Toward Ontology-Based Knowledge Extraction from Biomedical Literature

Satoshi Kamegai
kamegai@jaist.ac.jp

Kenji Satou
ken@jaist.ac.jp

Akihiko Konagaya
kona@jaist.ac.jp

School of Knowledge Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan

Keywords: ontology extension, information extraction, template discovery

1 Introduction
Research on natural language processing has been active in recent years because it is seen as an important source of information for advancing genome analysis research. In this context, bio-ontology (e.g., TAO, MBO, Gene Ontology [2], Interaction ontology) that arranged the technical terms about a genome is being built. In addition, research (e.g., EMPathIE, PASTA) is also being conducted which aims at extracting relevant information, such as a protein-protein interaction, from biomedical literature.

However, since the former approach is built with the help of domain experts, it only covers tens of thousands of terms at the maximum. The latter one extracts only a small proportion of the useful information because it uses template matching on limited verbs showing interaction information.

We are extending an existing ontology based on the extensional ontology [1] built in previous work. We are trying to link Gene Ontology with the extensional ontology. Moreover, although discovery of verbs and templates showing interaction information is being found out with the help of domain experts, we try to extract mechanically them using the extended ontology.

2 Method and Results
Before trying ontological extension, two preliminary experiments were conducted.

2.1 First experiment

number of words which technical terms recognized by ontology have covered First of all, we searched both Gene Ontology and extensional ontology to investigate how many technical terms can be recognized in a text. The text part of Journal Cell (Table 1) and the text part of OMIM (Table 2) were used for the experiment. We investigated how many technical terms the ontology could recognize in a text. The coverage is calculated from the following equation.

\[
\text{coverage} = \frac{\text{number of words which technical terms recognized by ontology have covered}}{\text{total number of words in a text}} \times 100
\]

From this result, extensional ontology showed the very high coverage when compared to Gene Ontology. Therefore, we think that extending Gene Ontology with extensional ontology will improve the quantity.

Table 1: Recognition of technical terms in the text part of Cell.

<table>
<thead>
<tr>
<th>ontology</th>
<th>the number of recognized words</th>
<th>total number of words</th>
<th>coverage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Ontology</td>
<td>37340</td>
<td>937589</td>
<td>4.0</td>
</tr>
<tr>
<td>extensional ontology</td>
<td>324637</td>
<td>937589</td>
<td>34.6</td>
</tr>
</tbody>
</table>
Table 2: Recognition of technical terms in the text part of OMIM.

<table>
<thead>
<tr>
<th>ontology</th>
<th>the number of recognized words / total number of words</th>
<th>coverage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Ontology</td>
<td>159972 / 6261479</td>
<td>2.6</td>
</tr>
<tr>
<td>extensional ontology</td>
<td>2221119 / 6261479</td>
<td>35.5</td>
</tr>
</tbody>
</table>

2.2 Second Experiment
We investigated how much technical terms could be associated between Gene Ontology and extensional ontology. The following methods were considered for associating two terms in different ontologies.

Method 1. Exact match.

\[
\text{cell death} = \text{cell death}
\]

Method 2. Word level inclusion with preserving continuation and order.

\[
\text{blood coagulation factor} \supset \text{coagulation factor}
\]

Method 3. Word level inclusion without preserving continuation and order.

\[
\text{alcohol dehydrogenase nadp} \supset \text{nadp dehydrogenase}
\]

Table 3 shows the result of term association, where GO(term) and EO(term) indicate terms in Gene Ontology and extensional ontology, respectively. Since the word level inclusion in Method 2 and 3 is directed, two lines are used for each method. Table 4 shows the total number of associated terms by using method 1, 2 and 3. From this result, it can be said that most of the terms in Gene Ontology could be associated with terms in extensional ontology. However, extensional ontology showed that only 1/6 is associated by these methods.

Table 3: Result of term association.

<table>
<thead>
<tr>
<th>cases</th>
<th>associated terms in GO and EO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1 (GO(term) = EO(term))</td>
<td>GO,EO: 4120</td>
</tr>
<tr>
<td>Method 2 (GO(term1) \supset EO(term2))</td>
<td>GO:10267, EO:5554</td>
</tr>
<tr>
<td>Method 2 (GO(term1) \subset EO(term2))</td>
<td>GO:4301, EO:310705</td>
</tr>
<tr>
<td>Method 3 (GO(term1) \supset EO(term2))</td>
<td>GO:3728, EO:3530</td>
</tr>
<tr>
<td>Method 3 (GO(term1) \subset EO(term2))</td>
<td>GO:2882, EO:48790</td>
</tr>
</tbody>
</table>

Table 4: Total number of associated terms.

<table>
<thead>
<tr>
<th>Ontology</th>
<th>associated terms / all terms in the ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Ontology</td>
<td>11378 / 11494</td>
</tr>
<tr>
<td>extensional ontology</td>
<td>330097 / 2138347</td>
</tr>
</tbody>
</table>

3 Conclusion and Future Work
Based on these results, now we are trying to construct an extended Gene Ontology. In poster presentation, we will show some results about template discovery for knowledge extraction, by using it.

Acknowledgment
This work was supported in part by Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan and BIRD of Japan Science and Technology Corporation.

References