Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

# NTCIR-9 総括と今後の展望

酒井　哲也[†1]　上保　秀夫[†2]　神門　典子[†3]
加藤　恒昭[†4]　相澤　彰子[†3]　秋葉　友良[†5]
後藤　功雄[†6]　木村　文則[†7]　三田村　照子[†12]
西崎　博光[†8]　嶋　秀樹[†12]　吉岡　真治[†9]
Shlomo Geva[†10]　Ling-Xiang Tang[†10]
Andrew Trotman[†11]　Yue Xu [†10]

NTCIR (エンティサイル) は国立情報学研究所 (NII) が主催する情報アクセス技術の評価型ワークショップシリーズで，1999 年以来一年半サイクルで開催されている．第 9 回ワークショップ (NTCIR-9) が 2011 年 12 月に終わり，第 10 回 (NTCIR-10) の採択タスクが今年 1 月にアナウンスされたところである．本稿では，NTCIR-9 の簡単な総括を行うとともに，NTCIR-10 の新たな取り組みについて紹介する．

## Summary of NTCIR-9 and A Look Forward

Tetsuya Sakai, Hideo Joho, Noriko Kando,
Tsuneaki Kato, Akiko Aizawa, Tomoyoshi Akiba,
Isao Goto, Fuminori Kimura, Teruko Mitamura,
Hiromitsu Nishizaki, Hideki Shima,
Masaharu Yoshioka, Shlomo Geva,
Ling-Xiang Tang, Andrew Trotman and Yue Xu

NTCIR is a sesquiannual evaluation workshop series on information access which started in 1999 and is run by National Institute of Informatics (NII). The ninth workshop (NTCIR-9) was concluded in December 2011 and the accepted tasks for the tenth workshop (NTCIR-10) have been announced in January 2012. This paper provides a brief summary of NTCIR-9, and introduces some latest developments from NTCIR-10.

## 1. はじめに

NTCIR (エンティサイル) はもともと NII Test Collection for Information Retrieval systems として 1999 年にスタートしたが，最近では NII Testbeds and Community for Information access Research とも呼ばれており，狭義の情報検索 (文書検索) のためのテストコレクション作成を伴う評価型ワークショップから，広義の情報検索 (ユーザを所望の情報に迅速かつ的確に導くための多様な技術，すなわち情報アクセス) の評価型ワークショップへと発展してきた[9]．米国の TREC (Text Retrieval Conference; 1992 年スタート) や欧州の CLEF (Cross-Language Evaluation Forum; 2000 年スタート) などと共に，一定期間内に同じ土俵の上で技術を競い合う，もしくは協力することを通じて情報アクセス技術を振興するための国際的な場となっている．2011 年 12 月に開催された第 9 回 NTCIR ワークショップ (NTCIR-9) には，7 つのタスク (オリンピックで言えば種目に相当する) が設けられ，14 の国もしくは地域から 90 の異なる研究機関が参加した．NTCIR は一年半サイクルで動いているので，次回の NTCIR-10 は 2013 年 6 月に開催される．これに向けて，NTCIR-10 の採択タスクが 2012 年 1 月に公表され，各タスクが始動したところである．本稿では，NTCIR-9 の簡単な総括を行うとともに，NTCIR-10 の新たな取り組みについて紹介する．

なお，今回の研究発表会では,「曖昧なクエリと (不) 明快なクエリ: NTCIR-10 INTENT-2 と 1CLICK-2 タスクへの誘い」という発表が別途あるので[8]，INTENT と 1CLICK の両タスクについてはそちらを参照いただきたい．また，9 章の Math タスクは NTCIR-10 にて導入された新しいパイロットタスクである．

†1 Microsoft Research Asia, China tetsuyasakai@acm.org
†2 筑波大学 hideo@slis.tsukuba.ac.jp
†3 国立情報学研究所
†4 東京大学
†5 豊橋技術科学大学
†6 情報通信研究機構
†7 立命館大学
†8 山梨大学
†9 北海道大学
†10 Queensland University of Technology, Australia
†11 University of Otago, New Zealand
†12 Carnegie Mellon University

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

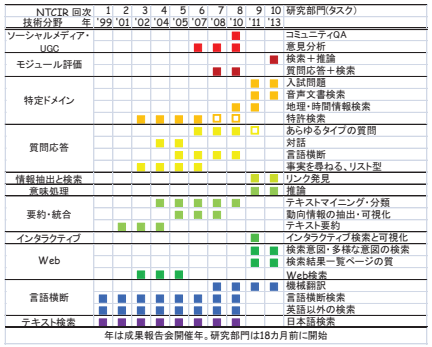
## 2. NTCIR 概観

### 2.1 NTCIR の概要

NTCIR プロジェクトは，大規模な評価実験用の研究基盤を提供することによって，情報アクセス技術研究を促進することを目的として，1997 年末に開始した．1998 年から評価型ワークショップ NTCIR-1 を開始し，その成果報告会は 1999 年 8 月 30 日〜9 月 1 日に開催した．以降，概ね 1 年半を 1 サイクルとして，活動を続け，学術論文，新聞記事，特許，WEB，Yahoo!知恵袋，音声文書，Wikipedia など多様な文書データを用い，再利用可能なテストコレクション（実験用データセット）を構築し，研究目的で公開している[*1]．いままでの NTCIR の対象技術分野は**図 1**(左) のとおりである．NTCIR ワークショップの手順は図 1(右) のとおりである．タスクと呼ばれる複数の研究部門を国際的なタスク選定委員会で選定し，タスクごとに実験を進めていく．タスクへの参加団体数は回を追うごとに増加し，NTCIR-9 では 120 団体が 1 つ以上のタスクに参加登録をし，そのうち 90 団体が実験結果を期限内に提出した (**図 2**(左) 参照)．成果報告会は，公開の国際ワークショップ (図 2(右)) として開催し，情報アクセスの評価手法に関する査読付きワークショップ EVIA (International Workshop on Evaluating Information Access) と連続開催である．EVIA と NTCIR ワークショップの Proceedings はオンラインで公開している[*2]．NTCIR の焦点を**図 3**(左) にしめした．初期には大規模なテストコレクション構築が最大の焦点であった．その重要性は今も変わらないが，次第に，新しい情報アクセス技術を提案し，集中して研究を進める場としての重要性が増している．さらに，近年，研究アイディアの交換，評価手法の研究，議論など，このフォーラムとしての機能がより重要になっている．

### 2.2 情報アクセス技術の評価

情報アクセス技術の評価には多様なレベルがあるが，NTCIR では，主に，処理の有効性の評価を主眼としている (図 3(右))．また，処理の有効性（Recall, Precision）を比較する場合，システム間の差よりも検索質問ごとの差の方が大きく，かつ，どの検索質問が難しいかはシステムごとに異なる (**図 4**)．より信頼性の高い評価を行うには多数の質問を用いる必要があり，評価指標やタスクの性質によっても必要な質問数は異なる．




**NTCIRワークショップの手順**

- 研究部門（タスク）のプロポーザル募集
- 研究部門・対象技術の選定 – タスク選定委員会
  – データ提供機関との協議・契約
- 研究部門ごとに、実験・評価手法の提案・議論
- 参加申込
  – 訓練用データ（文書、課題、正解）配布
    ・各参加者が実験、チューニング
  – 評価用データ（文書、課題）配布
- 検索結果の提出
- 提出結果から正解候補収集、人手判定
- 正解公開、評価結果発表
- 成果報告会　　次回について協議

