Creation of Technical Trend Map

- Extraction of viewpoints
  - Extracting of expressions for viewpoints from each research paper and patent
    - Patent 1
    - Paper m
- Analysis
  - Putting similar viewpoints together
    - Viewpoint A
    - Viewpoint B
    - Similar Viewpoint
- Visualization
  - Classifying in tabular form
    - Viewpoint A
    - Viewpoint B
    - Similar Viewpoint

Experiments for NTCIR-8 Technical Trend Map Creation Subtask at Hitachi

Yusuke Sato & Makoto Iwayama

- Extraction of expressions of the effect of a research paper and patent as a viewpoint
Purpose

- Difficulty to learn a model for assignment of NTCIR-defined tags
  - Grammatically inconsistent definition of the tags
  - Tendency to assign tags to long phrases
- Definition of a 3-tuple syntactic structure for an effect expression
  - Assigning our independently defined tag set and then converting to NTCIR-defined tag set

Our independently defined tags

<TARGET>圧壊</TARGET><SCALE>強度</SCALE>の<IMPACT>高い</IMPACT>

NTCIR-defined tags

<EFFECT><ATTRIBUTE>圧壊強度</ATTRIBUTE>の<VALUE>高い</VALUE></EFFECT>

Conversion by several rules
Our Approach

- Our independently defined tag set

  重金属イオンの 回収 効率 を 向上 させる

  - <TARGET> : verb or noun, <SCALE> : scale, <IMPACT> : words modifying TARGET and SCALE elements

  - <EFFECT> : A region including <TARGET>, <SCALE> and <IMPACT>
  - <TARGET> : verb or noun which represents an action
  - <SCALE> : words such as “速度”, “工程” and so on
  - <IMPACT> : words such as “向上”, “低減” and so on

- Difference with NTCIR-defined tags

  1. More consistent grammatical elements
     - <TARGET> : verb or noun, <SCALE> : scale, <IMPACT> : words modifying TARGET and SCALE elements

  2. Division into more common elements or not
     - 回収効率 → specific to some technology fields
     - 回収 → specific, 効率 → common

     \[\text{Specific : difficult to assign} \quad \text{Common : easy to assign}\]
The flow of our tag assignment

- Assignment in the order of <IMPACT>, <SCALE> and <TARGET>
- Tag assignment
  - Our tags: Learning by SVM using independently developed training data
  - NTCIR tags: Conversion rules of our tags to NTCIR tags

Assignment of tags by each models

• Assignment of our defined tag set
• More accurate than assigning each tag independently

Assignment in the order of <IMPACT>, <SCALE> and <TARGET>

Independently defined tagged documents

Training Data

Generation of models

Documents

Assignment of <IMPACT>

Assignment of <SCALE>

Assignment of <TARGET>

Assignment of <EFFECT> and Conversion to NTCIR tag set
Features

【要約】

【発明の効果】
AAAAAAA・・・A。BBBBBBB・・・BB。CCCCCCC・・・高速な脱リン処理が可能となる。
DDDD・・・DD。

【符号の説明】

1. Morphemes by using ChaSen 脱／リン／処理／が／可能
2. SCALE/IMPACT dictionary ⇒ “高速”
3. SCALE/IMPACT-expression prefix/suffix single-kanji ⇒ ”高”、”速”
4. Morpheme of head in modifying modified segment ⇒ “高速”、”可能”
5. Results of IMPACT/SCALE assignment ⇒ ”高速”
6. Information indicating to be effect sentence
   i. End-of-sentence clue-phrase match ⇒ ”可能となる”
   ii. Paragraph type ⇒ Effect（”効果”）
   iii. Sentence position ⇒ 3 / 4 = 0.75
   iv. Sentence length
   v. Numeric character ratio within sentence
Assignment of EFFECT tag and Conversion of our tag into NTCIR tag

- **<EFFECT> identification**
  - <I>適正</I>な →
  - <S>温度</S>に →
  - <T>制御</T>する

- **Conversion rules**
  - E.g.) {<I><S>}</T> → <V><A>
  - <EFFECT><ATTRIBUTE>適正な温度</ATTRIBUTE>に
  - <VALUE>制御</VALUE></EFFECT>するナビゲーション装置
Independently developed training data

- Training data manually assigned our independently defined tag set

<table>
<thead>
<tr>
<th>Common Data</th>
<th>Data1</th>
<th>Data2</th>
<th>Data3</th>
<th>Data4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstracts in patent specifications</td>
<td>Abstracts in patent specifications</td>
<td>Abstracts in patent specifications</td>
<td>Abstracts in patent specifications</td>
<td>Abstracts in patent specifications</td>
</tr>
<tr>
<td>• Water-purifying technology (C02F 1/28) : 100</td>
<td>• Learning and classification technology (G06F 17/30) : 98</td>
<td>• Mixed data A : A61B : 10, B41J : 20, C08L : 10, D01F : 10, E02D : 10, F02D : 10, G06T : 20, H04N : 20</td>
<td>• Mixed data B : 50, G : 50, H : 50</td>
<td>• Mixed data B+ : 50, G : 200, H : 200</td>
</tr>
</tbody>
</table>

Data1 : Covering more technology fields
Data2 : Larger volume, but lower reliability for tag assignment
Data3 : For paper
Data4 : Higher reliability, but smaller volume
Experiments

