BiTeM’s Experience at NTCIR-8

Douglas Teodoro
University of Geneva

http://eagl.unige.ch/bitem
Overview

- The task and data used
- Research Paper Classification
  - Methods
  - Results
- Technical Trend Map Creation
  - Methods
  - Results
- Conclusions
Task

• Patent Mining
  – Research Paper Classification: classification of paper abstracts into IPC codes
    • Subtasks:
      – English: test collection and corpus in English
      – J2E: test collection in Japanese and corpus in English
  – Technical Trend Map Creation: named entity recognition in abstracts
    • Tags: technology, effect, attribute and value
    • Subtasks:
      – English
        » Paper
        » Patent
Training Data

• Corpus:
  – PAJ: Japanese patent abstracts translated into English
    • 3M documents (2.38 used)
    • No citation information
  – USTPO: “complete” patent documents from USPTO office
    • 1.3M documents (0.89 used)
    • Only main IPC code

• Paper Classification
  – 976 English paper abstracts

• Technical Trend Map
  – 300 paper abstracts
  – 300 patent abstracts
Research Paper Classification
Classification System

• Used Terrier for IR
  – BM25 model
• 3 different indexes:
  – PAJ, USPTO and USPTO_CLAIM
• kNN based
  – Different k values tuned depending on the classifier
• Re-ranking methods [T. Xiao 2008]:
  – sim: \( S_i = \sum S_{d_k} \) if \( c_i \in d_k \)
  – freq: \( S_i = \sum 1 \) if \( c_i \in d_k \)
  – weak: \( S_i = (S_{sim_i} \times S_{freq_i}) / df_i \)
  – combined: \( S_i = \alpha S_{sim_i} + \beta S_{freq_i} + \gamma S_{weak_i} \)
  – multi-collection:
    \( S_i = \alpha S_{PAJ_i} + \beta S_{USPTO_i} + \gamma S_{USPTO_CLAIM_i} \)
• Query translator approach in the multi-lingual task (Google Language Tools)
Classification System Architecture

IR Engine

Abstract

Terrier

Translator

PAJ

USPTO

k-NN

k-documents

Mapping (doc->code)

Re-ranking

Re-ranking algorithms

Ranked code list
Results – Official Runs

**English subtask**

<table>
<thead>
<tr>
<th></th>
<th>subclass</th>
<th>main group</th>
<th>subgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiTeM_sim</td>
<td>0.683</td>
<td>0.497</td>
<td>0.299</td>
</tr>
<tr>
<td>BiTeM_combined</td>
<td>0.666</td>
<td>0.480</td>
<td>0.286</td>
</tr>
<tr>
<td>BiTeM_weak</td>
<td>0.661</td>
<td>0.469</td>
<td>0.282</td>
</tr>
</tbody>
</table>

**J2E subtask**

<table>
<thead>
<tr>
<th></th>
<th>subclass</th>
<th>main group</th>
<th>subgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiTeM_sim</td>
<td>0.705</td>
<td>0.500</td>
<td>0.303</td>
</tr>
<tr>
<td>BiTeM_combined</td>
<td>0.687</td>
<td>0.484</td>
<td>0.293</td>
</tr>
<tr>
<td>BiTeM_weak</td>
<td>0.680</td>
<td>0.474</td>
<td>0.282</td>
</tr>
</tbody>
</table>
Results – All English Runs

