# NTCIR-7速報

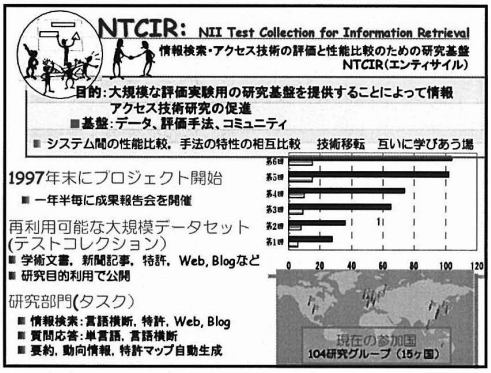
## Breaking News from NTCIR-7

http://research.nii.ac.jp/ntcir/index-ja.html

酒井哲也 (ニューズウォッチ)
加藤恒昭 (東京大学)
藤井敦 (筑波大学)
難波英嗣 (広島市立大学)
関洋平（豊橋技術科学大学）
三田村照子(CMU)
神門典子 (NII)

## TALK OUTLINE

1. NTCIRってなに
2. ACLIA速報
3. MOAT速報
4. PATMT/PATMN速報
5. MuST速報
6. EVIA速報
7. つづきはNTCIR-7 meeting (12/16-19)で…




NTCIR

## 情報検索　Information Retrieval (IR)

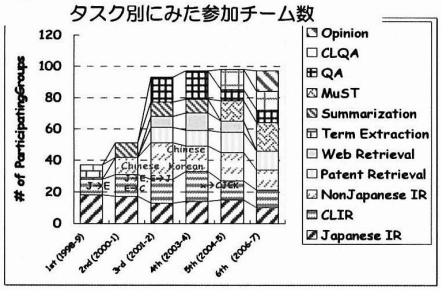
- *蓄積された大量の文書から利用者の情報要求にレレバントな(適合する）情報を選びだす*
  - *構造化されていない、自然言語テキスト*
  - *伝統的には文書検索-→文書中の情報活用支援へ*
- *コンピュータの使用　　１９５０年代から*
- *人間の主観的判定を成否の基準にした最初の計算機科学*

## *情報アクセス　Information Access (IA)*

- *蓄積された大量の文書中の情報を利用者が利用できるようにするまでの全過程*
- *検索、要約、質問応答、テキストマイニング、クラスタリングなど*

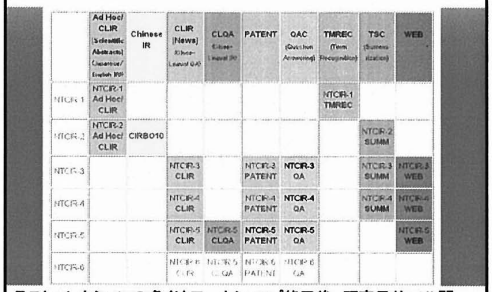
NTCIR

## タスク別にみた参加チーム数




NTCIR

## NTCIR テストコレクション




テストコレクションの多くはワークショップ終了後、研究目的で公開

NTCIR

## 文書データの例:

<DOC>
<DOCNO>ctg_xxx_19990110_0001</DOCNO>
<LANG>EN</LANG>
<HEADLINE> Asia Urged to Move Faster in Shoring Up Shaky Banks </HEADLINE>
<DATE>1999-01-10</DATE>
<TEXT>
<P>HONG KONG, Jan 10 (AFP) - Bank for International Settlements (BIS) general manager Andrew Crockett has urged Asian economies to move faster in reforming their shaky banking sectors, reports said Sunday. Speaking ahead of Monday's meeting at the BIS office here of international central bankers including US Federal Reserve chairman Alan Greenspan, Crockett said he was encouraged by regional banking reforms but "there is still some way to go." Asian banks shake off their burden of bad debt if they were to be able to finance recovery in the crisis-hit region, he said according to the Sunday Morning Post. Crockett added that more stable currency exchange rates and lower interest rates had paved the way for recovery. "Therefore I believe in the financial area, the crisis has in a sense been contained and that now it is possible to look forward to real economic recovery," he was quoted as saying by the Sunday Hong Kong Standard.</P>
<P>"It would not surprise me, given the interest I know certain governors have, if the subject of hedge funds was discussed during the meeting," Crockett said. </P>
<P>He reiterated comments by BIS officials here that the central bankers would stay tight-lipped about their meeting, the first to be held at the Hong Kong office of the Swiss-based institution since it opened last July. </P>
</TEXT>
</DOC>

NTCIR

## 検索課題・質問の例

検索語　検索質問文

<TOPIC>
<NUM>0008</NUM>
<TITLE CASE="b">サルサ, 学ぶ, 方法</TITLE>
<DESC>サルサを踊れるようになる方法が知りたい</DESC>
<NARR><BACK>最近はやっているサルサという踊りを学ぶためにどうすればよいか具体的な方法が知りたい。例えば教室に通うという場合には、その場所や授業形態など、具体的な内容を必要とする。</BACK><RELE>具体的な方法の表記のない、流行であることのみを扱った文書は不適合とする。</RELE></NARR>
<CONC>サルサ, 習う, 方法, 場所, カリキュラム</CONC>
<RDOC>NW011992774, NW011992731, NW011992734</RDOC>
<USER>大学1年, 女性, 検索歴2.5年</USER>
</TOPIC>

背景・目的、定義、判定基準など

## 正解判定の形式

0001 C gakkai-0000010176 0
0001 A gakkai-0000010187 1
0001 C gakkai-0000010218 0
0001 C gakkai-0000010219 0
0001 C gakkai-0000010220 0
0002 A gakkai-0000010187 1
0002 A gakkai-0000010221 1
0002 C gakkai-0000010222 0

