IntelliGenWiki: An Intelligent Semantic Wiki for Life Sciences

Bahar Sateli Marie-Jean Meurs Greg Butler
Justin Powlowski Adrian Tsang René Witte

Concordia University, Montréal, QC, Canada

Semantic Software Lab

NETTAB 2012 Nov. 15th, Como, Italy
Outline

1. Introduction
2. System Architecture
3. User Interface
4. Application
5. Evaluation
6. Conclusion
**Motivation:** Curation of Biomedical Literature

- Finding and extracting relevant knowledge from the domain literature
- Manually refining and updating bioinformatics databases

- Manual literature curation is
  - Expensive → requires domain experts
  - Labour-intensive → ever growing amount of scientific publications
  - Error-prone → critical knowledge can be easily missed

---

IntelliGenWiki: An Intelligent Semantic Wiki for Life Sciences  (Bahar Sateli et al.) 2 / 13
**Approach:** IntelliGenWiki

Enhanced Literature Curation Workflow Using IntelliGenWiki

- Text mining techniques integrated within the wiki environment
- Novel Human-AI collaboration patterns
- Producing semantic metadata
- Transform text into knowledge base
**Approach:** IntelliGenWiki

- Adopts the “Wiki” paradigm
  - Accessible via a web browser
  - Simple syntax (markup)
  - Open collaboration
- Based on the MediaWiki engine
  - Open source
  - Highly scalable
  - Extensible: Semantic MediaWiki
- Integrated Text Mining Assistants
- Provides semantic capabilities
  - Formalization of knowledge
  - Producing machine-readable content
- Open source software (AGPL3)
System Architecture

Introduction
System Architecture
User Interface
Application
Evaluation
Conclusion

System Overview

- **Front-end:** Semantic MediaWiki
- **Back-end:** Wiki-NLP Integration [Sateli and Witte, 2012]
  - Comprehensive architecture based on the Semantic Assistants Framework [Witte and Gitzinger, 2008]
  - Seamless integration of various NLP capabilities *within* a wiki environment

![System Architecture Diagram]

IntelliGenWiki: An Intelligent Semantic Wiki for Life Sciences (Bahar Sateli et al.) 5 / 13
IntelliGenWiki Pages

- Each wiki page corresponds to a literature instance, e.g., abstract of a paper
- Revision History
- Inquire text mining services via wiki toolbox
The NLP Interface

- The IntelliGenWiki NLP user interface offers various text mining services
- Customizing services at runtime
- Dynamically-generated interface

Text Mining Assistants inside the wiki

Step 1. Select the service your wish to execute on your collection.
Once you add this page to your collection, you can continue browsing as your collection is saved.

**Available Assistants**
- Select a service
  - mycolINE
  - IR Information Extractor
  - Information Extractor
  - OrganismTagger

**Runtime Parameters**

**Collection**

This page was last modified on 6 November 2012, at 23:21. This page has been accessed 4 times. Privacy policy. About
NLP Interface features

- Multi-document Analysis

![Image of NLP Interface features](image)

1. Select the service you wish to execute on your collection.
2. Once you add this page to your collection, you can continue browsing as your collection is saved.

- Flexible handling of results
  - Writing to the same page as the resource
  - Writing to a different page in the wiki
  - Writing to an external wiki

- Dynamic discovery of NLP services
Information Extraction

- Automatically extracting knowledge from text
- Various IE services
  - mycoMINE
  - OrganismTagger
  - Open Mutation Miner
  - ...
- Enrichment of literature content with semantic markup

Example:

[[hasType::Enzyme|cellobiohydrolase]]
Applications

Semantic Entity Retrieval

- Unadorned wikis offer only keyword-based search
- What if we want to discover what’s contained in the wiki?
  - e.g., “Which papers in this wiki mention an enzyme entity in their text?”
- Solution: Querying the semantic metadata in the wiki
  - Search the wiki by semantic properties, e.g., entity type, generated by NLP services
  - Using special Semantic MediaWiki markup, called inline queries

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

<table>
<thead>
<tr>
<th>Property: Enzyme</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Page Name</strong></td>
</tr>
<tr>
<td>PMID 20709852</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Enzyme Entities Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellbiohydrodrolase</td>
</tr>
<tr>
<td>Cellulases</td>
</tr>
<tr>
<td>Endoglucanases</td>
</tr>
<tr>
<td>β-glucosidases</td>
</tr>
<tr>
<td>Invrtogen</td>
</tr>
<tr>
<td>DNA polymerase</td>
</tr>
</tbody>
</table>
User Study

- Is the integration of text mining assistants in a wiki environment actually effective?
- User study within the Genozymes project context (www.fungalgenomics.ca)
  - **Goal**: Identifying and characterizing fungal enzymes
  - **Dataset**: 30 documents
  - **Users**: 2 expert biocurators
  - **NLP Service**: mycoMINE [Meurs et al, 2012]
  - **Measure**: Time spent on curation
  - **Method**: Comparison against time spent on manual curation

### Average Curation Time

<table>
<thead>
<tr>
<th></th>
<th>Abstract Selection</th>
<th>Full Paper Curation</th>
</tr>
</thead>
<tbody>
<tr>
<td>no support</td>
<td>IntelliGenWiki</td>
<td>no support</td>
</tr>
<tr>
<td>1 min.</td>
<td>0.3 min.</td>
<td>37.5 min.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.6 min.</td>
</tr>
</tbody>
</table>

- **Conclusion**: IntelliGenWiki was indeed efficient and reduced the paper selection and curation time by almost 70% and 20%, respectively.
Conclusion

What you can do now

- Install MediaWiki and Semantic MediaWiki extension
- Download and deploy the Wiki-NLP integration
- Use the existing text mining services in our public server
- Alternatively, setup your own Semantic Assistants services developed based on the GATE framework

What is next

- Cover other tasks, e.g.,
  - Quality assessment
  - Paper recommendation
  - Personalization
- Develop services for automatic import of literature, e.g., from PubMed
- Query the RDF in wiki from external applications
More Information

http://www.semanticsoftware.info/intelligenwiki

Acknowledgment

- Funding for this work was provided by NSERC, Genome Canada and Génome Québec.

- Caitlin Murphy and Sherry Wu, biocurators at the Centre for Structural and Functional Genomics (CSFG) at Concordia University, are acknowledged for their participation in the evaluation task.