**図 1** (左)NTCIR の対象技術分野の推移／ (右)NTCIR ワークショップの手順

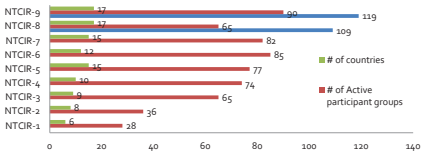



**図 2** (左) 研究部門（タスク）の参加団体数と国数の推移／ (右) 成果報告会の様子

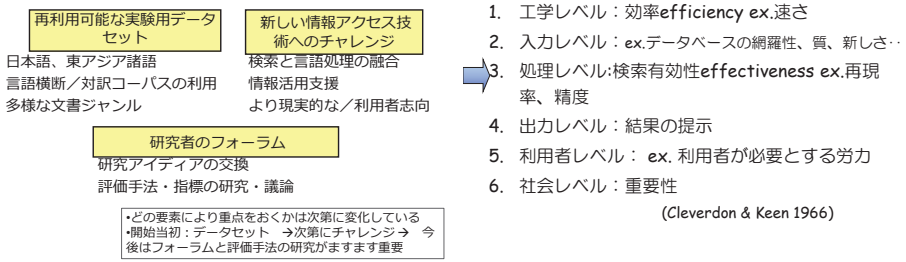



**図 3** (左)NTCIR の焦点／ (右) 情報アクセス技術の評価のレベル

---

⋆1 研究目的で公開しているテストコレクションとツールの一覧，概要説明，利用申し込み方法等は，http://research.nii.ac.jp/ntcir/data/data-ja.html をご参照ください．現在のテストコレクション利用者数は 2243.

⋆2 http://research.nii.ac.jp/ntcir/publication1-ja.html

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

他方，テストコレクションを用いた情報アクセスシステムの評価は薬の開発における Phase I すなわち試験管内実験にたとえられる．研究アイディアが浮かんだらすぐにそのアルゴリズムを実装し，その有効性を評価する．研究のごく初期に高価な利用者実験をしなくても，アルゴリズムの有効性をすぐに，繰り返し評価・検証できることにより研究がスピードアップすることが大きな価値である．研究者は，多様なアイディアを次々と評価し，有望なものが実用化に向けたステップを進んでいき，その過程で，多様なレベルで評価し，検証と改善を行う (**図 5**(左))．現在，情報アクセス技術は社会基盤であり，多様なアプリケーションがそれぞれの研究開発の過程で繰り返し評価・検証が必要である．個別の要件に適した実験用データセットを各研究グループが開発する必要性も高まっている．NTCIR などの評価ワークショップの活動を通して，情報アクセス技術の研究者は，評価手法についても研究を蓄積してきた．たとえば，信頼性，感受性の高い評価が可能な，データセットをより効率よく構築し，適切な評価指標と実験計画を用いるということについても多くの知見を積み重ねている．現在，情報アクセスシステムは，多くがユーザの介在を前提とし，評価検証はよりチャレンジングな課題となっている．

### 2.3 NTCIR-9 の概要

NTCIR-9 では，タスクを一新するとともに，1) 運営組織を一新し，強固な運営基盤を築くとともに，2) コミュニティ主導のタスク運営とし，研究の持続性を図った．運営では，General Chairs, Program Chairs とも複数体制とし，タスクを統括する Program Chairs を一新した．また，タスクの多様性を広げ，図 5(右) のように，幅広いコンテクストと多様なメディアを対象とした情報アクセス技術をとりあげ，実社会にインパクトを与える情報技術の研究をめざす方向をより明確にした．その結果，NTCIR-9 は 40 名ものタスクオーガナイザがいきいきと自律的に活動を進め，コミュニティを拡充し，オーガナイザ・参加者が一丸となって新たな課題に取り組んだ．海外強豪チームの参加も増加し，タスク参加団体と成果報告会出席者ともに過去最大となり，新たな運営体制の有望性が示唆された．

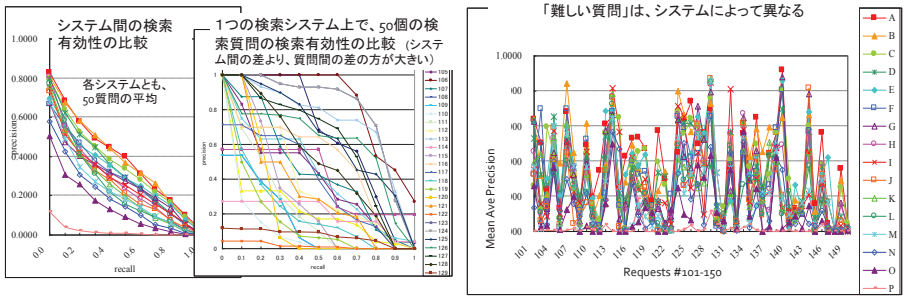



**図 4** 検索質問数と評価の信頼性

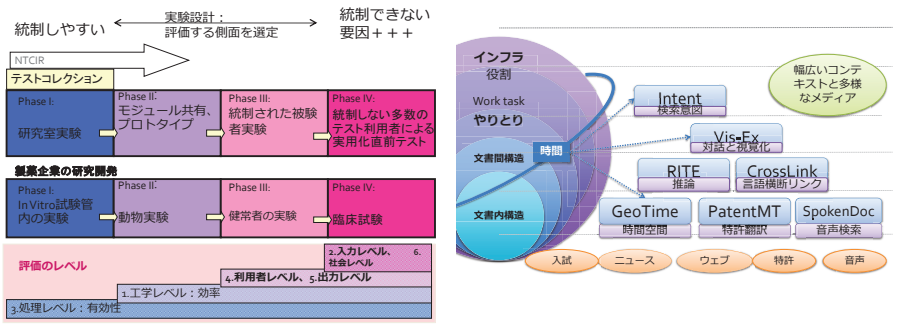



**図 5** (左) 実験のスペクトラム (製薬開発との対比) ／ (右)NTCIR-9 のタスク

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

## 3. CrossLink

Cross-lingual link discovery (CLLD) とは，異なった言語間の文書の間で, リンクを持つ可能性がある文書を自動的に発見することである．本タスクでは，英語の文書をソース文書とし，その文書に対応する文書を他の言語のコーパスより発見し，ソース文書から発見した対応文書に対してリンクを行う．発見対象となる言語は中国語，韓国語，日本語のいずれかである．

実際に異なった言語間においてリンクを持っている言語資源の例として，Wikipedia が挙げられる．Wikipedia は，多くの言語をカバーしている非常に多くの記事を含んだオンラインの多言語百科事典である．Wikipedia は同じ言語における文書間での広範にわたるハイパーリンクを持っており，それにより可読性が高く，参照も容易となっている．しかし，異なる言語間においては，同じ記事（項目）について記述されたページ間で言語を越えてリンクを持っている場合を除いては，言語を越えて文書間でリンクを持つことはまれである．このことは，利用者が情報や知識を異なる言語資源から探そうとする際に，重大な困難をもたらす．

**図 6** は，Cross-lingual link discovery タスクにおいて，発見したい CrossLink の例を示している．図 6 の左側のページは利用者が十分に理解できない言語版で記述された Wikipedia のある項目の記事である．その文中に “crema pasticcera” という単語が出現しているが，この単語からは，利用者が理解できる言語版へのリンクがないため，利用者はこの単語の意味を知ることができない．この単語に対し，利用者が理解できる言語版で記述されたページ（図 6 の右側のページ）への適切なリンクを発見したい，というのが Cross-lingual link discovery タスクの目的である．