0002 C gakkai-0000010231 0

正解判定結果リストは以下のような形式になっています。

検索課題番号　ダミーフィールド　文書番号　判定結果

正解判定には、検索要求に完全に適合する「A」と、部分的に適合する「B」の2段階 があります。人間の判定者が内容を検討した後、不適合と判定したものは、「C」となっています。正解判定結果のファイルに文書IDが含まれないものは、さまざまな手法で検索して も、正解候補として検索されなかったもので、判定者が直接内容を吟味したわけ ではありませんが、不適合であると想定しています。

これをつかって、検索システムが出力した検索結果の再現率（Recall）と精度（Precision） などの検索有効性の評価指標を算出します

NTCIR

## 評価ワークショップ

- 共通のデータセットと評価手順を提供
  - コンテスト、コンペティションではない
- 参加者は、共通データセットを用いて各自のシステムで実験遂行。独自の目的をもって参加。
- システム間の比較、技術移転、研究上のアイディアの交換を促進
- 評価型ワークショップは、多様なシステムから正解の候補を効率的に収集するよい機会でもある
- TREC、CLEF, DUC, INEX, FIRE　など

NTCIR

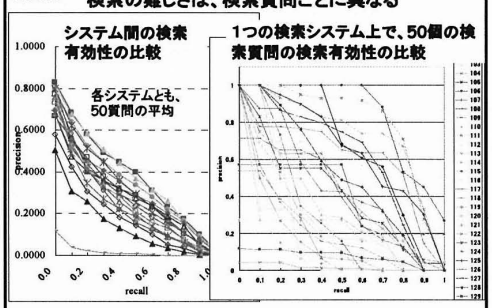
## 情報アクセス技術の評価の6レベル

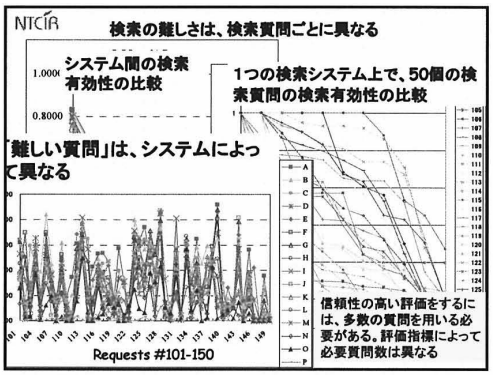
1. 工学レベル：効率efficiency ex.速さ
2. 入力レベル：ex.DBの網羅性、質、新しさなど
3. 処理レベル:検索有効性effectiveness ex.再現率、精度
4. 出力レベル：結果の提示
5. 利用者レベル： ex. 利用者が必要とする労力
6. 社会レベル：重要性

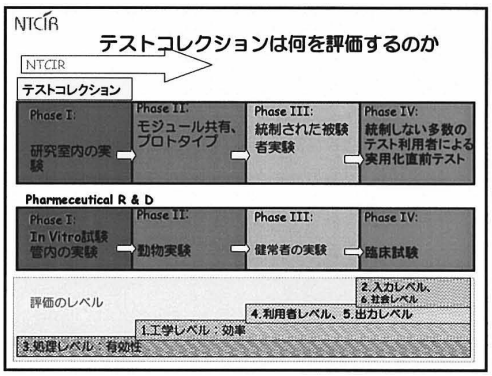
(Cleverdon & Keen 1966)

NTCIR

## 検索の難しさは、検索質問ごとに異なる






NTCIR-6 (2007) 口頭発表セッションの様子

NTCIR-6 (2007) ポスターセッションの様子

NTCIR-6 (2007) banquet の様子

## NTCIR-7

- クラスタ1: 高度言語横断情報アクセス (ACLIA=CCLQA + IR4QA)
- クラスタ2: 多言語意見分析(MOAT) + ~~CLIRB~~
- クラスタ3: 特許翻訳・マイニング (PATMT + PATMN)
- MuST: 動向情報の抽出・要約・可視化
- EVIA: 情報アクセス評価に関する NTCIR併設国際ワークショップ

## TALK OUTLINE

1. NTCIRってなに
2. ACLIA速報
3. MOAT速報
4. PATMT/PATMN速報
5. MuST速報
6. EVIA速報
7. つづきはNTCIR-7 meeting (12/16-19)で…

## ACLIAのねらい

多言語の評価:
日本語、中国語(簡体字、繁体字)
言語横断質問応答 vs 単言語質問応答

言語横断検索(CLIR)コミュニティと
言語横断質問応答(CLQA)コミュニティの協力促進
- QAシステム全体のーモジュールとしてのIRの評価
- QAシステム全体をもっていない研究チームでも
モジュール単位で参加可能

## ACLIA オーガナイザ

- **ACLIA task organizers:**
  - Teruko Mitamura (Carnegie Mellon University)
  - Eric Nyberg (Carnegie Mellon University)
- **IR Meta-Task coordinators:**
  - Tetsuya Sakai (NewsWatch, Inc.)
  - Fred Gey (UC Berkeley)
- **IR for QA coordinators:**
  - Noriko Kando (NII)
  - Donghong Ji (Wuhan University)
- **Japanese CLQA coordinators:**
  - Tsuneaki Kato (Tokyo University)
  - Tatsunori Mori (Yokohama National University)
- **Simplified Chinese CLQA coordinators:**
  - Chin-Yew Lin (Microsoft Research Asia)
  - Ruihua Song (Microsoft Research Asia)
- **Traditional Chinese CLQA coordinators:**
  - Chuan-Jie Lin (National Taiwan Ocean University)
- **ACLIA advisors:**
  - Noriko Kando (NII)
  - Kui-Lam Kwok (Queens College)

