NTCIR Patent:
A Test Collection for
Patent Retrieval / Classification

Makoto Iwayama (Hitachi / TIT)
Atsushi Fujii (Tsukuba Univ.)
Noriko Kando (NII)
Akihiko Takano (NII)
Outline

• Background
• Patent Documents and Patent Retrieval
• Overview of Patent Retrieval Task at NTCIR
• Introduction of each Task
  – Retrieval Task
  – Classification Task
• Summary and Future Directions
Background

• Large test collections for human language technology have been produced in TREC, CLEF, NTCIR
  – TREC (mainly for English)
    CLEF (for European languages)
    NTCIR (for east Asian languages)
  – Retrieval (mono/cross-lingual), QA, Summarization, etc.

• Targets are newspaper, technical paper (abstract), Web
Background (cont.)

• Commercial patent retrieval systems have operated for a long time
• But, less attention in IR and NLP research communities
• Needs for constructing a test collection for patents (at 2000 SIGIR Workshop on Patent Retrieval)
Characteristics of Patent Documents

• Structured documents
  – claim, purpose, effect, embodiment, etc.
• Many technical terms and new terms
  – Those are often defined by applicants
• General terms preferred in claims
  – “floppy disk” → “external storage”
• Unusual style for claims
  – legal aspect of patents
• Large variation of document length
  – The longest patent has about 30,000 unique terms!
• Having classifications
  – IPC: International Patent Classification
  – local classifications for EPO (ECLA), USPTO(USPC), JPO (FI, F-term)
Characteristics of Patent Retrieval

• Various search purposes
  – High recall is necessary in many situations
  – Trend analysis (making patent map) is important for deciding the patent strategy

• Large difference between technological fields
  – Chemical formulas/materials are important in chemistry
  – Images/diagrams are important in machinery
  – Searching by categories (ex. IPC) is effective in some fields, but not in others (ex. business method patents)
Overview of Patent Retrieval Task at NTCIR
Brief History of Patent Retrieval Tasks at NTCIR

- **2001**: Retrieval - Technology survey
- **2002**: Retrieval - Invalidity search; more practical
- **2003**: Classification - Patent map generation; well-defined
- **2004**: Classification - Patent classification; more search topics
- **2005**: Classification - more difficult
- **2006**: Classification - more documents

NTCIR-3, NTCIR-4, NTCIR-5, NTCIR-6
How to Construct Test Collection
(Basic Idea)

Participants

search topic

system1

system2

search target (doc. collection)

search results1

search results2

pooled results

relevance assessment

evaluation (recall/precision)

manual search

relevant docs.

assessors (human experts)

relevant docs.
Patent Corpora Released

- Japanese patents (full texts) published in 1993-2002 (3.5M patents)
- English abstracts (translated by experts) of the above patents
- List of categories (themes and F-terms) for the patents published in 1993-1997
- Other training data provided by PATOLIS Co. and JPO (Japan Patent Office)
Test Collections Constructed

• NTCIR-3
  – 31 search topics in Japanese, English, Korean and Chinese (traditional and simplified)
  – Relevant patents (about 75 / topic) judged by experts

• NTCIR-4
  – 34 search topics in Japanese, English, Korean and Chinese (traditional and simplified)
  – Relevant patents (about 10 / topic) judged by experts
  – 67 rejected patents and their citations used for the rejection
  – 6 patent map examples in Japanese

• NTCIR-5
  – 1189 rejected patents and their citations used for the rejection
Retrieval Task
NTCIR-3 (Technology Survey)

- Non-expert (manager): reading news articles
  - Clipping + memo for supplementing, focusing, etc.
- Expert (patent searcher): patent survey before product development
  - Patents as technical documents (cf. as legal documents)
  - Cross-genre retrieval
Cont.

• 31 topics created / assessed by patent experts (JIPA: Japan IP Association). They also performed manual search before the relevance assessment.

• Japanese, English, Korean, Chinese topics were created.

• 36 results were submitted by 8 participated groups (from Japan, USA, Sweden).

