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

	corpus	sub-class	main group	sub-group
codes	PAJ USPTO	420 428	4738 6588	30885 38491
avg codes/ doc	PAJ USPTO	1.5 1	1.9 1	2.3 1
max docs/ code	PAJ USPTO	180552 50103	123062 17880	24364 5026
min docs/ code	PAJ USPTO	13 1	1 1	1 1
avg docs/ code	PAJ USPTO	5673 5602	503 135	77 23
median docs/code	PAJ USPTO	3497 706	181 14	35 5

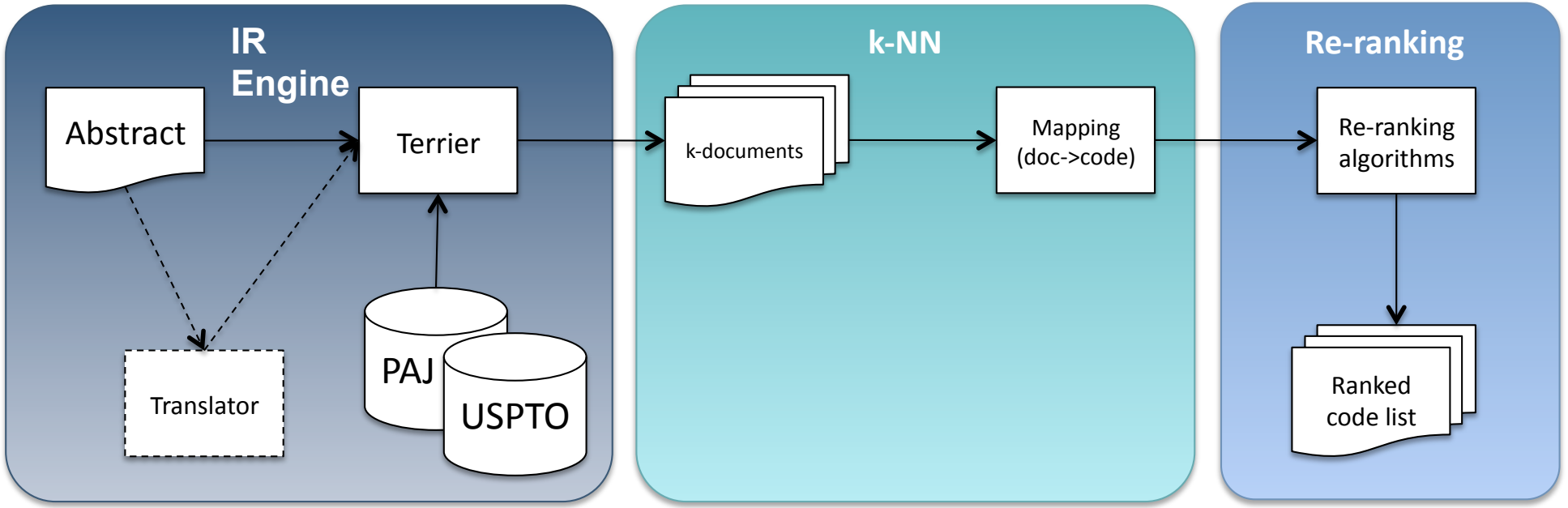
Research Paper Classification

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: $S_i = \sum S_{d_k} \text{ if } c_i \in d_k$
  - freq: $S_i = \sum 1 \text{ if } c_i \in d_k$
  - 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:

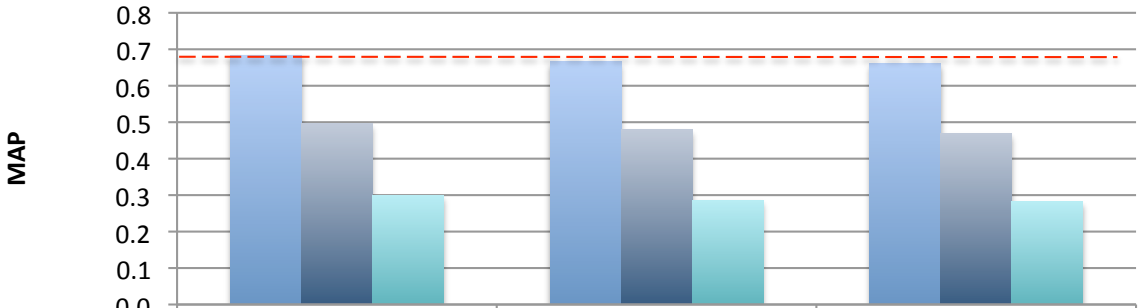
    $S_i = \alpha S_{PAJ\_i} + \beta S_{USPTO\_i} + \gamma S_{USPTO\_CLAIM\_i}$
- Query translator approach in the multi-lingual task (Google Language Tools)

