BiTeM’s experience at NTCIR-8

The Patent Mining task in NTCIR-8

The Patent Mining task challenges the participants to do multi-lingual and cross-genre IPC classification of research paper abstracts and to detect technology and effect passages in patent and research paper abstracts in order to create a technical trend map.

1. Research Papers Classification
   - English, Japanese, E2J and J2E subtasks

2. Technical Trend Map Creation
   - Japanese and English subtasks

Methods and Data

Research Papers Classification

- **Re-ranking methods:**
  - sim: \( S = \sum_i w_i \cdot e_i \cdot d_i \)
  - freq: \( S = \sum_i w_i \cdot e_i \cdot d_i \)
  - weak: \( S = \sum_i w_i \cdot \log \left( \frac{d_i}{\text{freq}(i)} \right) \cdot \text{idf}(i) \)
  - combined: \( S = \alpha \cdot \text{freq}(i) + \beta \cdot \text{weak}(i) + \gamma \cdot \text{freq}(i) \cdot \text{idf}(i) \)
  - multi collection: \( S = \sum_i \alpha \cdot \text{freq}(i) + \beta \cdot \text{weak}(i) + \gamma \cdot \text{freq}(i) \cdot \text{idf}(i) \)

Classification System

- Code distribution for PAJ and USPTO corpora
  - codes PAJ USPTO
  - Paj 420 4718 30885
  - USPTO 428 6598 38491
  - average codes/doc PAJ USPTO
  - PAJ 1.5 1.9 2.3
  - USPTO 1 1 1
  - median PAJ USPTO
  - PAJ 3497 181 35
  - USPTO 706 14 5

Technical Trend Map Creation

- **CRF models:**
  - 'token' model with dictionary
  - 'all' model without dictionary

Results

Research Papers Classification

- **Precision x Recall results for the English subtask (topic and corpus in English)**
  - subclass: 0.87 0.67
  - main group: 0.87 0.67
  - subgroup: 0.60 0.60
  - main group: 0.60 0.60

Precision x Recall results for the J2E subtask (topics in Japanese and corpus in English)

- subclass: 0.17 0.17
  - main group: 0.17 0.17
  - subgroup: 0.17 0.17

Results of the different re-ranking methods using English topics (MAP)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>freq</th>
<th>sim</th>
<th>weak</th>
<th>combined</th>
<th>multi-col</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>subclass</td>
<td>0.67</td>
<td>0.67</td>
<td>0.66</td>
<td>0.67</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td>main group</td>
<td>0.48</td>
<td>0.48</td>
<td>0.47</td>
<td>0.48</td>
<td>1.00</td>
<td>0.84</td>
</tr>
<tr>
<td>subgroup</td>
<td>0.29</td>
<td>0.29</td>
<td>0.28</td>
<td>0.29</td>
<td>1.00</td>
<td>0.53</td>
</tr>
<tr>
<td>subclass</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>1.00</td>
<td>0.57</td>
</tr>
<tr>
<td>main group</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.40</td>
<td>1.00</td>
<td>0.39</td>
</tr>
<tr>
<td>subgroup</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>1.00</td>
<td>0.19</td>
</tr>
<tr>
<td>subclass</td>
<td>0.59</td>
<td>0.59</td>
<td>0.60</td>
<td>0.60</td>
<td>0.01</td>
<td>0.57</td>
</tr>
<tr>
<td>main group</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.40</td>
<td>1.00</td>
<td>0.39</td>
</tr>
<tr>
<td>subgroup</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.17</td>
<td>0.01</td>
<td>0.19</td>
</tr>
<tr>
<td>subclass</td>
<td>0.69</td>
<td>0.68</td>
<td>0.68</td>
<td>1.00</td>
<td>1.00</td>
<td>0.69</td>
</tr>
<tr>
<td>main group</td>
<td>0.50</td>
<td>0.50</td>
<td>0.48</td>
<td>1.00</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>subgroup</td>
<td>0.31</td>
<td>0.30</td>
<td>0.29</td>
<td>1.00</td>
<td>1.00</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Conclusions

- The re-ranking methods proposed have similar performance.
- Their combination does not improve the results significantly.
- The combination of collections improves the results.
- Single and multi-lingual classifications have similar performances.
- Use of built in dictionary improves the performance of the NER engine, especially when detecting effect value passages.
- Technology passages are easier to detect in title than in abstract.

References


Acknowledgements

This research has been supported by the EU-IST FP7 DebugIT project # 71239.