

Semantic Assistants for Wiki Systems

- [Semantic Assistants](#)
- [Semantic Wiki](#)
- [NLP](#)

Introduction

Semantic Assistants for wikis are our novel architecture for the integration of Natural Language Processing (NLP) capabilities into wiki systems, based on the [Semantic Assistants](#) framework. The vision is that of a new generation of wikis that can help developing their own primary content and organize their structure by using state-of-the-art technologies from the NLP and Semantic Computing domains. The motivation for this integration is to enable wiki users – novice or expert – to benefit from modern text mining techniques directly within their wiki environment. You can find additional background and details in our publication [1], [Sateli, B.](#), and [R. Witte](#), "[Natural Language Processing for MediaWiki: The Semantic Assistants Approach](#)", *The 8th International Symposium on Wikis and Open Collaboration (WikiSym 2012)*, Linz, Austria : ACM, 08/2012. The full details of this work are described in [2].

For a number of real-world application examples have a look at our [Wiki-NLP integration showcase](#). A hands-on tutorial was presented at SMWCon Spring 2014 [3].

Natural Language Processing in Wiki Systems: Use Cases

In the design and implementation of our Wiki-NLP integration, we took three major use cases into account.

Text Mining Assistants for Wiki Users

The primary motivation for this integration is to enable wiki users – novice or expert – to benefit from modern text mining techniques directly within their wiki environment. Wikis have become powerful knowledge management platforms, while remaining easy to use and offering high customizability, from personal wikis to enterprise solutions. With a majority of content in natural language, wikis can greatly benefit from natural language processing techniques. Rather than relying on external NLP applications, we aim to bring NLP as an integrated feature to wiki systems, thereby creating new human/AI collaboration patterns, where users work together with semantic assistants on developing, structuring and improving wiki content.

Wikis as Corpora for NLP Researchers

Wikis have long been recognized as a useful source for natural language processing experiments. In this use case, the content of a wiki is accessed to develop or improve novel natural language processing tools or resources. Our integration facilitates this task through its highly flexible access methods that allow to extract a single page, a set of pages, or a whole namespace from a wiki, with metadata distinguishing content and discussion pages, as well as their history. Unlike other, existing solutions, our Wiki-NLP integration works on live content, rather than a static database or page dump.

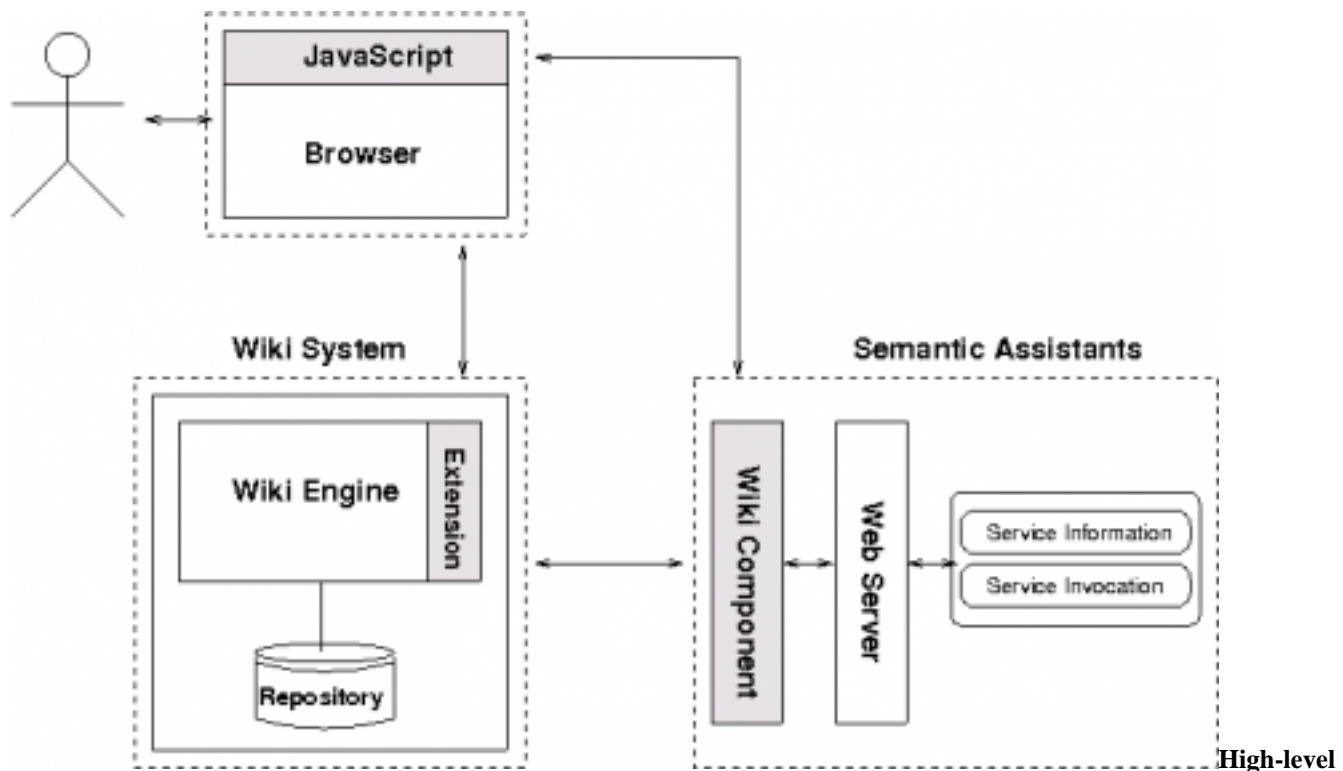
Wikis as new User Interfaces for Language Technology Experiments