## ACLIA – CCLQA (複雑な質問を扱う言語横断質問応答) 速報

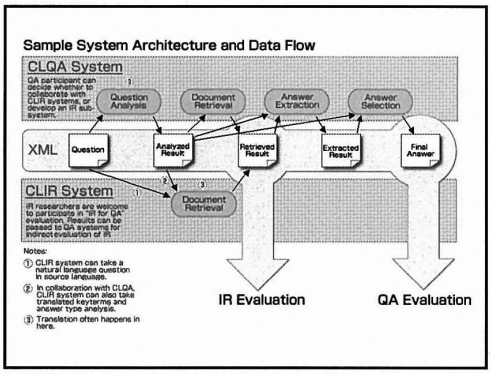
ACLIA=Advanced Cross-Lingual Information Access
CCLQA=Complex Cross-Lingual Question Answering

## CCLQA タスク概要

- ファクトイド質問(NTCIR-5,6)よりも高度で複雑な質問を扱う
- 4つの質問タイプ:定義、関係、出来事、人物情報
  - 定義: *What is the Human Genome Project?*
  - 関係: *What is the relationship between Saddam Hussein and Jacques Chirac?*
  - 出来事: *List major events in formation of European Union.*
  - 人物情報: *Who is Kim Jong-Il?*

## CCLQA タスク 概要 (2)

- 複雑な質問を扱うCLQA(質問:英語)
  - EN-JA (知識源:日本語)
  - EN-CS (知識源:中国語(簡体字))
  - EN-CT (知識源:中国語(繁体字))
- 同等の質問を扱う単言語QA
  - JA-JA (質問、情報源:日本語)
  - CS-CS (質問、情報源:中国語(簡体字))
  - CT-CT (質問、情報源:中国語(繁体字)
- QAシステムの検索モジュールを他チームのものに差し替えたコンビネーションラン




## 提出されたCLQA/単言語QA ランの数

| 参加チーム | 質問解析 | | | | | QA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | EN-CS | CS-CS | CT-CT | EN-JA | JA-JA | EN-CS | CS-CS | CT-CT | EN-JA | JA-JA |
| ATR/NiCT | | | | | | 3 | 3 | | | |
| Apath | 1 | 1 | | | | 2 | 2 | | | |
| CMUJAV | 1 | 1 | | 1 | 1 | 3 | 3 | | 3 | 3 |
| CSWHU | | 3 | | | | | 2 | | | |
| Forst | | | | | | | | | 1 | 1 |
| IASL | | | | | | | 2 | 3 | | |
| KECIR | | 1 | | | | 2 | 1 | | | |
| NTCQA | | | | | | | | | 1 | 1 |
| ORGANIZER | | | | | | 1 | 1 | | 1 | 1 |

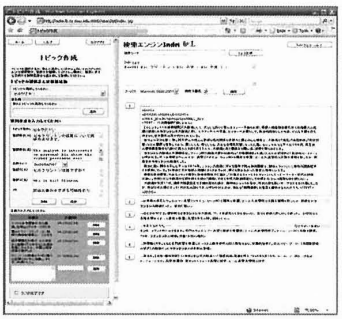
## 提出されたコンビネーションラン の数

| 参加チーム | 必須ラン | | | | | 任意提出ラン | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | EN-CS | CS-CS | CT-CT | EN-JA | JA-JA | EN-CS | CS-CS | CT-CT | EN-JA | JA-JA |
| ATR/NiCT | 6 | | | | | | | | | |
| Apath | | 2 | | | | | 2 | | | |
| CMUJAV | 14 | 20 | | 11 | 14 | 14 | 20 | | 11 | 14 |
| CSWHU | | | | | | | | | | |
| Forst | | | | 11 | | | | | | |
| IASL | | | | | | | | | | |
| KECIR | | | | | | 18 | 20 | | | |
| NTCQA | | | | | | | | | | 14 |
| ORGANIZER | | | | | | | | | | |

## 評価ツールキットEPAN
(Evaluation Package for ACLIA and NTCIR)

- トピック作成ツール
  - 質問や回答ナゲットの作成
- ラン提出ツール
- 正解作成ツール
  - IR4QAの適合性判定
  - CCLQAのナゲット(回答を構成する最小単位)判定
- XMLのimport/export
- マルチユーザによるワークスペース共有
- 管理者ツール(タスク進捗、ユーザ行動履歴)

## EPAN トピック作成ツール



## 評価について

- CLQA/単言語QAランを人手評価
  - ナゲット・ピラミッド評価:ナゲットベースで複数評価者に基づく「重みつきF値」を計算
    [Lin/Demner-Fushman 06]
- CLQA/単言語QA/コンビネーションランを全て自動評価
  - 日本語・中国語版POURPRE
    [Lin/Demner-Fushman 05]

**続きはNTCIR-7成果報告会で!**

# ACLIA – IR4QA
# (質問応答向け文書検索)
# 速報

ACLIA=Advanced Cross-Lingual Information Access
IR4QA=Information Retrieval for Question Answering

オーガナイザ: Tetsuya Sakai, Noriko Kando, Chuan-Jie Lin, Teruko Mitamura, Donghong Ji, Kuang-Hua Chen, Eric Nyberg

## IR4QAのねらい

- 質問応答に適した文書検索技術とは？

e.g. なるべく多様な適合文書を検索
vs 特定の高適合文書をずばり検索

- 質問解析は文書検索の有効性向上に役立つか？
- 質問応答の有効性と文書検索の有効性の関係は？
- 言語依存の問題と非依存の問題は何か？

中国語(簡体字) vs 中国語(繁体字) vs 日本語

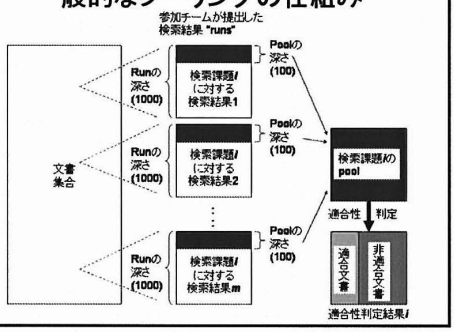
- IR4QAに適した評価方法とは?