NTCIR-10 では，NTCIR-9 と同じく，以下の 3 つのサブタスクが用意される．

- English to Chinese CLLD (E2C)
- English to Japanese CLLD (E2J)
- English to Korean CLLD (E2K)

NTCIR-9 では，11 グループが参加し，全部で 57run の結果が提出された．サブタスクごとでは，中国語が 25run，日本語が 11run，韓国語が 21run 提出された．

参加者は，参加するサブタスクにおいて対象となる言語で記述されたトピックの中から，与えられた英語版のトピックに対して最も適切である記事を見つけることとなる．参加するサブタスクに対する実行結果を提出するためには，参加者は以下のことを要求される．

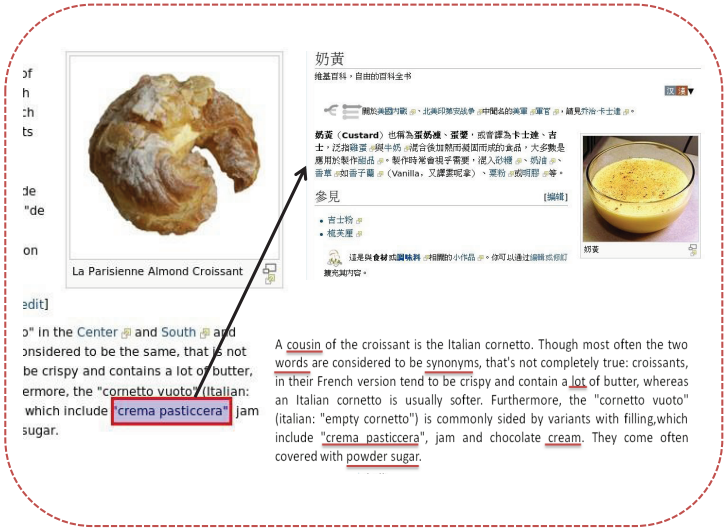



**図 6** 発見したい CrossLink の例

- 与えられた記事の英語版の文書中で，意味のあるアンカー（多言語へは未リンク）を特定する．
- 特定されたそれぞれのアンカーに対して，最も関連する文書を対象となる言語のコーパスから特定する．

評価は，File-to-File （記事間）評価および，Anchor-to-File （各アンカーと記事間）評価について行う．それぞれの評価項目に対し適合率と再現率を用いて評価を行う．

Cross-lingual link discovery の研究の結果として，タスクにおいて要求されたのとは逆方向のリンクを発見することも期待される．これにより，英語と中国語，韓国語，日本語を対象とした双方向 Cross-lingual link discovery へと発展することが期待される．このタスクの最終目的は，自動化された CLLD アプローチのための，再利用可能なリソースを創り出すことである．この研究の結果はリンク発見の自動化システムの構築および改良に用いることができる．また，本タスクにおいては Wikipedia を対象に Cross-lingual link discovery を行っているが，このタスクにおいて得られた知見を基に，将来的には自動化された CLLD アプローチを Wikipedia 以外の対象にも広く適用することも目標である．

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

## 4. GeoTime

### 4.1 タスクの概要

情報検索において，場所や時間に関する問い合わせは非常に多く，特に，地理情報に関する検索については，継続して研究が続けられている (Workshop on Geographic Information Retrieval GIR は 2010 年に第 6 回目となるワークショップを開催した[4]．GeoTime タスクは，これらの研究の流れを踏まえ，アジア言語における地理情報検索の評価に加え，時間に関する情報を含む検索に対する評価を目標としてタスクが提案された．

タスクの設計にあたっては，これまでの質問応答タスクにおいて，多くの地理時間情報に関する質問が存在していたことなどを踏まえ，地理時間情報に関する質問に対して，その質問への答を含む文書を検索するという質問応答のための情報検索の形式でタスクの設計が行われた．

### 4.2 タスクで用いるデータ

使用する言語は，英語と日本語で，対象とする検索文書は新聞記事である．NTCIR-8 では，毎日新聞と New York Times の 2002-2005 年の新聞記事を利用した．NTCIR-9 においては，収録期間を 1998-2001 年に広げるため，日本語については，毎日新聞の当該期間を追加し，英語については，Mainichi Daily,Xinhua English,Korea English の当該期間のデータを追加した．詳細については，NTCIR GeoTime[⋆1] のページを参照されたい．

各検索質問は，質問応答のための情報検索と同様に，**図 7** の形式で与えられる．特徴としては，時間情報を含む検索が存在するため，QUERYDATE の時点での判断を行う形式となっている[⋆2]．

各参加者は，日英 2 言語で与えられた検索質問と 2 言語による文書データベースの組み合わせをもちいて，単言語による情報検索タスクと言語横断の情報検索タスクに参加することができる．表 1 にタスクごとの参加者数を示す (E:英語，J:日本語で，検索質問-文書データベースの組み合わせ)．

### 4.3 タスクの特徴と考察

地理時間情報に対する情報検索の大きな特質としては，単純な単語ベースの類似度ではその適合度がはかりにくいという点があげられる．例えば，図 7 の例の場合を考えたとき，地理情報を表す有力なキーワードは「アフリカ」になるが，例えば，読者にとって「ナイジェリア」が「アフリカ」にあることが当たり前の場合に,「アフリカ」の「ナイジェリア」といった記載はされないため,「アフリカ」と「ナイジェリア」の関係を扱うための知識が必要となる．

一般の情報検索においても，同様の問題が発生するが，動詞や一般名詞のレベルの表記のぶれや類義語などについては，同義語や表記のバリエーションの数も限定されており，検索質問と実際の地名の記述の粒度の違いが大きい可能性のある地理情報とは，少し状況が異なる．特に，単純な動詞や一般名詞のレベルのバリエーションなどでは，疑似適合文書フィードバック＋検索語拡張などにより，ある程度カバーできる可能性があると考えられる．

この問題を扱うために，参加者の多くは，地理情報について，固有名抽出のシステムや Yahoo! Placemaker の利用や，GeoNames や Wikipedia といったオープンな地理情報データベースの利用方法の提案があった．

また，参加システムの多くは，既存の情報検索システムが作成するスコアによるランキングを，地理時間情報などを基準とした情報を用いてリランキングするという方法を利用して

**DESCRIPTION(EN)** When and where did a pipeline explosion occur in Africa killing over 500 people?

**DESCRIPTION(JA)** 500 人以上の死者を出したパイプライン事故は，アフリカのどこで，いつ起きましたか？

**NARRATIVE(EN)** An oil pipeline exploded in an African oil-producing country and the resulting fire killed more than 500 people. The user wants to know where this took place and when was the date of the accident.