Language technologies are rapidly entering modern applications. However, testing the real-world applicability of novel natural language processing algorithms on real-world tasks so far has required a large amount of software engineering work. Our solution

is based on a separation of concerns, where new NLP pipelines can be developed and easily deployed by a language engineer, without having to worry about front-end coding or web service invocations. Any new pipeline developed in GATE (which also integrates solutions from UIMA, OpenNLP, LingPipe, among others) can be automatically brokered to any connected wiki system. This greatly facilitates extrinsic experiments, where the impact of offering NLP to end users is measured on concrete tasks.

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 intermediary between the user's browser, the wiki engine and the [Semantic Assistants](#) framework [\[4\]](#) – our open source project that allows you to broker NLP pipelines as context-sensitive web services or *assistants*.



Design of the Semantic Assistants Wiki-NLP Integration

The Wiki Component shown above 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.

Features

The central goal of our Wiki-NLP integration is to provide a general architecture for enhancing various wiki systems with NLP techniques. Our solution is to offer a dynamically-generated user interface that is injected into the user's browser using JavaScript libraries. Such an approach allows the architecture's user interface to be portable among the multitude of existing wiki systems, providing a seamless integration of NLP capabilities within the wiki by creating the impression that the user is still interacting with the wiki's native interface.

The screenshot shows a MediaWiki page titled "Test Page" in a Mozilla Firefox browser. The page content includes a paragraph about Catherine's family and a table titled "Person and Location Extractor on Test_Page". The table lists entities like Catherine, Cilia, Rev. Thomas Davis, William, Leeds, Yorkshire, Amman, and Jordan with their types (Person or Location), start/end positions, and features (gender, locType). A sidebar on the left contains navigation links like "Main page", "Recent changes", and "Help". A bottom section titled "Available Assistants" shows a list of services including "Person and Location Extractor". A "Collection" field contains the URL "http://localhost/mediawiki-1.16/index.php/Test_Page".

MediaWiki Native Interface

NLP Results Embedded in Wiki Page

Wiki-NLP Integration Interface

The Wiki-NLP integration user interface embedded in a wiki page

Light-weight MediaWiki Extension

The Wiki/NLP integration is introduced to an existing MediaWiki engine through installing a light-weight extension. Without requiring major modifications on the wiki engine, the extension adds a link to the wiki toolbox menu through which users can load the Wiki/NLP interface. Using this interface, users can then inquire about and invoke NLP services through the dynamically generated Wiki/NLP interface within the wiki environment. Therefore, no context switching is needed by the wiki users in order to use the NLP services.

NLP Pipeline Independent Architecture

The Wiki/NLP integration is backed by the Semantic Assistants server, which provides a service-oriented solution to offer NLP capabilities in a wiki system. Therefore, any NLP service available in a given Semantic Assistants server can be invoked through the Wiki/NLP integration on a wiki's content.