## 提出されたIR4QA runsの数

CS=中国語(簡体字) CT=中国語(繁体字) JA=日本語

| team | CS-CS | EN-CS | CT-CT | EN-CT | JA-JA | EN-JA |
|---|---|---|---|---|---|---|
| BRKLY | | | | | 4 | |
| CMUJAV | 2 | 2 | | | 5 | 5 |
| CYUT | | 3 | | 3 | | 3 |
| HIT | | 4 | | | | |
| KECIR | 3 | | | | | |
| MITEL | | 5 | 4 | | | |
| NLPAI | 5 | | | | | |
| NTUBROWS | | | 5 | | | |
| OT | 5 | | 5 | | 5 | |
| RALI | 5 | 4 | 5 | 4 | | |
| TA | | | | | | 3 |
| WHUCC | 2 | | | | | |
| total by lang. pair | 22 | 18* | 19 | 7 | 14 | 11 |
| total by document lang. | 40 | | 26 | | 25 | |

## 一般的なプーリングの仕組み




## IR4QAにおけるプーリング

1. 各トピックについて深さ*d*=30, 50, 70, 90, 100のプールを作成
2. まず深さ30のプールを判定、次に深さ50のプールの未判定文書を判定、etc.

- 約100トピック×3言語について2週間で判定!
- トピック毎に最終的なプールの深さは異なる
- 多くのrunsの上位*d*位以内に検索され、かつ上位に検索されたものを優先的に判定者に提示
- 今後、判定結果をversion upする予定

## IR4QAの評価指標

- 平均精度(AP=Average Precision)

検索研究コミュニティで最も広く用いられているが
多値適合性(高適合、部分適合etc)に対応できない

- Q-measure [Sakai04]

多値適合性に対応したAPの拡張
ユーザモデル [Sakai/Robertson08]

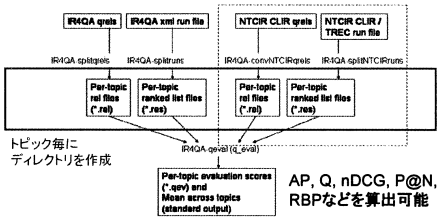
- nDCG [Jarvelin02] のMicrosoft版 [Burges05]

多値適合性に対応した最も広く用いられている指標
APやQよりも、再現率の低いシステムに甘い

## 公開評価ツール: ir4qa_eval

http://research.nii.ac.jp/ntcir/tools/ir4qa_eval-ja

よりダウンロード可能。IR4QAだけでなく、NTCIR CLIRなどTREC-likeなタスクに適用可能。




AP, Q, nDCG, P@N, RBPなどを算出可能

## NTCIR-7成果報告会では…

以下について報告します。

- 各参加チームのアプローチ
- AP, Q, nDCGによる評価や、これらの指標の違い
- 適合性判定結果を全く使わずにシステムをランクづけした場合と、適合性判定結果に基づきシステムをランクづけした場合の相関
- 擬似適合性フィードバックがうまくいったチーム、いかなかったチーム
- その他…

一緒に議論しましょう!

## TALK OUTLINE

1. NTCIRってなに
2. ACLIA速報
3. MOAT速報
4. PATMT/PATMN速報
5. MuST速報
6. EVIA速報
7. つづきはNTCIR-7 meeting (12/16-19)で…

## MOAT (多言語意見分析)速報

MOAT=Multilingual Opinion Analysis Task

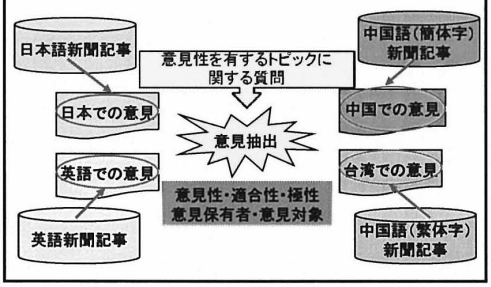
Organizers:

関洋平 (豊橋技術科学大学)
David Kirk Evans (アマゾン・ジャパン)
Hsin-Hsi Chen, Lun-Wei Ku(国立台湾大学)
Le Sun(中国科学院ソフトウェア研究所)
神門典子(国立情報学研究所)

## 動機

- 成熟しつつある意見抽出技術を評価するための共通基盤の提供。
- 多言語の意見分析の傾向の違いを明らかにする。
  - 資源:
    多言語(日本語・英語・簡体字中国語・繁体字中国語)を対象とした新聞記事。
  - 言語を横断した共通トピックの提供
    (comparable opinion corpus)。

## タスクモデル



## NTCIR-7 タスク定義

サブタスク一覧

| サブタスク | 値 | 判定単位 |
|---|---|---|
| 意見性判定 | Yes/No | 文 |
| 適合文判定 | Yes/No | 文 |
| 極性判定 | Pos/Neg/Neu | 意見節 |
| 意見保有者 | 文字列 | 意見節 |
| 意見対象 | 文字列 | 意見節 |