**NARRATIVE(JA)** アフリカの産油国で起きたパイプラインの爆発で 500 人以上の死者を出す火災が起きた．ユーザはこの爆発が起きた場所と日付を知りたい．

**QUERYDATE** 20051231

**図 7** 検索質問の例

**表 1** タスク参加者数

| | E-E | E-J | J-J | J-E |
|---|---|---|---|---|
| NTCIR-8 | 6 | 3 | 8 | 1 |
| NTCIR-9 | 7 | 2 | 6 | 1 |

⋆1 http://metadata.berkeley.edu/NTCIR-GeoTime/
⋆2 本タスクの質問のほとんどでは，文書データベースの収録期間の最終日 2005 年 12 月 31 日となっている

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

いる．また，質問に対する答を質問応答システムで作成して検索に生かす方法や，地理時間情報の記述が多い記事には，正解文書が多い可能性があるといった正解文書が持つべき特質を考慮したランキングなども提案されている．

時間情報の取り扱いについては，まだ，検討が十分行われている段階ではなく，相対的な日付の情報 (例えば,「明日」など) の表記の正規化を行う方法や，新聞記事においては，後日の記事ほど，整理されている情報が記載されやすいといった方法などが提案されるに留まっている．

他言語の比較という点においては，先にも述べた適合文書中における地名の記載方法の差異による課題の難しさの差異という現象が見受けられた．具体的には,「アフリカの火山」に関する質問に対し，日本語の記事では,「アフリカ」と明示されていない「コンゴ民主共和国」の「火山」に関する記事が適合文書であるのに対し，英語の記事では,「Democratic Republic of Congo, Africa」と Africa であることが明記されていたため，難易度に大きく差異が出たというケースがあった．

また，一般的に，トピックの中心となる固有名詞が検索質問中に与えられている場合と与えられていない場合では，難易度が大きく異なることが確認されている．これは，多くのシステムが文書検索に重きをおいたシステムであり，文書検索において，検索に有用な固有名詞の存在が大きな影響を与えていることが推定される．

### 4.4 まとめと今後の展開

NTCIR-8,NTCIR-9 の 2 回の GeoTime タスクを通じ，主に地理情報を用いた情報検索についての様々な技術の開発が行われた．特に，近年，容易に利用できるようになってきたオープンな地理情報などを活用する方法や，従来型の検索システムのスコアリングに，地理情報などの文書へのアノテーション情報に基づいて算定されたスコアを融合させる方法などについては，一定レベルの手法が提案されたのではないかと考えている．

また，検索質問の難易度を推定するための基準についての議論を行うための基盤はできつつあり，今後の同様なタスクを設計する際には，この知見が活用できると考えている．

今後は，Twitter や Blog など新聞記事以外の情報への展開についても検討を考えている．その際には，テキスト以外から獲得できる様々な地理時間情報の活用などを検討する必要があり，そこで用いる手法などについての調査を踏まえて，新たなタスク提案を行う方向で検討を行っている．

# 5. PatentMT

## The Patent Machine Translation Task
### — Summary of NTCIR-9 and Plans for NTCIR-10 —

Isao Goto (NICT)
Bin Lu (City Univ. of Hong Kong / Hong Kong Institute of Education)
Ka Po Chow (Hong Kong Institute of Education)
Eiichiro Sumita (NICT)
Benjamin K. Tsou (Hong Kong Institute of Education / City Univ. of Hong Kong)

1

## Summary of NTCIR-9

2

## Motivation

- There is a significant **practical need** for patent translation.
  - to understand patent information written in foreign languages
  - to apply for patents in foreign countries
- Patents constitute one of the **challenging domains**.
  - Patent sentences can be quite **long** and contain **complex structures**

3

## Goals of PatentMT

- To develop **challenging** and **significant practical** research into patent machine translation.
- To **investigate** the **performance** of state-of-the-art machine translation systems in terms of patent translations involving Japanese, English, and Chinese.
- To **compare** the effects of **different methods** of patent translation by applying them to the same test data.
- To **create** publicly-available **parallel corpora of patent documents** and human evaluations of MT results for patent information processing research.
- To **drive machine translation research**, which is an important technology for cross-lingual access of information written in unknown languages.
- The ultimate goal is **fostering scientific cooperation**.

4

## Findings of Previous Patent Translation Tasks

| | | |
|---|---|---|
| NTCIR-7 | Human evaluation | **RBMT** was **better** than **SMT** for **JE** and **EJ**. |
| | CLIR evaluation | SMT was better than RBMT for EJ.<br>▪ The translations were used as bag-of-words.<br>▪ This means that **word selection** by SMT was better than that by RBMT. |
| NTCIR-8 | Automatic evaluation | A hybrid system (RBMT with statistical post edit) achieved the best score for JE. |

5

## Comparison of NTCIR-7, 8, and 9

| | NTCIR-7 | NTCIR-8 | NTCIR-9 (New) |
|---|---|---|---|
| Language | **Japanese to English**<br>**English to Japanese** | **Japanese to English**<br>**English to Japanese** | **Chinese to English**<br>**Japanese to English**<br>**English to Japanese** |
| Human evaluation | **Adequacy**<br>**Fluency** | No human evaluation | **Adequacy**<br>**Acceptability** (New) |
| Extrinsic evaluation | CLIR | CLIR | No extrinsic evaluation |
| Number of participants | 15 | 8 | **21** |

At NTCIR-9, participants can choose subtasks from three language directions, including **Chinese to English**.

6

## Features of PatentMT at NTCIR-9

- Provided data

| | | |
|---|---|---|
| Training | CE | **1 million** patent **parallel** sentence pairs |
| | | Over **300 million** patent **monolingual** sentences in English |
| | JE | Approximately **3.2 million** patent **parallel** sentence pairs |
| | | Over **300 million** patent **monolingual** sentences in English |
| | EJ | Approximately **3.2 million** patent **parallel** sentence pairs |
| | | Over **400 million** patent **monolingual** sentences in Japanese |
| Development | All | 2,000 patent description parallel sentence pairs |
| Test | All | 2,000 patent description sentences |
| | | 2,000 reference translations |

  - The **periods** for the training and test data are **different** (Training data: 2005 or before, Test data: 2006 or later)
- **Human** evaluation (Primary evaluation)
  - **Adequacy** and **Acceptability**

7

## Notable Findings at NTCIR-9

- **SMT** was the **best** system for **Chinese to English** and **English to Japanese** patent translation.
  - This is the **first time** for **SMT** to be **demonstrated equal or better** quality than that of the top-level RBMT for **English to Japanese** patent translation.
  - The **pre-ordering** method of NTT-UT for SMT is very effective for English to Japanese patent translation.
- **80%** of patent sentences could be understood in the best system for **Chinese to English** patent translation.
- **RBMT** was the best system for **Japanese to English** patent translation.

8

## Remaining Issues of NTCIR-9

- Practical evaluation
  - The quality of translated sentences was evaluated at NTCIR-9.
  - More practical evaluations are also expected.

9

## Plans for NTCIR-10

10

## Outline of the Plans for NTCIR-10

- Three subtasks:

| Subtasks | Training data |
|---|---|
| Chinese to English | **1 million** sentence pairs |
| Japanese to English | Approximately **3.2 million** sentence pairs |
| English to Japanese | |

(Subtasks and training data are the same as at NTCIR-9)

- Participants select subtasks in which they wish to participate.
- Large scale **parallel corpora** and **new test sets** will be **provided**.
- **Practical evaluation** will be added (under consideration).
- **Human evaluation** will be carried out.

11

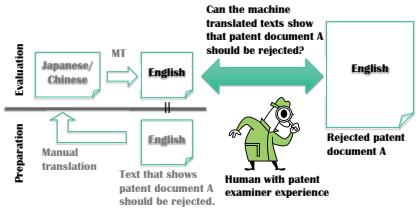
## Differences from NTCIR-9

| | | |
|---|---|---|
| Practical Evaluation (under consideration) | **New**: To explore **practical** MT performance in appropriate fields for patent machine translation. | |
| Intrinsic Evaluation | **Similar to the NTCIR-9 evaluation**.<br>Quality of translated sentences will be evaluated.<br>Additions: | |
| | ***Chronological evaluation*** | Comparison between NTCIR-10 and NTCIR-9 to **measure progress**. |
| | ***Multilingual evaluation*** | Comparison of CE and JE translations using the **same English reference** will be added. |

12

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

## Possible Approach to Practical Evaluation




(The feasibility of this is under investigation.
We are working hard to make necessary arrangements.)