Classification System Architecture



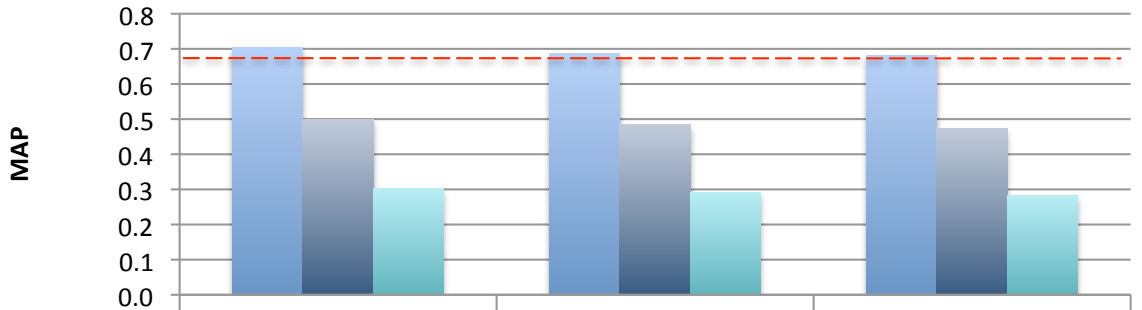
Results – Official Runs

English subtask



	BiTeM_sim	BiTeM_combined	BiTeM_weak
subclass	0.683	0.666	0.661
main group	0.497	0.480	0.469
subgroup	0.299	0.286	0.282

J2E subtask



	BiTeM_sim	BiTeM_combined	BiTeM_weak
subclass	0.705	0.687	0.680
main group	0.500	0.484	0.474
subgroup	0.303	0.293	0.282

Results – All English Runs

- Re-ranking approaches

Classifier	freq	sim	weak	combined	multi-coll parameters	
subclass	0.59	0.59	0.60	0.60	0;0;1	USPTO_CLAIM
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	0.60	0.60	0.60	0.60	0;1;0	USPTO
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	0.67	0.67	0.66*	0.67*	1;0;0	PAJ
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	0.69	0.68*	0.68	-	1;0.1;0.01	3 indexes
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	

~5% better

*Official runs

Technical Trend Map Creation

Technical Map System

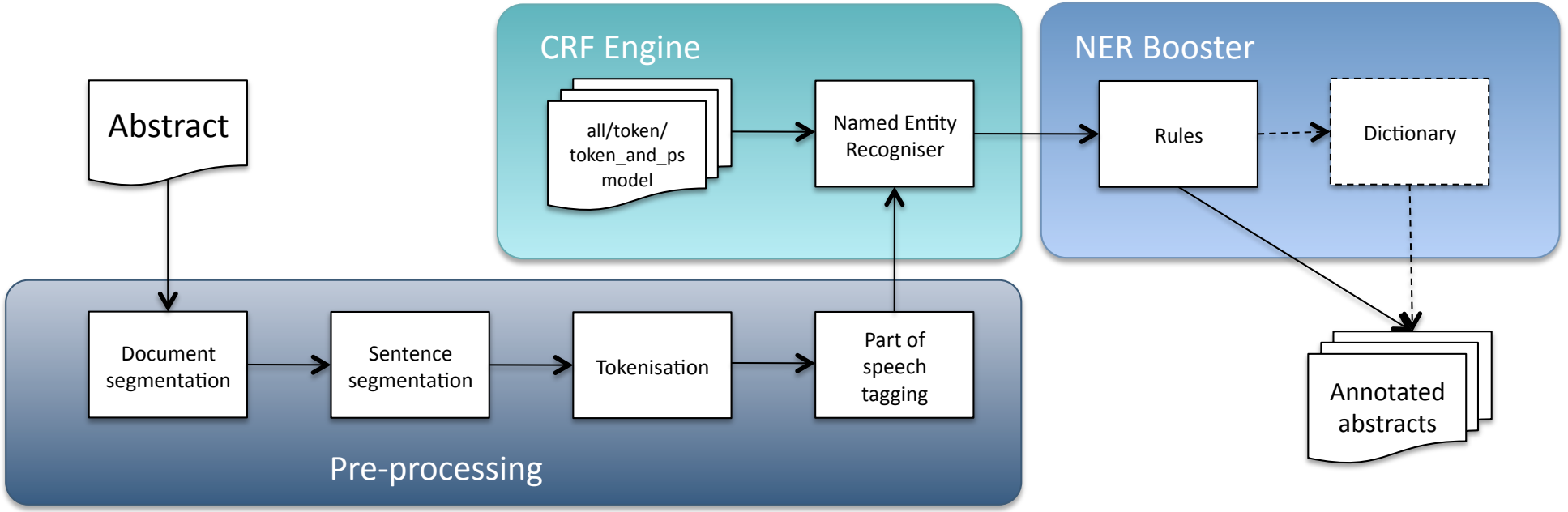
- Use openNLP for pre-processing and Mallet for NER
- CRF based
  - Models:
    - token
    - token and part of speech
    - all
- Post-processing:
  - Rule-based
  - Dictionary