訓練データ
(NTCIR-6 OAT corpus)

| 言語 | 日本語 | 英語 | 中国語 |
|---|---|---|---|
| 文書数 | 490 | 438 | 843 |
| 文数 | 15,279 | 8,356 | 11,907 |
| トピック数 | 30 | 28 | 32 |

## NTCIR-7 MOATの狙い

- アノテーションツールを使い，
  精度の高い意見分析コーパスを作成する。
- 文より細かい粒度の
  意見節の単位を設定し，
  極性(肯定・否定・中立)，
  意見保有者(意見を保有・表明する主体)，
  意見対象(意見の対象)の抽出技術を評価。

## アノテーションツール

- Web ブラウザベースの
  アノテーションツールを開発し，高い κ 係数
  (意見性0.7以上，極性0.6以上，適合性0.55以上)を達成。

| 言語 | 判定属性 | κ平均 | |
|---|---|---|---|
| | | NTCIR-6 | NTCIR-7 |
| 日本語 | 意見文判定 | 0.6740 | 0.7135 |
| | 極性判定 | 0.6153 | 0.6341 |
| | 適合文判定 | 0.5415 | 0.5905 |

(Cohenの) κ 係数:
判定者2名の付与の一致度を表す指標。
-1から1の間の値を取り，1に近いほど一致しており，
0.4以上であればほどほどに一致，
0.6以上であればかなり一致しているとされる。

## アノテーションツール


## アノテーション

- 日本語テストデータ(formal run):
  5,885文に対して6,221意見節をアノテート。
- 意見保有者・意見対象は，
  照応関係を明示的にアノテート。
- CSV, XML フォーマットの提供。

## アノテーション

極性
(POS, NEG, NEU)

意見保有者
(照応あり, なし)

意見対象
(照応あり, なし)

```
<OpinionAnalysisResult topic="N03" document="JA-010913079"
 sentence="0029" opinionated="YES" relevant="NO" polarity="POS"
 holder="(寺島実郎・三井物産戦略研究所長)@3" target="米国">
◆反論、示戦とも予様でオープンな米国社会の強みを狙われた不幸な出来事
だが、平面、柔軟な社会構造が人、モノ、金、技術を世界中から吸収する米国
の成長要素だ。
 <OpinionExpression startChar="1186" endChar="1222" polarity="NEG"
 holder="(寺島実郎・三井物産戦略研究所長)@3" target="同時多発テロ">
事件と、予様で、示戦とも予様でオープンな米国社会の強みを狙われた不幸
な出来事だが、</OpinionExpression>
 <OpinionExpression startChar="1223" endChar="1261" polarity="POS"
 holder="(寺島実郎・三井物産戦略研究所長)@3" target="米国">平面、
柔軟な社会構造が人、モノ、金、技術を世界中から吸収する米国の成長要素
だ。</OpinionExpression>
</OpinionAnalysisResult>
<OpinionAnalysisResult topic="N03" document="JA-010913079"
 sentence="0030" opinionated="YES" relevant="NO" polarity="NEU"
 holder="(寺島実郎・三井物産戦略研究所長)@3" target="米国">
先制にひんしたときの米国の強さは大変なもので、数カ月以内に米国は正常な
状態に戻るはずだ。
 <OpinionExpression startChar="1262" endChar="1284" polarity="POS"
 holder="(寺島実郎・三井物産戦略研究所長)@3" target="米国">先制に
ひんしたときの米国の強さは大変なもので、</OpinionExpression>
 <OpinionExpression startChar="1285" endChar="1305" polarity="NEU"
 holder="(寺島実郎・三井物産戦略研究所長)@3" target="米国">数カ月
以内に米国は正常な状態に戻るはずだ。</OpinionExpression>
</OpinionAnalysisResult>
```

## 参加者数

| 言語 | 日本語 | 英語 | 中国語 | |
|---|---|---|---|---|
| | | | 繁体字 | 簡体字 |
| 参加者数 | 8 | 9 | 7 | 8 |

- 多言語タスクへの参加者チーム数
  2チーム(NTCIR-6)→5チーム(NTCIR-7)
- それ以外に，簡体字中国語と繁体字中国語の双方に参加したチームが3チーム。

## 評価

- 自動評価ツールの提供。
- 多様なオプションにより，
  意見性・極性・適合性について
  様々な基準での分析が可能。
- 前回と比べて参加システムの技術が向上。
- 続きは成果報告会で…

## TALK OUTLINE

1. NTCIRってなに
2. ACLIA速報
3. MOAT速報
4. PATMT/PATMN速報
5. MuST速報
6. EVIA速報
7. つづきはNTCIR-7 meeting (12/16-19)で…

## PATMT(特許翻訳)速報

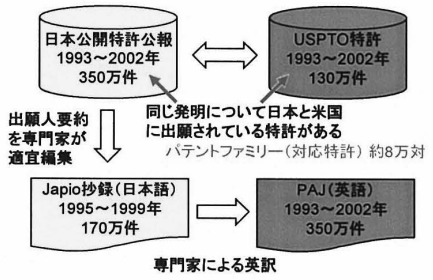
PATMT=Patent Translation

## PATMT(特許翻訳タスク)

藤井敦，山本幹雄，宇津呂武仁(筑波大学)，内山将夫 (NICT)

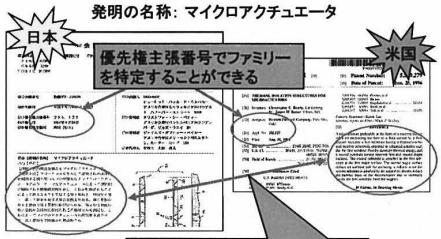
- 動機
  - エンジンとデータが整備されつつある
- エンジン
  - 統計的機械翻訳のデコーダ
- データ
  - 日本公開公報と米国特許(1993～2002年)から抽出したパテントファミリー 8万対
    - 上記パテントファミリーから抽出した対訳文200万対
    - 文翻訳の訓練データとテストデータとして使用可能
- 意義
  - 学術研究と産業上の価値

## NTCIRで配布している特許データ



## パテントファミリーの例

発明の名称：マイクロアクチュエータ




原文：cpu 1 performs the control of the whole electronic musical instrument such as key assigning and tone generating control .