13

## Chronological Evaluation

- In addition to the new NTCIR-10 test sets, the NTCIR-9 test sets will be also translated.
- Translations of the NTCIR-9 test sets at NTCIR-10 will be compared to the NTCIR-9 submissions.
- This allows **measurement** of **the progress from NTCIR-9**.

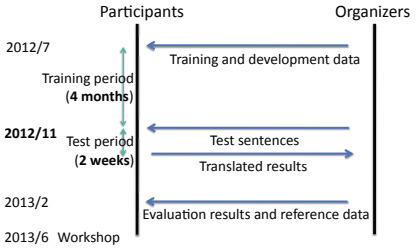
14

## Multilingual Evaluation

- We will produce a **C-J-E multilingual** test set.
- A C-J-E multilingual test set enables **comparison** of the **CE** translations and the **JE** translations using the **same English reference** data.

15

## The Flow and Tentative Schedule of the Task




16

## Why is it so exciting to participate in?

- **Patents** are one of the **challenging domains** for MT.
  - Patent sentences could be quite **long** and contain **complex structures.**
  - Translation between **languages with largely different word order** is difficult for **long** sentences.
- Participants will receive **evaluation results** for their MT quality.
- Participants can use **large-scale patent parallel** and **monolingual corpora**.
- Participants can choose subtasks from three language directions, including **the language direction of Chinese to English**.
- We look forward to many groups participating in PatentMT at NTCIR-10!

17

# 6. RITE

## Overview of NTCIR-9 RITE
(Recognizing Inference in TExt)

Hideki Shima* Hiroshi Kanayama**
Cheng-Wei Lee# Chuan-Jie Lin†
Teruko Mitamura* Yusuke Miyao‡
Shuming Shi* Koichi Takeda**

*Carnegie Mellon University, USA **IBM Research – Tokyo, Japan
#Academia Sinica, Taiwan †National Taiwan Ocean University, Taiwan
‡National Institute of Informatics, Japan *Microsoft Research Asia, P.R. China



## Overview of RITE

RITE is a generic benchmark task that addresses common semantic inference needs in various NLP/Information Access research areas.

- $t_1$: Taro was born in Tokyo.
- $t_2$: Taro was born in Japan.

- $t_1$: Yasunari Kawabata won the Nobel Prize in Literature for his novel "Snow Country"
- $t_2$: Yasunari Kawabata is the writer of "Snow Country"

Given $t_1$, can a computer infer that $t_2$ is most likely true?

Target languages: Japanese, Simplified Chinese, Traditional Chinese



## Motivation

### Information Access applications

- Question Answering; Information Retrieval; Information Extraction; Text Summarization; Automatic evaluation for Machine Translation, Text Summarization, Complex Question Answering

### Success in previous shared tasks

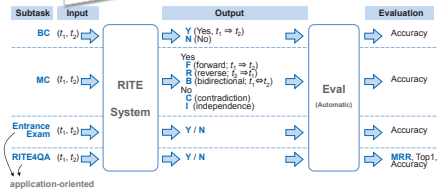
- TREC, CLEF and NTCIR are modern examples of the "Cranfield evaluation paradigm" (Voorhees, 2002)
  - Abstraction of a real Information Access (IA) task is done in a system-centric lab evaluation approach to avoid affects from uncontrollable variables.
  - We'd like to abstract away complexities further and focus on a key semantic processing need that commonly exists in various IA problems



## RITE (Recognizing Inference in TExt)

$t_1$: Yasunari Kawabata won the Nobel Prize in Literature for his novel "Snow Country".
$t_2$: Yasunari Kawabata is the writer of "Snow Country".

**Does $t_1$ entail (infer) $t_2$?**






## Definition of Textual Entailment

- The premise $t_1$ entails the hypothesis $t_2$ if a human (with a common knowledge) reading $t_1$ would infer that $t_2$ is *most likely* true.

- Note that *logical entailment* and *textual entailment* are different.
  - $t_1$: The temperature is only 5 degrees outside.
  - $t_2$: It's cold outside.



## Binary-class (BC) Subtask

### Development process (JA)

1) RITE organizers proposed a small set of sample dataset on an online collaborative spreadsheet to participants.
2) Participants gave feedbacks to the samples and proposed additional samples.
3) Ten college students studied general trends from the sample, and then built training/test data. Sentences were collected from Mainichi newspaper corpus in (somewhat random) various topics. Minimum post-edits are allowed. Controlled to be difficult to solve.
4) Four students independently annotated labels on the collected pairs.
5) Pairs with agreement < 3 are discarded. Inter-annotator agreement: 0.829 (Fleiss' Kappa).
6) Organizers randomly split the dataset into dev and test, with label distribution balanced.



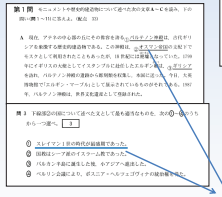
## Multi-class (MC) Subtask

- A system needs to classify a pair into one of five categories considering entailment direction, paraphrase and contradiction.
- Output labels
  - F: forward entailment ($t_1$ entails $t_2$ AND $t_2$ does not entail $t_1$).
  - R: reverse entailment ($t_2$ entails $t_1$ AND $t_1$ does not entail $t_2$).
  - B: bidirectional entailment ($t_1$ entails t2 AND $t_2$ entails $t_1$).
  - C: contradiction ($t_1$ and $t_2$ contradict, or cannot be true at the same time).
  - I: independence (otherwise)
- Motivation: in Text Summarization, knowing textual entailment direction helps to choose one from multiple summary candidate sentences. Contradiction detection is also useful for finding contradicting opinions.
- Sentence length are controlled.



## Entrance Exam Subtask

**Entrance exam problem**
National Center Test for University Admission
(*Daigaku Nyushi Center Shiken*)



**Wikipedia**

$t_1$: スレイマン1世は数多くの軍事的成功を収めてオスマン帝国を最盛期に導いた。(Suleiman I contributed in a lot of military successes and led the Ottoman Empire to its peak.

$t_2$: オスマン帝国ではスレイマン1世の時代が最盛期であった。(The Ottoman Empire's peak was during the reign of Suleiman I).



## Entrance Exam Subtask

- Covers wide range subjects where different problem arises
  - Domestic and World History, Politics, Economy, and Modern Society
  - In History, geo-temporal reasoning may be required.
- Source difference in $t_1$-$t_2$ causes vocabulary mismatch (e.g. "bin Laden" and "bin Ladin")
- Has a natural distribution of linguistic phenomena as seen in an exam-solver application
- Social impact - can wow the public

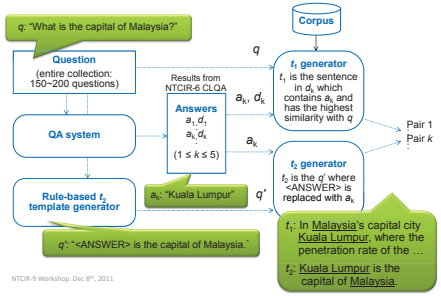


## RITE4QA Subtask

- Can a RITE system rank a set of unordered answer candidates in QA?
- The dataset is created fully automatically from Japanese monolingual data at NTCIR-6 CLQA (Factoid Question Answering)
  - t1: answer-candidate-bearing sentence
  - t2: a question in an affirmative form
- A system is required to generate an additional confidence score used for the ranking process
- Also has a natural distribution of linguistic phenomena
- Uses QA evaluation metrics for result comparability



