Annotating Scientific Papers for Mathematical Formula Search

Giovanni Yoko Kristianto¹, Goran Topic², Minh-Quoc Nghiem³, Akiko Aizawa¹,²

¹Department of Computer Science, The University of Tokyo, Tokyo, Japan
²National Institute of Informatics, Tokyo, Japan
³Department of Informatics, The Graduate University for Advanced Studies, Tokyo, Japan

Goal

Development of a knowledge base providing relationships between mathematical formulas and corresponding descriptions.

Automatic Extraction

Methods:
- pattern matching: consists of seven predefined sentence patterns
- CRF:
  - noun phrase as description candidate
  - syntactic features: sentence patterns, relative position of descriptions toward expressions, POS tags
  - lexical features: word unigrams, bigrams, and trigrams

<table>
<thead>
<tr>
<th>Method</th>
<th>Strict Matching</th>
<th></th>
<th>Soft Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1-score</td>
</tr>
<tr>
<td>pattern</td>
<td>25.53</td>
<td>20.84</td>
<td>22.91</td>
</tr>
<tr>
<td>CRF</td>
<td>73.60</td>
<td>30.09</td>
<td>42.46</td>
</tr>
</tbody>
</table>

Using the Annotation Result

Semantic Search
- Input: natural language description
- Output: related mathematical formula

<table>
<thead>
<tr>
<th>Input</th>
<th>Formulas</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;entropy&quot;</td>
<td>$H(T</td>
<td>L) = - \sum_{t \in T} P(t</td>
</tr>
<tr>
<td></td>
<td>$\bar{H}<em>T(L) = \sum</em>{L \in T} P(L) H(T</td>
<td>L)$</td>
</tr>
</tbody>
</table>

Semantic Browsing
- Input: mathematical formula
- Output: descriptions of the formula, including explanation of variables and subexpressions

Conclusion

- Our annotation design supports the annotation of continuous, discontinuous, and complex descriptions
- Annotation results can be used as a training data in automatic description extraction. The subsequently extracted data can be used for semantic searching and semantic browsing of mathematical expressions

Evaluation

- strict matching: extracted descriptions must be exactly the same as annotation result
- soft matching: accepting extracted descriptions that contain, are contained, or overlap with annotation result

Example:

- mathematical expression
- description set
- continuous part
- discontinuous part
- short version
- condition part

Example:
cite: arXiv:0806.4135v1 [math.GR]