① ・単語やフレーズの単位で英語から日本語に翻訳する
・英日の対応は大量のテキストデータから学習しておく

cpu 1 | performs the control of the | and tone generating control | CPU1は | whole electronic musical instrument such as | key assigning | など電子楽器全体 | キーアサイン

統計的機械翻訳のイメージ

② ・日本語として自然な語順に並べ替える
・日本語の語順も大量のテキストデータから学習しておく

日本語訳：CPU1はキーアサイン、発音制御など電子楽器全体の制御を行う。

## 参加システムの評価方法

- intrinsic
  - BLEUによる自動評価
    - 英仏を対象とした機械翻訳とほぼ同等
  - 人手判定による評価
    - 自動評価との相関を調査中
- extrinsic
  - 言語横断特許検索の精度(MAP)による評価
  (1) NTCIR-5の検索課題(英訳)を日本語に翻訳
  (2) 日本語の特許を検索

ドライラン(予備試験)とフォーマルラン(本試験)を行った

## BLEU: BiLingual Evaluation Understudy

- 参照訳と比較してシステムの翻訳結果を評価する
  - 完全一致で比較すると，ほとんどの翻訳結果が0点
  - そこで，*n*-gram単位の一致率を幾何平均する(*n*=1～4)
  - 値の範囲は [0,1] で，大きいほど良い
- 短い翻訳結果には *BP* でペナルティを与える

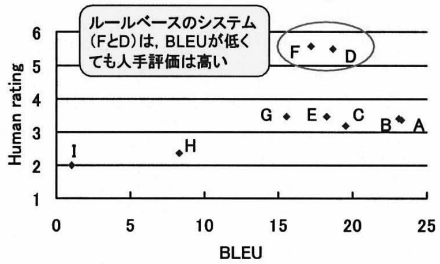
$$BLEU = BP \times \sqrt[4]{\prod_{n=1}^{4} p_n}$$

$$BP = \min(1, e^{1-r/c})$$

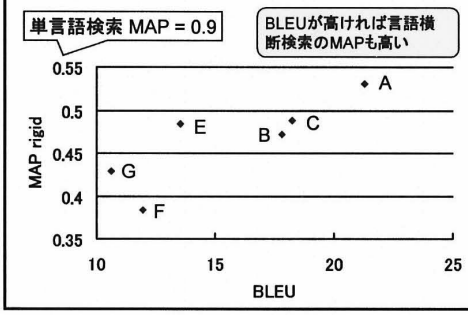
$p_n$ *n*-gramの一致率
$r$ 参照訳の長さ
$c$ システム翻訳の長さ

## ドライランの結果：日英intrinsic

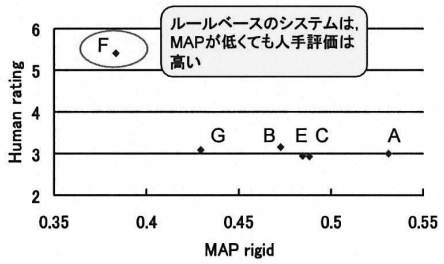
- 人手評価 = Adequacy + Fluency




## ドライランの結果：英日extrinsic



## ドライランの結果：英日extrinsic（つづき）




## ドライランから分かったこと

- BLEUとMAPは相関が高い
- 人手評価は，ルールベースシステムを除けば，BLEUやMAPと相関がある
- フォーマルランでは，ドライランよりも多くのチームが参加し，さらに詳細な分析を行った
  - 詳しくは12月の成果報告会で

## 統計的機械翻訳の講習会 ＠筑波大学（2008.08.28）

- 講義＋計算機による実習
- 講師：PATMTオーガナイザ
- 受講者：大学や企業から42名


## PATMN(特許マイニング)速報

PATMN=Patent Mining

## PATMN(特許マイニングタスク)

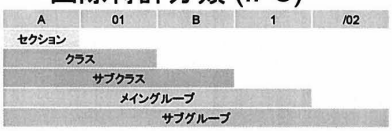
難波英嗣(広島市立大学) 藤井敦(筑波大学)
岩山真(日立製作所/東京工業大学) 橋本泰一(東京工業大学)

**目的**

**特許と論文を対象にした検索や技術動向分析など、様々な目的に利用可能な言語処理技術の開発**

**その第一歩として、**論文抄録を「国際特許分類」(IPC)に自動分類

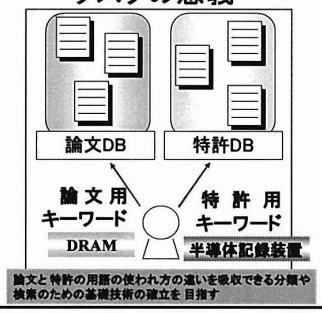
## 国際特許分類 (IPC)




| | | |
|---|---|---|
| A | セクション | 生活必需品 |
| A01 | クラス | 農業、林業、畜産、狩猟、捕獲、漁業 |
| A01B | サブクラス | 農業または林業における土作業、農業機械または器具の部品、細部または附属具一般 |
| A01B 1/00 | メイングループ | 手作業具 |
| A01B 1/02 | サブグループ | 鋤、ショベル |

