Proposal of Patent Mining Task at NTCIR-8

2009.6.24

Hidetsugu Nanba (Hiroshima City University)
Atsushi Fujii (University of Tsukuba)
Makoto Iwayama (Hitachi, Ltd.)
Taiichi Hashimoto (Tokyo Institute of Technology)

1

Technical Trend Analysis from Research Papers and Patents

	Effect 1	Effect 2	Effect 3
Technology 1	[AAA 1993] [US Pat. XX-XXX]		[BBB 2002]
Technology 2			
Technology 3	[CCC 2000]	[US Pat. YY-YYY]	[US Pat. ZZ-ZZZ] [US Pat. WW-W]

2

Creation of Technical Trend Maps

- **(Step 1) For a given field, research papers and patents written in various languages are collected.**

  Subtask 1: Research Papers Classification

- **(Step 2) Elemental technologies and their effects are extracted from the documents collected in Step 1.**

  Subtask 2: Technical Trend Map Creation

3

Example of an Abstract

<TOPIC>
<TOPIC-ID> 100 **</TOPIC-ID>**
<TITLE> DTMF (Dual Tone Multi-Frequency) transmission method for a mobile communication system **</TITLE>**
<ABSTRACT> High efficient speech encoding scheme called VSELP, is adopted for Japanese digital mobile communication systems. However, DTMP (Dual Tone Multi-Frequency) signals are distorted by using this encoding scheme. This paper presents a DTMF signal transmission scheme. DTMF signals are transmitted in the form of call control messages from mobile station (MS) to mobile control center (MCC). In addition, necessary control capabilities in MS and MCC is described. **</ABSTRACT>**
</TOPIC>

4

An example of a data for the Subtask of Technical Trend Map Creation

[Japanese]

PM磁束制御用コイルを設けて<Technology>閉ループフィードバック制御</Technology>を施すため、<Effect><Attribution>電力損失</Attribution>を<Value>最小化</Value></Effect>できる。

[English]

Through <Technology> closed-loop feedback control </Technology>, the system could <Effect><Value> minimize </Value> the <Attribution> power loss </Attribution> </Effect>.

5

Challenges at PAT-MN

Different terms between research papers and patents

"floppy disc" - "removable recording medium"
"floppy disc" - "magnetic recording device"
"word processor" - "document editing device"

Cross-genre Cross-lingual Information Access !

6

Technical Trend Analysis from Research Papers and Patents

	Effect 1	Effect 2	Effect 3
Technology 1	[AAA 1993] [US Pat. XX-XXX]		[BBB 2002]
Technology 2			
Technology 3	[CCC 2000]	[US Pat. YY-YYY]	[US Pat. ZZ-ZZZ] [US Pat. WW-W]

7

Subtask 1: Research Papers Classification

Each system is requested to classify research papers written in either Japanese or English in terms of the International Patent Classification (IPC) system.

- **Japanese: classification of Japanese research papers using patent data written in Japanese.**
- **English: classification of English research papers using patent data written in English.**
- **Cross-lingual subtask: classification of Japanese (English) research papers using patent data written in English (Japanese).**

8

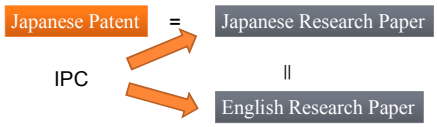
The Article 30 in Japanese Patent Law

- **The article 30 in Japanese patent law provides a six-month grace period for disclosures made through a publication or a presentation at a conference or an exhibition.**
- **The applicants need to mention the proceeding titles and the date it was published in "Indication of exceptions to lack of novelty" field in the patent.**

[Indication of exceptions to lack of novelty] The provisions set forth in Article 30, Paragraph 1 in Japanese patent law. Proceedings (Volume 4) of the 60th Annual Meeting of the Information Processing Society of Japan, published in March 14, 2000.

9

Assignment of IPC Codes into Research Papers



10

Comment from Participants in NTCIR-7 PAT-MN (1)

Isn't it possible to use more data for evaluation? The topics used in the formal run and the dry run were not much enough in terms of variation of the IPCs. Though there were about 50,000 IPC codes in the pseudo-training data, most of them were not contained in the data for evaluation.