## RITE4QA Subtask






## Comparison with Related Works

| | Lang | Entail-ment | Entail-ment Direction | Para-phrase | Contra-diction | Answer validation for QA |
|---|---|---|---|---|---|---|
| TAC RTE (2-way) | EN | X | | | | |
| TAC RTE (3-way) | EN | X | | | X | |
| MSR Paraphrase Corpus | EN | | | X | | |
| CLEF AVE | EN | | | | | X |
| Kurohashi Lab's | JA | X | | (X) | | |
| **NTCIR-9 RITE** | **JA CS CT** | **X** | **X** | **X** | **X** | **X** |
| SemEval-2012 CLTE | Cross-lingual | X | X | X | | |

## Formal Run Participation

Number of submitted runs

| Subtask | Language | | | Total |
|---|---|---|---|---|
| | JA | CS | CT | |
| BC | 24 | 33 | 32 | **89** |
| MC | 10 | 27 | 22 | **59** |
| Entrance Exam | 18 | - | - | **18** |
| RITE4QA | 13 | 17 | 16 | **46** |
| **Total** | **65** | **77** | **70** | **212** |

- 24 active participants from 5 countries



## Result Highlights (BC)

JA

| Run | Accuracy |
|---|---|
| JAIST-01 | 0.5800 |
| JAIST-02 | 0.5660 |
| JAIST-03 | 0.5520 |
| NTTCS-03 | 0.5480 |
| LTI-03 | 0.5460 |
| LTI-02 | 0.5420 |
| LTI-01 | 0.5340 |
| NTTCS-01 | 0.5320 |
| IBM-02 | 0.5260 |
| FX-02 | 0.5240 |
| *Average* | *0.5233* |
| *Baseline (char overlap)* | *0.5160* |

CS

| Run | Accuracy |
|---|---|
| UIOWA-01 | *0.9705 |
| UIOWA-03 | *0.9631 |
| UIOWA-02 | *0.9361 |
| ICRC_HITSZ-03 | 0.7764 |
| FudanNLP-02 | 0.7617 |
| ICRC_HITSZ-02 | 0.7568 |
| FudanNLP-01 | 0.7469 |
| WHUTE-03 | 0.7371 |
| NTU-01 | 0.7346 |
| WHUTE-02 | 0.7322 |
| WUST-01 | 0.7248 |
| NTU-02 | 0.7224 |
| NTU-03 | 0.7199 |
| ZSWSL-01 | 0.7199 |
| IASLD-01 | 0.7150 |
| ICL-01 | 0.7150 |
| *Average* | *0.7135* |
| *Baseline (char overlap)* | *0.7617* |

CT

| Run | Accuracy |
|---|---|
| UIOWA-01 | *0.9078 |
| UIOWA-02 | *0.8844 |
| IASLD-03 | 0.6611 |
| IASLD-02 | 0.6533 |
| III_CYUT_NTHU-02 | 0.6500 |
| IASLD-01 | 0.6478 |
| NTOUA-02 | 0.6422 |
| *Average* | *0.6222* |
| *Baseline (char overlap)* | *0.6667* |

* UIOWA Systems contain manual intervention (not fully automatic).

Showing runs above the average.



## Summary of Features Explored

- Overlap (character, word, bigram, trigram, head-word, POS, NE, numerical expression)
- String Similarity (Jaro distance, Jaro–Winkler distance, Jaccard Coefficient, Chebyshev Distance, Dice Coefficient, Manhattan Distance, Longest Common Subsequence, Cosine similarity, Levenshtein Edit Distance, BLEU score)
- Structural matching (predicate-argument matching, subtree matching, Tree Edit Distance)
- Verbs number mismatch
- Antonyms
- Negation / Polarity matching
- Temporal matching (5% improvement in EXAM [IBM])
- Quantification (*all*, *only*, *most* , *every*…)
- Quote (something just said might not be true…)

: : :



## Conclusion

- Best runs were able to outperform the strong character-overlap baseline
- Diverse techniques were explored – e.g. supervised machine learning, crowdsource-driven rule-based approach, predicate-argument matching, bilingual enrichment, LFG-based inference etc.
- Simple core challenge allowed participants to focus on developing textual entailment components that are potentially applicable to various IA problems
- Fast automatic evaluation enabled participants to report additional experimental results (e.g. ablation study).
- Attracted many participants including new comers as a first NTCIR task – indicating there's a research need.

**RITE was successful as a first attempt in NTCIR!**



## Result Highlights (MC)

JA

| Run | Accuracy |
|---|---|
| IBM-02 | 0.5114 |
| KYOTO-03 | 0.4841 |
| KYOTO-02 | 0.4795 |
| IBM-01 | 0.4545 |
| NTTCS-03 | 0.4523 |
| NTTCS-01 | 0.4477 |
| IBM-03 | 0.4455 |
| *Average* | *0.4124* |
| *Baseline (char overlap)* | *0.4682* |

CS

| Run | Accuracy |
|---|---|
| UIOWA-01 | *0.8919 |
| UIOWA-02 | *0.8919 |
| UIOWA-03 | *0.8870 |
| ICRC_HITSZ-03 | 0.6413 |
| ICRC_HITSZ-02 | 0.6241 |
| ZSWSL-02 | 0.6192 |
| WHUTE-02 | 0.6093 |
| *Average* | *0.5971* |
| *Baseline (char overlap)* | *0.5315* |

CT

| Run | Accuracy |
|---|---|
| UIOWA-01 | *0.7867 |
| UIOWA-02 | *0.7744 |
| UIOWA-03 | *0.7244 |
| MCU-01 | 0.5356 |
| IMTKU-01 | 0.5222 |
| IMTKU-02 | 0.5067 |
| *Average* | *0.5019* |
| *Baseline (char overlap)* | *0.4885* |

* UIOWA Systems contain manual intervention (not fully automatic).



## Result Highlights (EXAM)

JA

| Run | Accuracy |
|---|---|
| IBM-01 | 0.7217 |
| TU-02 | 0.7183 |
| TU-03 | 0.7042 |
| IBM-02 | 0.6742 |
| LTI-03 | 0.6674 |
| KYOTO-02 | 0.6561 |
| KYOTO-03 | 0.6561 |
| LTI-02 | 0.6538 |
| JAIST-02 | 0.6516 |
| JAIST-03 | 0.6516 |
| TU-01 | 0.6493 |
| JAIST-01 | 0.6222 |
| LTI-01 | 0.6018 |
| KYOTO-01 | 0.5928 |
| *Average* | *0.5863* |
| *Baseline (char overlap)* | *0.6516* |



## Proposal of NTCIR-10 RITE-2

(Recognizing Inference in TExt)



## NTCIR-9 RITE Subtasks

- **Binary-class (BC) subtask** - Given a text pair ($t_1$,$t_2$) identify whether $t_1$ entails (infers) a hypothesis $t_2$ or not.
- **Multi-class (MC) subtask** - 5-way labeling task to detect (forward / reverse / bidirection) entailment or no entailment (contradiction / independence) in a text pair.
- **Entrance Exam subtask** (Japanese Only) - Same as the BC subtask in terms of input and output. All the data is created from actual college-level entrance exams.
- **RITE4QA subtask** - Same as the BC subtask in terms of input and output, but as an embedded answer validation component in Question Answering system. This way, the impact of RITE to an overall end-to-end application can be measured.