- Models

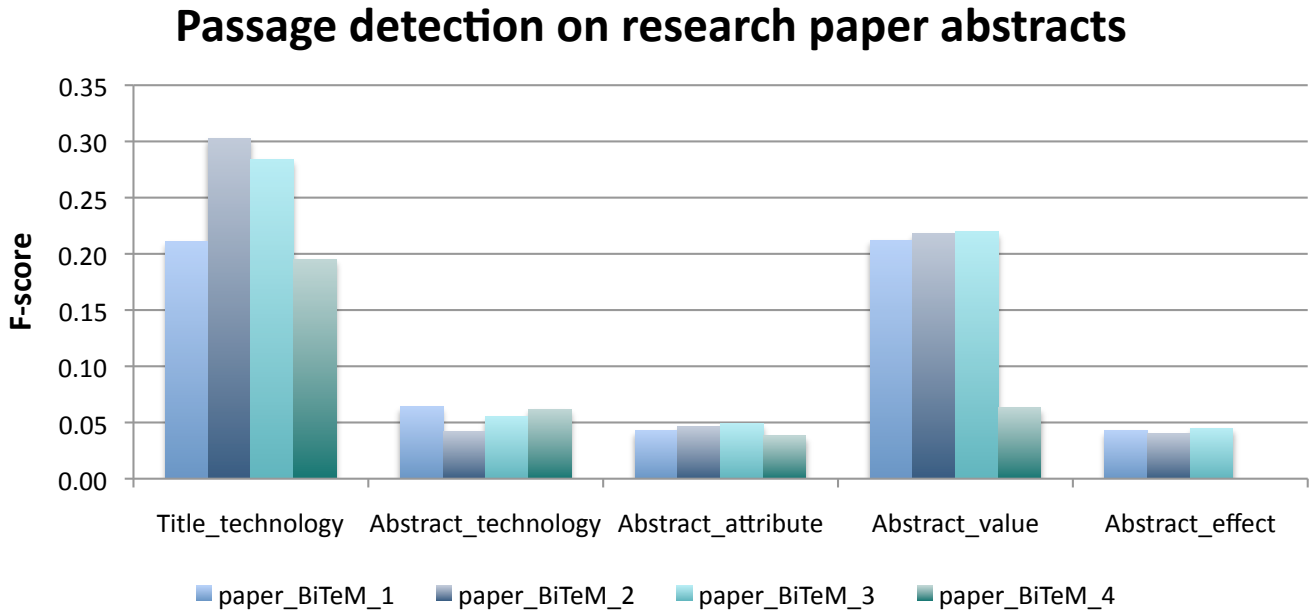
Features \ Models	token	token and ps	all
token	x	x	x
hasPreviousToken	x	x	x
hasNextToken	x	x	x
partOfSpeech	-	x	x
hasPreviousPS	-	x	x
hasNextPS	-	x	x
sentencePosition	-	-	x
isCapital	-	-	x
isAlphanumeric	-	-	x
isInCounterPart	-	-	x
paragraphSize	-	-	x
paragraphPosition	-	-	x
sentenceFeatures	-	-	x
sentenceLength	-	-	x
sentenceParenthesis	-	-	x
sentencePunctuation	-	-	x

- 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.