国際特許分類第6版ではサブグループのレベルで約50,000。
「サブグループ」レベルのIPCコードを論文抄録に付与することを目的とする。

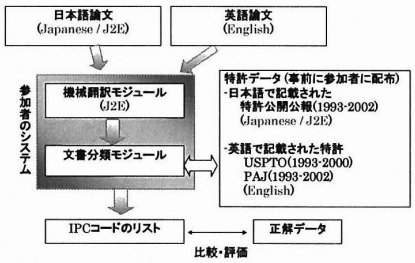
## タスクの意義




## 実施サブタスク

- 日本語サブタスク(Japanese): 日本語の論文を日本語で記載された特許データを用いて分類。
- 英語サブタスク(English): 英語の論文を英語で記載された特許データを用いて分類。
- 言語横断サブタスク(J2E): 日本語の論文を英語で記載された特許データを用いて分類。

## サブタスクのまとめ




## 特許マイニングタスク参加者数

| | 日本 | 日本以外のアジア | 欧州 | 北米 |
|---|---|---|---|---|
| 大学 | 3 | 4 | 0 | 2 |
| 企業 | 2 | 0 | 1 | 0 |

### 参加システム数

日本語サブタスク :24
英語サブタスク :20
言語横断サブタスク: 5

## 評価

正解データ
日英論文抄録にIPCコードを付与したデータを準備

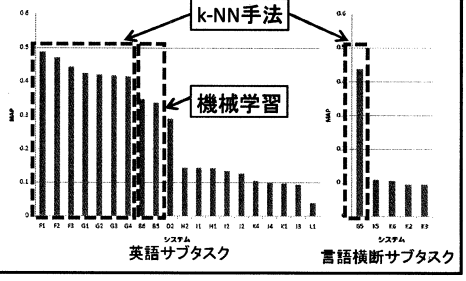
同一内容の特許と論文の対:976
ドライラン : 97トピック
フォーマルラン: 879トピック

1トピックあたり平均 正解数:2.2

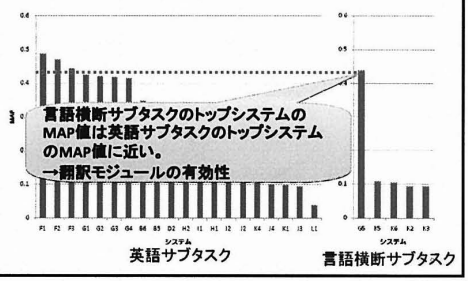
評価尺度
MAP / 再現率 / 精度

## フォー

k-NNとは、入力された論文と最も類似したk個の特許を見つけ、それらに付与されたIPCから、論文のIPCを推定する方法。



## フォーマルラン結果(一部)




## フォーマルランの結果から分かったこと

- 各サブタスクにおいて、成績トップのシステムは、いずれもk-NN手法を採用。
- 機械学習(ナイーブベイズ+ロジスティック回帰)を用いたグループも、トップのシステムに近い成績を得ている。
- 言語横断サブタスクトップのシステムは、日本語サブタスクや英語サブタスクと同等のMAP値を得ている。

## TALK OUTLINE

1. NTCIRってなに
2. ACLIA速報
3. MOAT速報
4. PATMT/PATMN速報
5. MuST速報
6. EVIA速報
7. つづきはNTCIR-7 meeting (12/16-19)で…

## MuST
## (動向情報の抽出・要約・可視化)
## 速報

MuST=Multimodal Summarization for Trend Information

Organizers:
加藤恒昭 (東京大学)
松下光範 (関西大学)

## 位置づけ

- 対話的かつ探索的な情報活用の支援
  - 収集された情報の全体像の概観
  - 関心の絞り込み・具体化や変更
  - 必要な詳細情報へのアクセス
- 言語情報と非言語情報の活用
  - 言語情報と非言語情報の横断的利用
  - マルチモーダルプレゼンテーション

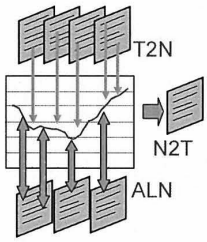
⇨ 動向情報の要約と可視化に着目
–一定期間にわたる情報の総合的まとめあげ
–時間, 地理的空間に止まらない様々な軸
–テキスト情報＋数値情報

## タスクの概要

- 情報活用を目的に, 動向情報を中心としたマルチモーダル要約についての研究を加速し, その検討・評価を行う
- 関心・目的の具体化に応じたふたつのアプローチをとる
  - 評価課題:具体化された研究課題について, 定量的な評価を検討し, 研究を進める
  - 自由課題:萌芽的な研究課題について, 模索的, 提案的な研究を進める

## 評価課題(提案)

- T2N:テキストからの時系列統計量の抽出と可視化
- N2T:時系列数値情報の言語化
- ALN:時系列数値情報とテキスト情報とのアラインメント




## 資源

- MuST データセット
  - 1998,1999年の毎日新聞記事より27トピック581記事について注釈付けを行ったもの
- 可視化プラットホーム
  - 情報活用を目的とした可視化システム構築ためのプラットホーム
  - オープンソースで提供予定
- 変化表現コーパス
  - テキスト中で表現される統計量の値やその変化を抽出し，一定の形式で整理(コーディング)
  - 9トピックについて219記事より1,789件の情報を抽出