## Result Highlights (RITE4QA)

JA

| Run | Acc | MRR |
|---|---|---|
| LTI-03 | 0.6753 | 0.2982 |
| JAIST-01 | 0.5602 | 0.2765 |
| JAIST-03 | 0.6940 | 0.2731 |
| JAIST-02 | 0.6763 | 0.2604 |
| LTI-02 | 0.6411 | 0.2563 |
| JUCS-01 | 0.5954 | 0.2490 |
| *Average* | *0.6148* | *0.2424* |
| *Baseline1 (char overlap)* | *0.4180* | *0.3192* |
| *Baseline2 (all yes)* | *0.1100* | *0.1657* |
| *Baseline3 (random)* | *0.5000* | *0.2320* |
| *Baseline4 (QA system)* | *0.1100* | *0.3917* |
| *Oracle* | *1.0000* | *0.5326* |

CS

| Run | Acc | MRR |
|---|---|---|
| UIOWA-01 | *0.9010 | 0.4272 |
| IMTKU-02 | 0.4090 | 0.3998 |
| WHUTE-02 | 0.4876 | 0.3979 |
| WHUTE-01 | 0.3886 | 0.3773 |
| IMTKU-03 | 0.4716 | 0.3768 |
| IMTKU-01 | 0.3319 | 0.3744 |
| ICL-01 | 0.3231 | 0.3545 |
| ICRC_HITSZ-01 | 0.6390 | 0.3520 |
| WHUTE-03 | 0.3275 | 0.3494 |
| ICRC_HITSZ-03 | 0.7293 | 0.3398 |
| *Average* | *0.5192* | *0.3367* |

CT

| Run | Acc | MRR |
|---|---|---|
| UIOWA-01 | *0.9010 | 0.4272 |
| IMTKU-03 | 0.4003 | 0.3992 |
| NTOUA-03 | 0.6346 | 0.3824 |
| NTOUA-01 | 0.5459 | 0.3803 |
| IMTKU-01 | 0.3246 | 0.3772 |
| IMTKU-02 | 0.3392 | 0.3736 |
| NTOUA-02 | 0.5124 | 0.3572 |
| ICRC_HITSZ-01 | 0.6390 | 0.3520 |
| ICRC_HITSZ-03 | 0.7293 | 0.3398 |
| *Average* | *0.5514* | *0.3352* |
| *Baseline1 (char overlap)* | *0.2317* | *0.3844* |
| *Baseline2 (all yes)* | *0.1906* | *0.2378* |
| *Baseline3 (random)* | *0.5000* | *0.3454* |
| *Baseline4 (QA system)* | *0.1906* | *0.4852* |
| *Oracle* | *1.0000* | *0.5906* |

* UIOWA Systems contain manual intervention (not fully automatic).



## Summary of Ideas Explored

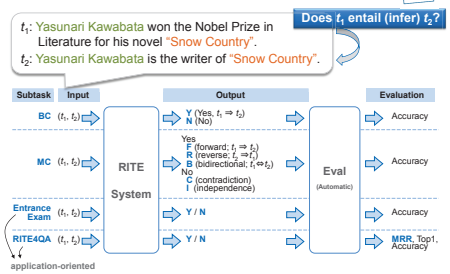
- Machine learning [many teams]
- Predicate-argument matching [KYOTO, LTI, NTTCS, SITLP, WHUTE, ZSWSL]
- Bilingual enrichment [JAIST, JUCS]
- Crowdsource-driven rule-based approach [UIOWA]
- Inference on Lexical Functional Grammar [FX]
- Alignment [TU]

**Ideas NOT Explored…**

- Monolingual Machine Translation



## RITE (Recognizing Inference in TExt)






## NTCIR-10 RITE Subtasks

- **BC Subtask**
  - Additional label / subsubtask for a specific entailment relation (e.g. syntactic, lexical, world-knowledge, inference)
- **MC Subtask**
  - No more entailment direction
  - Bidirectional, Contradiction, Independence

Application Oriented Subtasks

- **Entrance Exam Subtask**
- **RITE4QA Subtask**

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

## 7. SpokenDoc

音声・画像・ビデオの記録・編集機器の拡大，およびインターネットをはじめとする情報通信網の発展により，誰でも気軽にコンテンツを作成・公開することが可能となり，マルチメディアコンテンツの増大が加速している．これらのコンテンツには，ファイル名やタイトル以外にはメタデータが付与されていないことが多く，従来のテキストベースの検索技術だけでは，目的のコンテンツにたどり着くことは困難である．一方，音声を含むコンテンツの場合には，大語彙連続音声認識技術を利用することで言語情報を抽出し，テキスト検索技術を利用した検索が可能である．このような音声言語情報を対象とした検索技術は「音声ドキュメント検索」と呼ばれ，マルチメディアコンテンツの情報爆発時代に必要不可欠な技術になると考えられる[1]．NTCIR-9 音声ドキュメント検索タスク “IR for Spoken Documents (SpokenDoc)”[2] では，実際の検索環境に近い条件 (自由発話音声を対象，未知語を含む検索課題) における共通タスクを設定し，日本で最初の大規模な音声ドキュメント検索タスクの評価を行った．

### 7.1 タスク概要

NTCIR-9 SpokenDoc では，次の 2 つのサブタスクを行った．

**Spoken Term Detection (STD)** 語をクエリとして与え，音声ドキュメント中からクエリが現れる位置を特定するタスク．計算効率 (索引に必要な空間コスト，検索時間コスト，など) と検索性能 (精度と再現率) の 2 つの観点から評価を行う．

**Spoken Document Retrieval (SDR)** 自然言語文によるクエリを与え，クエリと関連するパッセージあるいは講演を見つけるタスク．テキストを対象とした検索における内容検索に相当するが，検索対象が音声データである点が異なる．

### 7.2 音声ドキュメント

SpokenDoc タスクでは，国立国語研究所から公開されている「日本語話し言葉コーパス (Corpus of Spontaneous Japanese; CSJ)」[7] を音声ドキュメントとして利用した．CSJ に含まれるデータのうち，学会講演および模擬講演を検索対象文書とする．両講演データを合わせると，2702 講演となる．STD サブタスクでは，コアと呼ばれるサブセット 177 講演を対象とした評価も行う．CSJ の各講演は 200 ミリ秒以上のポーズで分割された転記基本単位 (Inter Pausal Unit; IPU) で書き起こしが行われている．この IPU を正解判定の基本単位として用い，正解 IPU を見つけるタスクとして STD および SDR サブタスクを定義した．

### 7.3 書き起し

オーガナイザは，参加者共通で利用できる，単語ベース音声認識による認識結果，および，音節ベース音声認識による認識結果，の 2 種類の CSJ 自動書き起しを提供した．これによって，参加者は自前で音声認識を行わなくてもタスクに参加可能となる．また，共通の認識性能のもとで，参加者の検索手法の比較が可能になる．一方，認識率の改善に焦点を当てる参加者は，自前の音声認識システムを用いて書き起しを作成することも可能とした．

### 7.4 STD サブタスク

#### 7.4.1 検索課題と結果提出

検索課題として，全講演用，コア講演用の 2 種類のクエリ語リストを提供した．課題数は，それぞれ 50 語である．参加者が提出する 1 つの結果提出 (run) には，検出結果およびシステムが検出に要した周辺情報を記述する．検出結果は，クエリ語リスト中のクエリ語毎に，検出候補のリストを指定する．検出候補は，文書 ID，IPU の ID，検出の尤らしさを表すスコア，最終的な検出結果として出力するか否かのバイナリフラグ (YES または NO)，の 4 つ組で指定する．周辺情報は，オフライン処理 (索引付け等) およびオフライン処理 (検索処理) それぞれについて，使用したマシンスペック，処理に要した時間，空間コスト (索引のサイズ，メモリ使用量) を記述する．

