BiTeM's experience at NTCIR-8

Hôpitaux Universitaires de Genève

h e g
Haute école de gestion de Genève
Geneva School of Business Administration

Douglas Teodoro[1], Julien Gobeill[2], Emilie Pasche[1] Dina Vishnyakova[1], Patrick Ruch[2] and Christian Lovis[1]
Bibliomics and Text Mining (BiTeM) group
[1] University and Hospitals of Geneva, Switzerland
[2] University of Applied Sciences Geneva, Switzerland

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

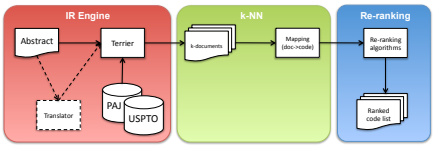
Methods and Data

Research Papers Classification

❑Re-ranking methods:

- sim: $S_i = \sum S_{d_k}$ if $c_i \in d_k$
- freq: $S_i = \sum 1$ if $c_i \in d_k$
- weak: $S_i = (S_{sim\_i} \times S_{freq\_i}) \ / \ idf_i$
- combined: $S_i = \alpha S_{sim\_i} + \beta S_{freq\_i} + \gamma S_{weak\_i}$
- multi-collection:
  $S_i = \alpha S_{PAJ\_i} + \beta S_{USPTO\_i} + \gamma S_{USPTO\_CLAIM\_i}$

Classification System



Code distribution for PAJ and USPTO corpora

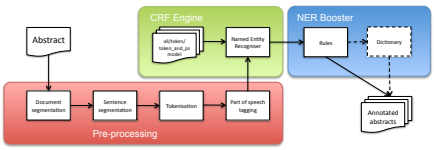
	corpus	subclass	maingroup	subgroup
codes	PAJ USPTO	420 428	4738 6588	30885 38491
average codes/doc	PAJ USPTO	1.5 1	1.9 1	2.3 1
median docs/code	PAJ USPTO	3497 706	181 14	35 5

Technical Trend Map Creation

❑CRF models:

- token: 3 features
- token and part of speech: 6 features
- all: 16 features

Technical Map Creation System



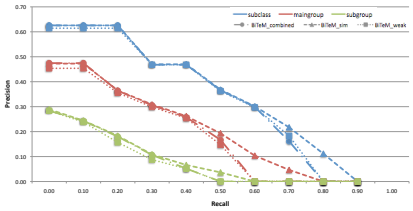
Annotated abstract example

<TECHNOLOGY>a processor LSI</TECHNOLOGY> and <TECHNOLOGY>a DRAM</TECHNOLOGY> divided into banks, for <EFFECT> <VALUE> increasing</VALUE><ATTRIBUTE>a ratio of using a fast operation mode </ATTRIBUTE></EFFECT>

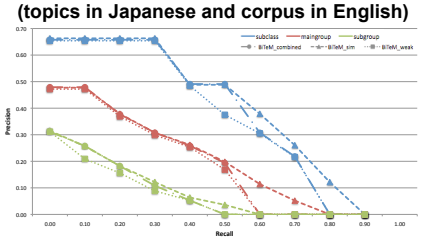
Results

Research Papers Classification

Precision x Recall results for the English subtask (topic and corpus in English)



Precision x Recall results for the J2N subtask (topics in Japanese and corpus in English)

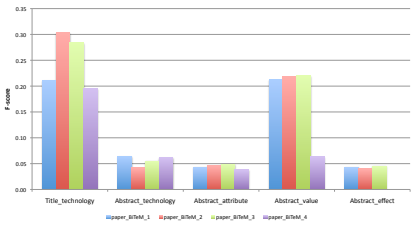


Results of the different re-ranking methods using English topics (MAP)

Classifier	freq	sim	weak	combined	multi-coll parameters
subclass	0.67	0.67	0.66	0.67	1;0;0
main group	0.48	0.48	0.47	0.48	1;0;0
subgroup	0.29	0.29	0.28	0.29	1;0;0
subclass	0.60	0.60	0.60	0.60	0;1;0
main group	0.39	0.39	0.39	0.40	0;1;0
subgroup	0.17	0.17	0.17	0.17	0;1;0
subclass	0.59	0.59	0.60	0.60	0;0;1
main group	0.39	0.39	0.39	0.39	0;0;1
subgroup	0.16	0.16	0.16	0.17	0;0;1
subclass	0.69	0.68	0.68	-	1;0.1;0.01
main group	0.50	0.50	0.48	-	1;0.1;0.01
subgroup	0.31	0.30	0.29	-	1;0.1;0.01

Technical Trend Map Creation

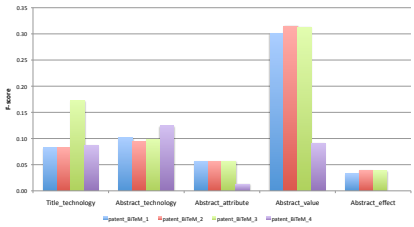
Results of the English subtask using research paper abstracts



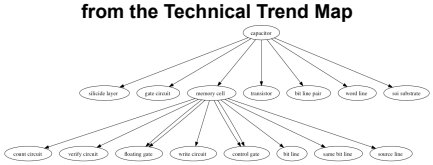
BiTeM_1 -> 'all' model with dictionary
BiTeM_2 -> 'token' model with dictionary
BiTeM_3 -> 'token and part of speech model' with dictionary
BiTeM_4 -> 'all' model without dictionary

BiTeM's website: http://eagl.unige.ch/BiTeM
IPCCat: http://pingu.unige.ch:8080/IPCCat

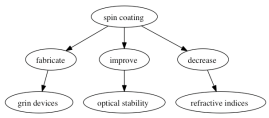
Results of the English subtask using patent abstracts



Example of technological trend extracted from the Technical Trend Map



Example of ontology extracted from the Technical Trend Map



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

F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1-47, 2002.

T. Xiao, F. Cao, T. Li, G. Song, K. Zhou, J. Zhu, and H. Wang. Knn and re-ranking models for english patent mining at ntcir-7. *In Proceedings of NTCIR-7 Workshop Meeting*, 2008.

Y. G. Kim, J. H. Suh, and S. C. Park. Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, 34:1804-1812, 2008.

Acknowlegements

This research has been supported by the EU-IST-FP7 DebugIT project # 712139.