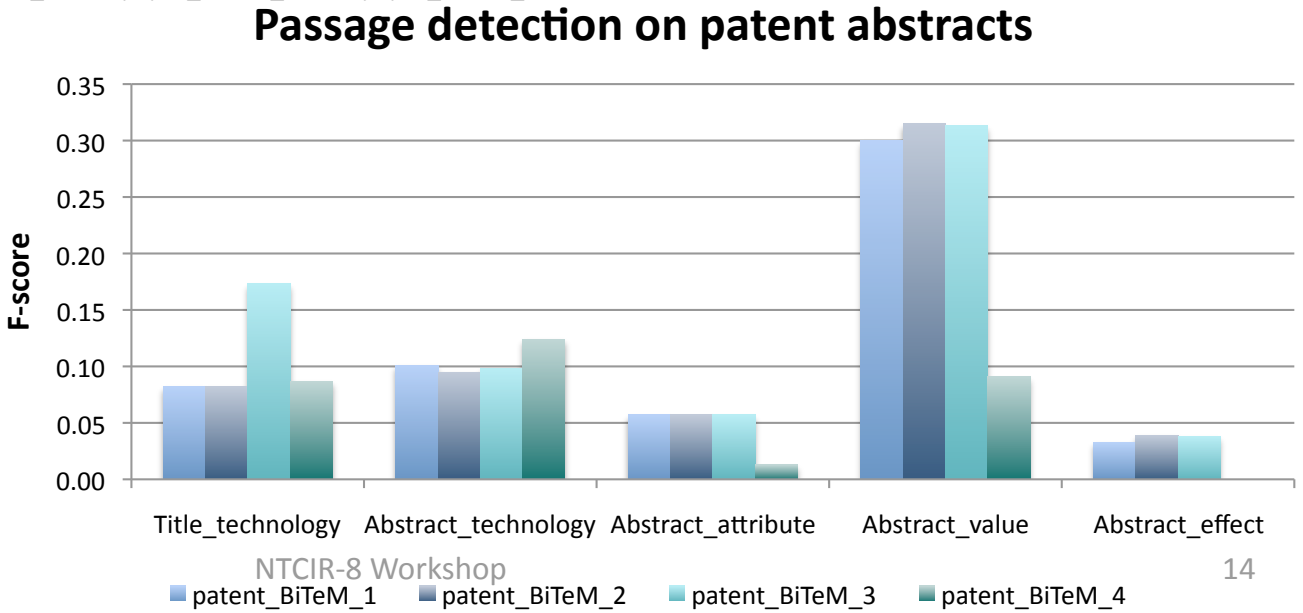
Technical Map System Architecture



Results



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



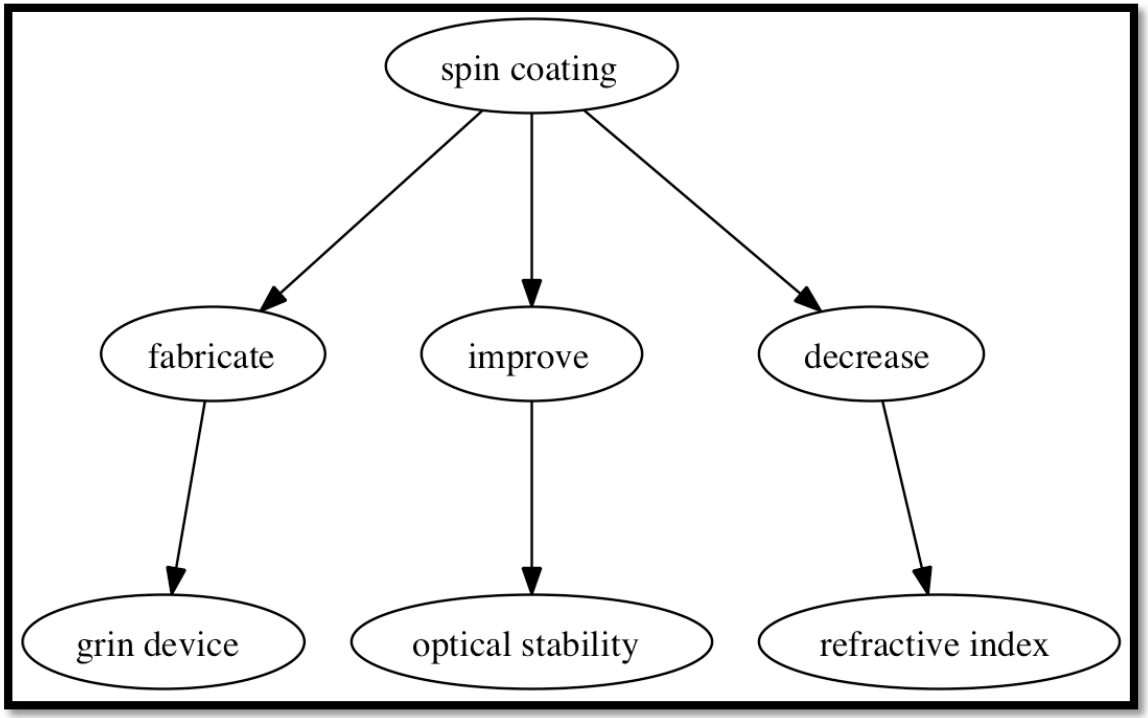
- Simpler models performs better
- Effect dependent on attribute and value