11

Comment from Participants in NTCIR-7 PAT-MN (1)

- **(participant) If we regard that an IPC code of a patent is the same as that of a cited paper in the patent, large number of topics can be prepared.**
- **(organizers) However, we might not get full text data corresponding to the cited paper. Even if we could obtain full text data by negotiating with particular associations, such as ACM or IEEE, these data are biased to particular IPC codes.**
- **(participant) Instead of using full text data, how about using only bibliographic information, such as title, authors' names, and conference names? Some participant groups used only titles in NTCIR-7 PAT-MN.**
- **(organizers) As we have developed a system that extracts citation information in patents, it is possible to prepare such data.**

12

Related Papers

Extraction of cited papers from Japanese patents

Nanba, H., Anzen, N., and Okumura, M. (2008) "Automatic Extraction of Citation Information in Japanese Patent Applications" International Journal on Digital Libraries, Vol.9, No.2, 151–161.

Extraction of cited papers from US patents

Koguri, Y. and Nanba, H. (2007) "Automatic Extraction of Citation Information in US patents" Proceedings of the 13th Annual Meeting of the Association for Natural Language Processing, 582–585. (in Japanese)

13

Topics

(1) TITLE
(2) TITLE+AUTHORS
(3) TITLE+AUTHORS+CONFERENCE
(4) TITLE+AUTHORS+CONFERENCE+KEYWORDS
(5) TITLE+ABST (NTCIR-7 PAT-MN)
(6) TITLE+AUTHORS+ABST
(7) TITLE+AUTHORS+CONFERENCE+ABST
(8) TITLE+AUTHORS+CONFERENCE+KEYWORDS+ABST

(1)–(3) Instead of using the NTCIR-1, 2 corpora, use bibliographic information) of cited papers or the database of office action.

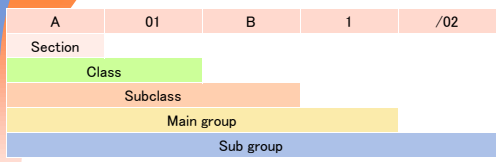
14

Comment from Participants in NTCIR-7 PAT-MN (2)

When a system output is "A01B 1/02", and a correct IPC is "A01B 1/01", the current evaluation method (MAP) does not give any score, though the system output and the correct IPC match at main group level ("A01B 1"). In addition to the current evaluation method, how about employing another method that gives additional partial score, if a system output and a correct IPC match at higher level, such as section, class, subclass, and main group?

15

Evaluation



A	Section	Human necessities
A01	Class	Agriculture; Forestry; Hunting; Fishing; …
A01B	Subclass	Soil working in agriculture or forestry; parts, details or accessories of agricultural machines or implements, in general
A01B 1/00	Main group	Hand tools
A01B 1/02	Sub group	Spades; Shovels

16

Subtask 2: Technical Trend Maps Creation

Each system is requested to extract expressions of element technologies and their effects from research papers and patents.

17

An example of a data for the Subtask of Technical Trend Map Creation

[Japanese]

PM磁束制御用コイルを設けて<Technology>閉ループフィードバック制御</Technology>を施すため、<Effect><Attribution>電力損失</Attribution>を<Value>最小化</Value></Effect>できる。

[English]

Through <Technology> closed-loop feedback control </Technology>, the system could <Effect><Value> minimize </Value> the <Attribution> power loss </Attribution> </Effect>.

18

Technical Trend Analysis from Research Papers and Patents

	Effect 1	Effect 2	Effect 3
Technology 1	[AAA 1993] [US Pat. XX-XXX]		[BBB 2002]
Technology 2			
Technology 3	[CCC 2000]	[US Pat. YY-YYY]	[US Pat. ZZ-ZZZ] [US Pat. WW-W]

19

Technical Trend Visualization

[English]
<Technology>CRF-based approach</Technology> obtained a
<Effect><Attribute>precision</Attribute> of <Value>0.935</Value></Effect>.

Numerical expression

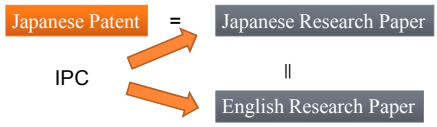
Subtasks

- Japanese: extraction of technologies and their effects from Japanese documents.
- English: extraction of technologies and their effects from English documents.

21

対象文書

- 日英論文抄録(ntcir-1, 2)
- 日本語特許



22

Tag Definition

TOPIC
GOAL
TECHNOLOGY
SUBTITLE (Title only)
EFFECT
　-ATTRIBUTION
　-VALUE
　-CONDITION

Example of CONDITION (Jp)