- Features #1 - #3 are commonly used in all runs
- Learning and assignment by SVM (Linear kernel)
  - Giving “+1” if a morpheme is assigned any tag, otherwise “-1”
- No NTCIR-provided training data

<table>
<thead>
<tr>
<th>#</th>
<th>Type</th>
<th>ID</th>
<th>Training data (Our defined data)</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>Patent</td>
<td>HTC_1_1</td>
<td>Data1</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>ii</td>
<td>HTC_1_2</td>
<td>Data1</td>
<td>✓ ✓ ✓</td>
<td></td>
</tr>
<tr>
<td>iii</td>
<td>HTC_2_1</td>
<td>Data2</td>
<td>✓ ✓ ✓</td>
<td></td>
</tr>
<tr>
<td>iv</td>
<td>HTC_2_2</td>
<td>Data2</td>
<td>✓ ✓</td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>Paper</td>
<td>HTC_1</td>
<td>Data3</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>vi</td>
<td>HTC_2</td>
<td>Data4</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
**Results of NTCIR-defined tag set**

<table>
<thead>
<tr>
<th>ATTR.</th>
<th>Patent</th>
<th></th>
<th>Paper</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#i</td>
<td>#ii</td>
<td>#iii</td>
<td>#iv</td>
<td>#v</td>
<td>#vi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>25.1%</td>
<td>24.1%</td>
<td>24.7%</td>
<td>23.7%</td>
<td>14.9%</td>
<td>11.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>24.1%</td>
<td>23.6%</td>
<td>28.2%</td>
<td>27.3%</td>
<td>16.4%</td>
<td>11.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>24.6%</td>
<td>23.9%</td>
<td>26.3%</td>
<td>25.4%</td>
<td>15.6%</td>
<td>11.3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VALUE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>58.0%</td>
<td>57.2%</td>
<td>52.1%</td>
<td>50.8%</td>
<td>20.7%</td>
<td>23.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>43.4%</td>
<td>43.2%</td>
<td>46.2%</td>
<td>45.5%</td>
<td>21.0%</td>
<td>20.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>49.6%</td>
<td>49.2%</td>
<td>49.0%</td>
<td>48.0%</td>
<td>20.9%</td>
<td>22.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFFECT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>16.4%</td>
<td>15.5%</td>
<td>15.3%</td>
<td>14.5%</td>
<td>5.5%</td>
<td>5.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>22.3%</td>
<td>21.7%</td>
<td>23.6%</td>
<td>22.8%</td>
<td>11.2%</td>
<td>9.9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>18.9%</td>
<td>18.1%</td>
<td>18.6%</td>
<td>17.7%</td>
<td>7.3%</td>
<td>7.3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ave.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>23.3%</td>
<td>22.7%</td>
<td>21.5%</td>
<td>20.9%</td>
<td>10.0%</td>
<td>10.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>34.6%</td>
<td>34.4%</td>
<td>38.0%</td>
<td>37.3%</td>
<td>18.8%</td>
<td>16.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>27.8%</td>
<td>27.4%</td>
<td>27.5%</td>
<td>26.8%</td>
<td>13.1%</td>
<td>12.3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results of our independently defined tag set

<table>
<thead>
<tr>
<th></th>
<th>Patent (Data1)</th>
<th>Paper (Abstracts in 200 papers)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TARGET</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>45.0%</td>
<td>7.9%</td>
</tr>
<tr>
<td>P</td>
<td>58.7%</td>
<td>19.6%</td>
</tr>
<tr>
<td>F</td>
<td>50.9%</td>
<td>11.3%</td>
</tr>
<tr>
<td><strong>SCALE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>54.3%</td>
<td>19.5%</td>
</tr>
<tr>
<td>P</td>
<td>63.4%</td>
<td>33.8%</td>
</tr>
<tr>
<td>F</td>
<td>58.5%</td>
<td>24.7%</td>
</tr>
<tr>
<td><strong>IMPACT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>64.9%</td>
<td>28.0%</td>
</tr>
<tr>
<td>P</td>
<td>68.4%</td>
<td>38.4%</td>
</tr>
<tr>
<td>F</td>
<td>66.6%</td>
<td>32.4%</td>
</tr>
</tbody>
</table>
Discussion

- NTCIR defined tag set
  - The results of Data 1 has slightly higher F-value than those of Data 2
    - Need of higher reliability to tag set rather than a larger volume of data
  - Lower accuracy for papers than patents
    - End-of-sentence clue-phrase in effect sentence are NOT used frequently

- Our independently defined tag set
  - Accuracy of TARGET was low, for which there are relatively few words common to diverse technology fields
Conclusion

- Independent definition of syntactic structure of effect expressions
  - TARGET / SCALE / IMPACT
    - <EFFECT><TARGET>建築</TARGET><SCALE>コスト</SCALE>の<VALUE>低減</VALUE></EFFECT>
  - Assignment of our defined tags data by using SVM according to independently developed training
  - Conversion of our defined tag set to NTCIR defined tag set by eight rules based on dependency relations
  - ATTR. : 24.6%, VALUE : 49.6%, EFFECT : 18.9%
  - “Effect sentence” feature (#6) is very effective for patent data
  - Lower accuracy to assign to long phrases