Flexible Wiki Input Handling

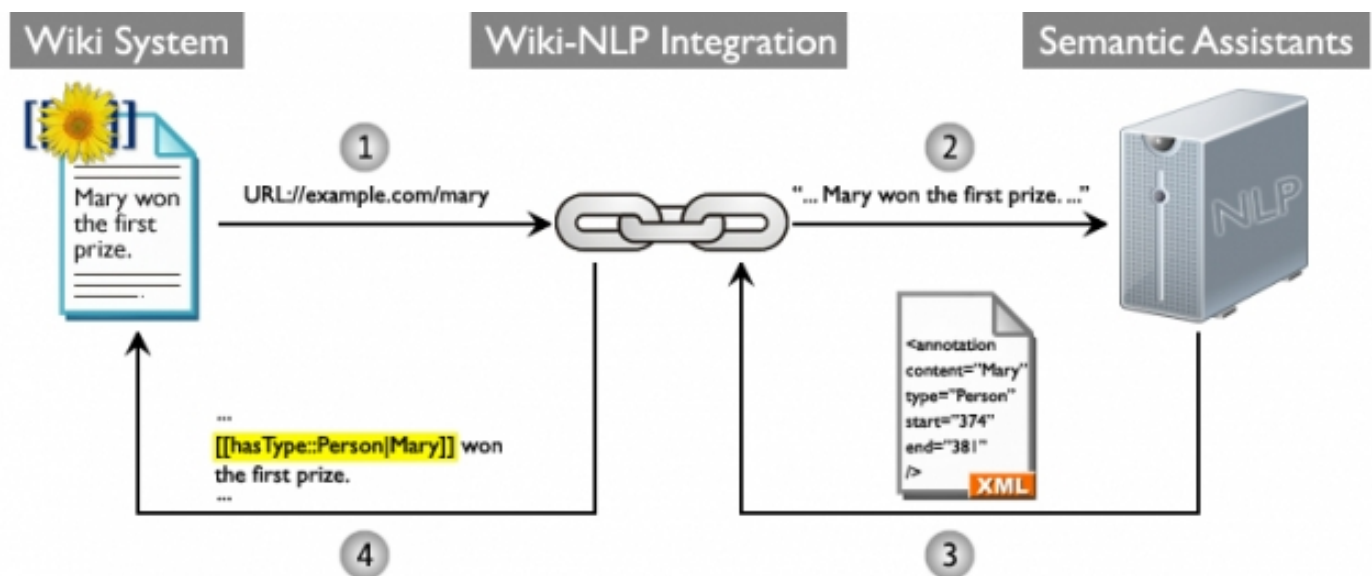
At times, a user's information need is scattered across multiple pages in the wiki. To address this problem, our Wiki/NLP integration allows wiki users to collect one or multiple pages of the wiki in a so-called "collection" and run an NLP service on the collected pages at once. This feature allows batch-processing of wiki pages, as well as gathering multiple input pages for pipelines analyzing multi-documents.

Flexible NLP Result Handling

The Wiki/NLP integration is also flexible in terms of where the NLP pipelines' output can be written. Upon a user's request, the pipeline results can be appended to an existing page body or its associated discussion page, create a new page, as well as writing to a wiki page in an external wiki, provided that it is supported by the Wiki/NLP integration architecture. Based on the type of results generated by the NLP pipeline, e.g., annotations or new files, the Wiki/NLP integration offers a simple template-based visualization capability that can be easily customized. Upon each successful NLP service execution, the Wiki/NLP integration automatically updates the existing results on the specified wiki page, where applicable.

Semantic Markup Generation

Where semantic metadata is generated by an NLP pipeline, the Wiki/NLP integration takes care of representing it in a formal language using the Semantic MediaWiki special markup. For generated metadata, the Wiki/NLP integration enriches the text with its equivalent markup and makes it permanent in the wiki repository. Therefore, for each generated result, both a user-friendly and machine-processable representation of the result is made available in the page. These markups are, in turn, transformed to RDF triples by the Semantic MediaWiki parsing engine, making them available for querying purposes as well as externalization to other applications.



Semantic enrichment of wiki text through the Wiki-NLP integration

For example, when the sentence *"Mary won the first prize."* is contained in a wiki page and processed by a Named Entity Detection pipeline, an XML document is generated by the Semantic Assistants server and returned back to the Wiki/NLP integration, which indicates "Mary" as an entity of type "Person". This XML document is then processed by our integration and transformed for Semantic MediaWiki into a formal representation in the form of