<TOPIC>
<TOPIC-ID>202</TOPIC-ID>
<TITLE><TOPIC>LSI化回線品質検出回路</TOPIC>の構成と特性</TITLE>
<ABSTRACT>TDMA装置の一層の小形・経済化および高速化を図るために<TECHNOLOGY>1μmゲートアレイ</TECHNOLOGY>を用い開発した第2世代汎用高速・高機能TDMA LSIのうち<HEAD>回路品質検出LSI(OLM LSI:On Line Monitor LSI)</HEAD>の機能と構成について述べる。また、本LSIの<TECHNOLOGY>再符号化法</TECHNOLOGY>を用いた場合の<EFFECT><ATTRIBUTION>誤り検出の精度</ATTRIBUTION>を求め、<CONDITION>サンプル数が10^3個以上</CONDITION>ならば<VALUE>高精度な検出</VALUE></EFFECT>を行えることを示した。また、本LSIを誤検出率低減ユニークワード検出回路に適用した場合の特性も明らかにした。</ABSTRACT>
</TOPIC>

24

Example of CONDITION (En)

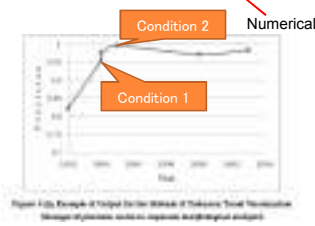
<TOPIC>
<TOPIC-ID>117</TOPIC-ID>
<TITLE><TOPIC>AVserver"KABUKI"</TOPIC>:<SUBTITLE>Distributed Multimedia Environment</SUBTITLE></TOPIC>
<ABSTRACT>This paper presents <TOPIC>the AVserver"KABUKI"</TOPIC>.It is developed to <EFFECT>provide clients with an integrated environment to handle multimedia data</EFFECT> on <CONDITION>Unix workstation</CONDITION>. It offers the functions of inter-media synchronization,data transfer,data compression, decompression and resource management.Requirement,architecture model and implementation of the system are described. Also,the results of performance measurements in a network environment and the evaluation are discussed.</ABSTRACT>
</TOPIC>

25

Usage of CONDITION Tag

[English]
<Technology>CRF-based approach</Technology> obtained a
<Effect><Attribute>precision</Attribute> of <Value>0.935</Value></Effect>.

Condition 2
Numerical expression
Condition 1



Usage of CONDITION Tag

	Effect 1	Effect 2	Effect 3
Technology 1	[AAA 1993] [US Pat. XX-XXX]		[BBB 2002]
Technology 2			
Technology 3	[CCC 2000]	[US Pat. YY-YYY]	[US Pat. ZZ-ZZZ] [US Pat. WW-W]

Condition 2
Condition 1

27

Example of GOAL Tag

<TITLE><TOPIC>DTMF(Dual Tone Multi-Frequency)transmission method</HEAD> for <GOAL>a mobile communication system</GOAL></TOPIC>

<TITLE>衛星画像による<GOAL>植生識別</GOAL>のための<TOPIC>大気・地形の影響の除去</TOPIC>について</TITLE>

28

タグ付けが困難な例(1)

- 本手法を用いることによって,相互接続試験における試験系列生成のコストの削減,試験の信頼性の向上などが期待される.

ATTRIBUTION 試験系列生成のコスト
VALUE 削減
ATTRIBUTION 試験の信頼性
VALUE 向上

「～が可能となる。」→OK
「～が期待される。」→？

タグ付けが困難な例(2)

今回の実験では1）精度と2）再現率についてβ手法と比較する。結果、1）では3.5%向上し、2）では下がった。

ATTRIBUTIONとVALUEが入れ子構造になっているため、EFFECTタグが付与できない。

タグ付けが困難な例(3)

- MIL状態で、パーフルオロアルキル基をもつポリフェニルアセチレンはP_{O_2}が10gtを超えかつαが2.7以上と優れた酸素選択透過性を有していることが分かった。

CONDITION MIL状態
ATTRIBUTION P_{O_2}
VALUE 10gtを超え
ATTRIBUTION α
VALUE 2.7以上
ATTRIBUTION 酸素選択透過性
VALUE 優れた

タグ付けが困難な例(4)

- 従来手法では、ABCDには耐電性が付加できなかった。…以上のことにより、本研究で従来手法が改善された。

EFFECT「ABCDには耐電性が付加」?

既存の研究の問題点を指摘。

その他の問題

日本語特許データへのタグ付け付与

- 全文を対象
- 一部の項目のみ対象
  (「要約」と「発明の効果」)
- PAJ

33

Document Sets

Data	Year	Size	Number	Language
(1) Unexamined Japanese patent applications	1993–2002	100 GB	3.50M	Japanese
(2) USPTO patent data	1993–2000	33 GB	0.99M	English
(3) Patent Abstracts of Japan (translated into English)	1993–2002	4.2 GB	3.50M	English
(4) NTCIR-1 and NTCIR-2 CLIR Task test collection (Abstracts of research papers)	1988–1999	1.4 GB	0.26M	Japanese/ English

34