**Text Mining Assistants in Wikis for Biocuration**

Bahar Sateli, Caitlin Murphy, René Witte, Marie-Jean Meurs, and Adrian Tsang

1. Semantic Software Lab. & 2. Centre For Structural and Functional Genomics

[ Email addresses: m.j.meurs@encs.concordia.ca, c.murphy, a.tsang@gene.concordia.ca, r.witte@cse.concordia.ca ]

---

**Literature Curation**
- Manually refining and updating bioinformatics databases
- Labor-intensive, error-prone and expensive task

**mycoMINE**
- Natural Language Processing (NLP) techniques supporting Biofuel Research

**GenWiki**
- An NLP-enhanced wiki system based on the MediaWiki engine
- Seamlessly integrates NLP capabilities in a wiki environment
- Allows curators to discover knowledge embodied in the wiki content

**mycoMINE [1]:** a text mining pipeline based on GATE [2].

**Semantic Infrastructure**
- NLP capabilities provided by the Semantic Assistants Framework [3]
- Brokers NLP pipelines as Web services
- Wiki-NLP communication through wiki plug-ins

**Application**
- Templating mechanism to present the NLP results to users
- Semantic metadata embedded in each page
- Using the Semantic MediaWiki (SMW) markup

**Semantic Entity Retrieval**
- Retrieve literature that contain certain type of entities
- Using SMW inline queries

---

**Evaluation**
- Methodology:
  - Corpus of 15 documents
  - Manually pre-filled GeneWiki with full-text papers
  - Used mycoMINE to automatically extract relevant entities
  - Keeping the track of time

**Average Curation Time: No Semantic Support vs. GenWiki**

<table>
<thead>
<tr>
<th></th>
<th>No Support</th>
<th>GenWiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 min.</td>
<td>30.6 min.</td>
<td>20 sec.</td>
</tr>
<tr>
<td>20 sec.</td>
<td>37.5 min.</td>
<td>67%</td>
</tr>
<tr>
<td>30.6 min.</td>
<td>30.6 min.</td>
<td>20%</td>
</tr>
</tbody>
</table>

**Conclusion**
GenWiki proves that seamless integration of NLP capabilities inside a wiki is indeed efficient, in terms of the time curators need to annotate and retrieve entities of interest from the domain literature.

---

**Acknowledgments**

Funding: Genome Canada & Génome Québec. Collaboration: Sherry Wu.