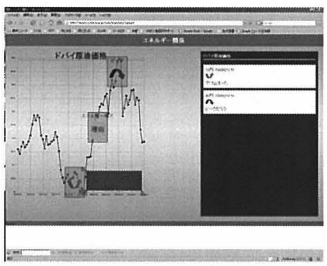
## MuSTデータセット注釈例

統計量名　統計量名　日付

<unit stat="レギュラーガソリンの全国平均店頭価格"><date gra="週" abs="19990617">今週</date>**調査の**<name part="head">ガソリン価格(レギュラー1リットル)</name>**は**<name part="foot">全国平均</name>**で**<val>92円</val>**となり、**<date gra="週" abs="19990610">**前週**</date>**の**<name part="foot">**平均**</name>**に比べ**<rel type="diff">1円</rel>**上昇した**</unit>。<unit stat="レギュラーガソリンの全国平均店頭価格"><name part="foot">**全国平均**</name>**の上昇は**<date gra="旬" abs="19990201">**2月上旬**</date>**以来**、<dur gra="月">4カ月</dur>**ぶりだ**</unit>。

値　差分　期間

## 可視化プラットホーム



## 変化表現コーパス

**97年(1～12月)の国内本体出荷台数は、前年比3・4%増の704万2000台、本体出荷金額は同1・7%増の1兆7147億円と、消費税駆け込み需要前の4月以前の「特需」に支えられ、かろうじて前年を上回った。**

**統計量 = 国内本体出荷台数**
**時点 = <date gra="年" abs="1997">97年(1～12月)</date>**
**Type0.val = 704万2000台**
**Type1.change = upward**
**Type1.ref = <date gra="年" abs="1996">前年</date>**
**Type1.diff = <rel stat="国内出荷台数" type="prop">3・4%</rel>**

**統計量 =本体出荷金額**
**時点 = <date gra="年" abs="1997">97年(1～12月)</date>**
**Type0.val = 1兆7147億円**
**Type1.change = upward**
**Type1.ref = <pro ref="前年比" id="980204080_2">同</pro>**
**Type1.diff = <rel stat="国内出荷額" type="prop">1・7%</rel>**

## 実施状況

- 14グループ(若干の変動あり)が参加
- 様々な自由課題
  - 08年3月末の成果進捗報告会にて中間報告
- 評価課題はT2N課題のみ実施
  - T2N:テキストからの時系列統計量の抽出
  - 8トピック25統計量
    - ガソリン　ガソリン価格，原油価格
    - ネット携帯加入者　I-mode, EG-web, J-Sky
    - デジカメ　出荷台数，出荷額　等々
  - 5グループ，9システムが結果提出

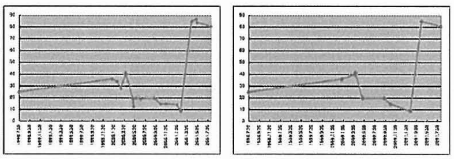
## 自由課題の例

- 表形式データ可視化手法による新聞記事コーパスの可視化
- 動向情報に基づくオンラインニュースフィルタリング
- 刑事事件の関連記事抽出に基づく量刑推移の可視化
- 数値固有表現情報に関わるテキストマイニングと可視化
- 時系列データの全体傾向のより分かりやすい言語表現の生成
- 動向情報分析の多言語化

http://must.c.u-tokyo.ac.jp/ 成果進捗報告会 参照

## T2N課題の実施例

T2N010401 内閣支持率



対象記事中のすべての情報を用いて描いたグラフ

あるシステムの抽出結果に基づく再現

## NTCIR-7 Meetingでは...

- 公開予定の資源を紹介
- 様々な自由課題の成果と進捗
- 評価課題の傾向分析と利用技術を紹介

於 MuST Session
乞う！ご期待

## TALK OUTLINE

1. NTCIRってなに
2. ACLIA速報
3. MOAT速報
4. PATMT/PATMN速報
5. MuST速報
6. EVIA速報
7. つづきはNTCIR-7 meeting (12/16-19)で...

## EVIA(情報アクセス評価に関する国際ワークショップ)速報

EVIA=International Workshop on Evaluating Information Access

Chairs: Mark Sanderson and Tetsuya Sakai

## EVIA

- 情報アクセス評価の国際ワークショップ
- プログラム委員会による査読あり。
  オンライン予稿集あり。NTCIRと関係なくても可
- 第一回: 2007年5月15日(NTCIR-6 Day 1)
  米TREC・東アジアからの招待講演
  +論文+ポスター+EVJF(Evaluating Japanese Food)
- 第二回: 2008年12月16日(NTCIR-7 Day 1)

TREC, NTCIR, CLEF, INEX, CHORUSほか多様なコミュニティ(日本, 中国, スペイン, 米国, 英国, オーストラリア, インド, スウェーデン, アイルランド)から12件の投稿あり。

## TALK OUTLINE

1. NTCIRってなに
2. ACLIA速報
3. MOAT速報
4. PATMT/PATMN速報
5. MuST速報
6. EVIA速報
7. つづきはNTCIR-7 meeting (12/16-19)で…

## NTCIR-7でお会いしましょう

- 会場:国立情報学研究所(神保町／竹橋)
- 日時:2008年12月16-19日
- 学生さんの論文集なしの参加は無料です!
- 参加登録:http://research.nii.ac.jp/ntcir/ntcir-ws7/meeting/

| | 12/16(火) | 12/17(水) | 12/18(木) | 12/19(金) |
|---|---|---|---|---|
| 午前 | NTCIR 概要紹介 | ACLIA – CCLQA | PATMT 招待講演 | 招待講演 MOAT |
| 午後 | EVIA 2008 | 招待講演 PATMN | ACLIA – IR4QA 招待講演 | MuST |

宴

ご清聴ありがとうございました