Technical Trend Map Application

- Trend detection in technological field
  - Important to have timestamp when we talk about trends
- Ontology generation
- Technology fusion

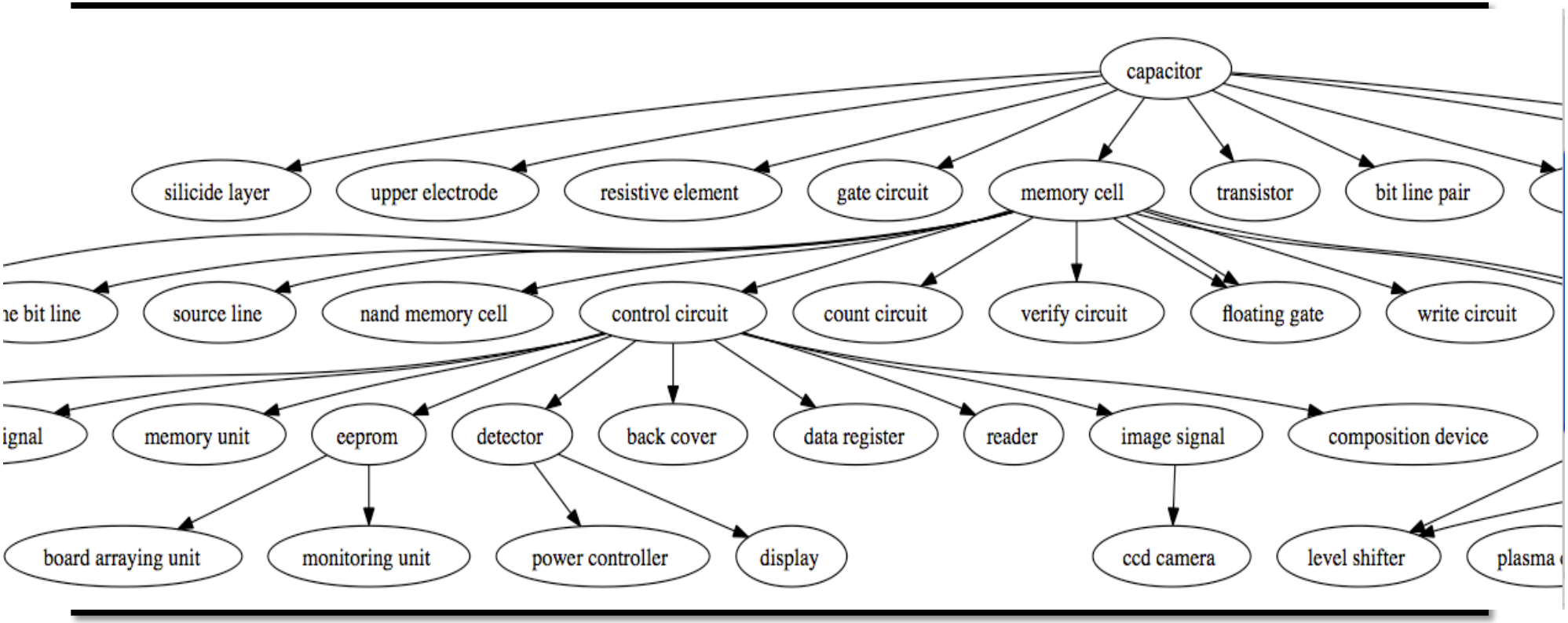
Automatic Ontology Generation

- 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