitable. *ReqWiki* is our third scenario wiki, where we showcase the impact of NLP services on the *quality* of wiki content. In ReqWiki, users can invoke various generic or domain-specific quality assurance NLP services on the SRS documents using the Wiki-NLP user interface, in order to detect and amend the extracted defects. Figure 5 shows the results of a readability and a writing quality analysis service invoked on a use case document excerpt.



Pre-Conditions	The manager must be identified and authenticated in the application
Success end condition	The tasks is created and assigned to the technicians with status Assigned.

Readability Metrics on UC/Manage_Tasks (View)

Content	Type	Start	End	Features
The tasks is created and assigned to the technicians with status Assigned.	Passive Voice	686	760	▪ The sentence has been detected as passive and can be improved by changing the verb phrase

Writing Quality on UC/Manage_Tasks (View)

Content	Type	Start	End	Features
The tasks is	Grammar	686	698	▪ problem: Wrong Auxiliary Verb ▪ suggestion: The task is

Figure 5: Quality Assurance of Wiki Content in ReqWiki

During the demonstration, we will also present our experiments with Software Engineering students that used ReqWiki for their course assignments, which corroborate our hypothesis that employing NLP techniques in a wiki installation can significantly improve the quality of its content [1]. Moreover, a usability study with the same group showed that users unfamiliar with NLP technology can easily apply the offered Semantic Assistants [1].

Conclusion

Natural language processing has become an important tool for information and knowledge management. NLP techniques, such as question-answering, automatic summarization, information extraction, or classification can offer tremendous benefits in the context of wikis: Humans can now work collaboratively with semantic assistants that help them analysing, editing, and creating textual wiki content. This demo highlights some application scenarios, which we hope will inspire other users to adopt NLP techniques for their wiki of choice.

References

[1] B. Sateli. A General Architecture to Enhance Wiki Systems with Natural Language Processing Techniques. Master's thesis, Concordia University, Montréal, QC, Canada, 2012.

[2] B. Sateli, C. Murphy, R. Witte, M.-J. Meurs, and A. Tsang. Text Mining Assistants in Wikis for Biocuration. In *5th International Biocuration Conference*, page 126, Washington DC, USA, 04/2012 2012. International Society for Biocuration, International Society for Biocuration.

[3] B. Sateli and R. Witte. Natural Language Processing for MediaWiki: The Semantic Assistants Approach. In *Proceedings of the 8th International Symposium on Wikis and Open Collaboration (WikiSym'12)*, Linz, Austria, 2012.

[4] R. Witte and T. Gitzinger. Connecting Wikis and Natural Language Processing Systems. In *WikiSym '07: Proceedings of the 2007 International Symposium on Wikis*, pages 165–176, Montréal, Québec, Canada, 2007.

[5] R. Witte and T. Gitzinger. Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients. In *3rd Asian Semantic Web Conference (ASWC 2008)*, volume 5367 of *LNCS*, pages 360–374. Springer, 2008.

[6] R. Witte, T. Kappler, R. Krestel, and P. C. Lockemann. *Integrating Wiki Systems, Natural Language Processing,*

and Semantic Technologies for Cultural Heritage Data Management, pages 213–230. Theory and Applications of Natural Language Processing. Springer, 2011.