• Recall-Precision based evaluation (MAP).
Findings

• Cross-genre retrieval was difficult due to term mismatch
  – “president” in news articles vs. patents
• Cross-lingual retrieval was more difficult than mono-lingual retrieval
• Participating systems could find many relevant documents \( \frac{667}{2311} = 30\% \) that human experts could not find
NTCIR-4 (Invalidity Search)

- examiner in patent office
- searcher in IP division

inventor (applicant)

application

claim 1

claim 2

claim 3

relevant document

prior art

participating system

patent document collection

document collection

invalidate the demand in claim 1
34 topics created / assessed by patent experts (JIPA: Japan IP Association). They also performed manual search before the relevance assessment.

- Japanese, English, Korean, Chinese topics were created.
- 110 results were submitted by 8 participated groups (only from Japan).
- Recall-Precision based evaluation (MAP).
- Passage retrieval was also evaluated (patents are long!).
Findings

- Interesting approaches specific to patent retrieval were proposed
  - Term weighting considering the term position in claim ("X comprising of Y which …; Z which …")
  - Query expansion using the corresponding descriptive part to a claim
  - Two stage retrieval (recall-oriented → precision-oriented)
- Participating systems could find many relevant documents (129 / 344 = 38%) that human experts could not find
NTCIR-5 (Invalidity Search)

• Same as NTCIR-4 Invalidity Search Task
• We automatically created a large number of topics (1,189), those were extracted from rejected patents
• Cited patents used for the rejection were regarded as relevant patents
• No human assessment was done
• English topics were created
• 84 results were submitted by 8 participated groups (from Japan, Korea)
• Recall-Precision based evaluation (MAP)
• Passage retrieval was evaluated
Findings

- We had similar results to those in NTCIR-4
- Search task was easy if
  - query patent and cited patent were filed by the same applicant
  - query patent and cited patent has the same IPC
NTCIR-6 (Invalidity Search)

• We will continue invalidity search task
• We will focus on difficult topics
• US patents will be used
Classification Task
NTCIR-4 (Patent Map Generation)

Topic: optical disk

Target patents:

<table>
<thead>
<tr>
<th>problems</th>
<th>high density</th>
<th>erasing</th>
<th>rewriting</th>
</tr>
</thead>
<tbody>
<tr>
<td>managing the number of rewriting</td>
<td></td>
<td></td>
<td>1993-000003</td>
</tr>
<tr>
<td>Shifting the writing position</td>
<td></td>
<td></td>
<td>1994-000002</td>
</tr>
<tr>
<td>Laser power pulse waveform</td>
<td>1994-000008</td>
<td>1996-000005</td>
<td></td>
</tr>
</tbody>
</table>

• Target patents and viewpoints (“problems” and “solutions”) are given.
• Participants determine categories (“high density”, “erasing”, …) under each viewpoint (“problems” or “solutions”), and classify the target patents.
Review

• Hard task
  – Cluster labeling was especially difficult
• Subjective evaluation only (by JIPA)

• Simple and well-defined task
• Objective evaluation by many topics
NTCIR-5 (Classification)

- Classify patents into F-terms (File forming terms)
- F-terms are multi-faceted categories for patents
- F-terms are used in JPO
- Every patent has F-terms

- 5 years’ patents with F-terms were the training documents
- About 2,000 patents were the test documents
- 49 results were submitted by 5 participated groups (from Japan, Korea)
F-term (File Forming Term)

5B001
Detection and correction of errors

AB
PURPOSE

AC
MEAN

AE
ERROR LOCATION

01
Code operations

04
Comparison

05
Interleaving

theme
2,577

viewpoint
(facet)

element
(category)

F-term
18 ~ 986
(ave.190)

NTCIR-4 → NTCIR-5 → NTCIR-6
Findings

• K-NN was generally effective
• Document surrogates were used in almost all submissions (abstract, claim, prior art, problems to be solved, method, etc)
• Calculating component-by-component similarities was effective in K-NN (abstract vs. abstract + claim vs. claim + …)
NTCIR-6 (Classification)

• We will continue F-term classification task with more test documents
• We will release annotated patents with grounds of categorization
Summary

• NTCIR patent is a project to create test collections for patent processing to enhance the technology of patent processing
• Our test collection is one of the largest test collections for patents
Future Directions

• Automatic patent map generation
• Searching both patents and technical papers
• Citation analysis
URL

or
search by “NTCIR”