• Re-ranking approaches

<table>
<thead>
<tr>
<th>Classifier</th>
<th>freq</th>
<th>sim</th>
<th>weak</th>
<th>combined</th>
<th>multi-coll parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>subclass</td>
<td>0.59</td>
<td>0.59</td>
<td>0.60</td>
<td>0.60</td>
<td>0;0;1</td>
</tr>
<tr>
<td>main group</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0;0;1</td>
</tr>
<tr>
<td>subgroup</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.17</td>
<td>0;0;1</td>
</tr>
<tr>
<td>subclass</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0;1;0</td>
</tr>
<tr>
<td>main group</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.40</td>
<td>0;1;0</td>
</tr>
<tr>
<td>subgroup</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0;1;0</td>
</tr>
<tr>
<td>subclass</td>
<td>0.67</td>
<td>0.67</td>
<td>0.66</td>
<td>0.67</td>
<td>1;0;0</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;0;0</td>
</tr>
<tr>
<td>subgroup</td>
<td>0.29</td>
<td>0.29</td>
<td>0.28</td>
<td>0.29</td>
<td>1;0;0</td>
</tr>
<tr>
<td>subclass</td>
<td>0.69</td>
<td>0.68</td>
<td>0.68</td>
<td>-</td>
<td>1;0;1;0.01</td>
</tr>
<tr>
<td>main group</td>
<td>0.50</td>
<td>0.50</td>
<td>0.48</td>
<td>-</td>
<td>1;0;1;0.01</td>
</tr>
<tr>
<td>subgroup</td>
<td>0.31</td>
<td>0.30</td>
<td>0.29</td>
<td>-</td>
<td>1;0;1;0.01</td>
</tr>
</tbody>
</table>

*Official runs

~5% better

{USPTOCLAIM, USPTO, PAJ, 3 indexes}
Technical Trend Map Creation
Technical Map System

• Use openNLP for pre-processing and Mallet for NER

• CRF based
  – Models:
    • token
    • token and part of speech
    • all

• Post-processing:
  – Rule-based
  – Dictionary
### Models

<table>
<thead>
<tr>
<th>Features</th>
<th>Models</th>
<th>token</th>
<th>token and ps</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>token</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>hasPreviousToken</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>hasNextToken</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>partOfSpeech</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>hasPreviousPS</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>hasNextPS</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>sentencePosition</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>isCapital</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>isAlphanumeric</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>isInCounterPart</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>paragraphSize</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>paragraphPosition</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>sentenceFeatures</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>sentenceLength</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>sentenceParenthesis</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>sentencePunctuation</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

### Rules
- isLinkingWord (‘and’, ‘or’, etc.)
- isEndTag (‘;’, ‘’, etc.)
- isOpenParenthesis
- isCloseParenthesis
- isFalseTech (‘methods’, ‘models’, etc.)
- hasRelevantWord (‘a’, ‘the’, etc.)

### Dictionary
- Technology:
  - Advanced digital modulation technique
  - Artificial Reality
  - Breathing Wall
  - etc.
- Attribute:
  - animation
  - anti-abrasion
  - anti-cracking property
  - etc.
- Value:
  - absorbing
  - accurate
  - accurately
  - achieved
  - etc.
Technical Map System Architecture

Pre-processing

Abstract

Document segmentation → Sentence segmentation → Tokenisation → Part of speech tagging → Named Entity Recogniser

CRF Engine

NER Booster

Rules

Dictionary

Annotated abstracts
Results

Passage detection on research paper abstracts

- BiTeM_1 -> ‘all’ model with dictionary
- BiTeM_2 -> ‘token’ model with dictionary
- BiTeM_3 -> ‘token and part of speech model’ with dictionary
- BiTeM_4 -> ‘all’ model without dictionary

• Simpler models perform better
• Effect dependent on attribute and value

June 17, 2010
Technical Trend Map Application

• Trend detection in technological field
  – Important to have timestamp when we talk about trends
• Ontology generation
• Technology fusion
Automatic Ontology Generation

- Extracted directly from the technology/effect tags
- It depends on sentence’s voice:
  - Active: attribute->object, aux verb + value->predicate
  - Passive: attribute->object, value->predicate
Technology Fusion

- Based on co-occurrence
- Ranked according to term frequency
Conclusions

- Multi-lingual classification has the same performance as monolingual.
- The re-ranking methods proposed have similar performance and their combination does not improve the results significantly.
- The combination of collections improves the results.
- NER in paper and patent documents show roughly the same performance in our system*.
- 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.

*Other groups achieved much better performance in patent documents