#### 7.4.2 評価指標

検出した IPU を単位とした Recall と Precision をクエリ毎に算出し，全クエリで平均した値を基本とした評価指標を用いる．バイナリフラグで参加者が指定した検出での F-measure，F-measure が最大になるしきい値での F-measure，などが分析のために利用される．

#### 7.4.3 評価結果

STD サブタスクには 7 チームが参加した．全グループがコア講演を対象としたタスクに参加し，合計 12 の run が提出された．全講演を対象としたタスクには 2 チームが参加し，合計 4 つの run が提出された．コア講演に対する提出結果の Recall-Precision 曲線を，**図 8** に示す．タスク参加者は多様な視点でタスクに臨んでいるので，このグラフのみから結果を読み取るのは難しい．例えば，12 のうち 5 つの run は音声認識にも独自の工夫を行い自ら書き起こしを生成しているが，他の run はオーガナイザ提供のリファイレンス書き起こしのみを用いて検索手法に焦点を当てた評価を行っている．また，5 つの run は索引を使った検出の高速化を実装しており，他の run よりも数千倍程度の速度向上に成功している．一方で，純粋に性能の観点から見た場合，高い検出性能を示した run は，共通して複数の音声認識結果を使用する手法を用いていた．例えば，図 8 で最も高い性能を示した run は，合

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

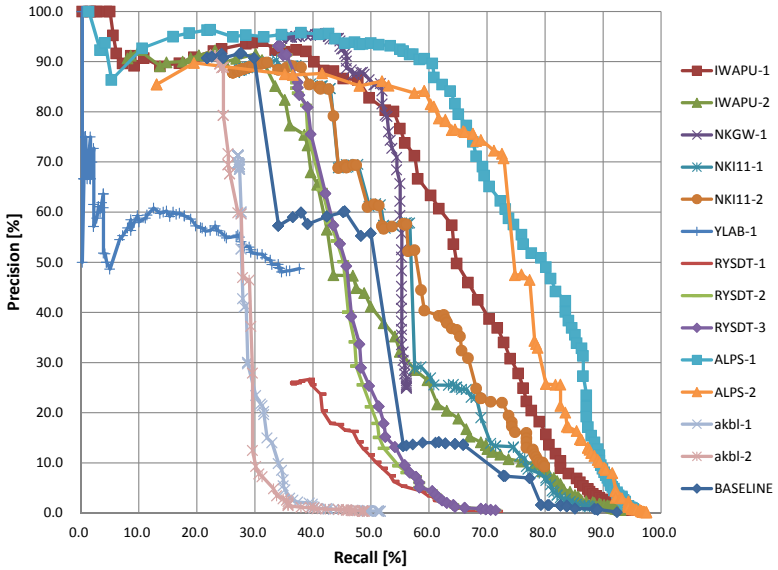



**図 8** Recall-precision curves for the CORE query sets.

計 10 種類の認識結果を利用して高精度かつ高再現率の検出を実現している．

### 7.5 SDR サブタスク

#### 7.5.1 検索課題と結果提出

86 件検索課題を提供した．これらは，講演の一部に含まれる内容を探す質問文である．これらの検索課題に対して，次の 2 種類のタスクを設定した．

**講演検索タスク** 正解区間が含まれる講演を見つけるタスク．

**パッセージ検索タスク** 正解区間そのものを見つけるタスク．

参加者が提出する 1 つの結果提出 (run) には，検出結果およびシステムが検出に要した周辺情報を記述する．検出結果は，講演検索タスクでは講演，パッセージ検索タスクでは可変長パッセージを，それぞれ順序づけて 1000 件を上限に記述する．周辺情報は，STD サブタスクと同様である．

#### 7.5.2 評 価 指 標

適合性判定により，各クエリトピックについて，可変長区間 (IPU 列) の集合が正解として与えられる．講演検索タスクについては，可変長区間が含まれる講演を正解文書とみなして Recall および Precision を計算する．評価尺度としては，Mean Average Precision(MAP) を用いる．一方，パッセージ検索タスクでは正解文書が可変長区間となるため，文書検索タスクで用いられる評価指標がそのまま利用できない．そのため，MAP をベースにした新たな 3 つの評価指標，uMAP，pwMAP，fMAP を設計した．[2)]

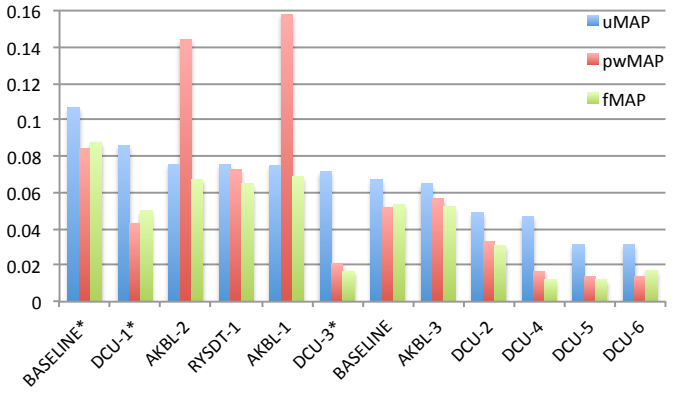



**図 9** Evaluation results for the passage retrieval task.

#### 7.5.3 評 価 結 果

SDR サブタスクには 5 チームが参加し 21 の run が提出された．このうち，講演検索タスクには 4 チームが参加し 11 の run が提出された．パッセージ検索タスクには 3 チームが参加し，10 の run が提出された．パッセージ検索タスクの評価結果を，**図 9** に示す．図中の* は人手書き起こしを用いた run を示す．ベースライン手法は，あらかじめ固定区間 (15IPU) に分割した講演を検索対象に，単語認識の 1-best 候補に対して典型的なテキストベース検索手法を単純に適用したものである．幾つかの run はベースラインの性能を改善することに成功している．また特定の評価指標について大幅に性能を完全できた run も存在する．しかし，全体的に評価指標の値は低く，本タスクの難しさが示されており，課題が残されている．NTCIR-10 で実施する SpokenDoc-2 では，パッセージ検索タスクについてより詳しく評価を行う予定である．

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

## 8. VisEx

パイロットタスクとして実施した VisEx の目的は，対話的・探索的な情報アクセスを支援する環境を評価するための方法論を構築することにある．

**図 10** に VisEx における評価の枠組みを示す．VisEx においては，評価される情報アクセス環境は図の中央に示した構成を持つものと仮定される．この環境のうち，核部が参加者によって作成され，提出される．それ以外の外枠部分はオーガナイザより提供され，すべての参加者で共有される．そこには，実際の検索を行う情報検索エンジンや集められた知識の編集や記録に用いられるエディタ部分が含まれる．更に，全体が Web ブラウザの下で動作し，すべてのインタラクションがブラウザ経由で行われる．情報アクセス核部は，これら外枠の機能を有機的に組み合わせて，情報アクセス環境を構成する．その役割は，利用者の情報ニーズを受け取って情報検索エンジンに送り，その結果得られた情報をわかりやすく，課題の達成を容易にする形で利用者に提示する等して，情報アクセス行為を支援することにある．このような核部のひとつとして，オーガナイザも比較のために用いるベースラインを提供している．

このような構成を仮定することで．

- 情報アクセス行為全体，つまり，情報を集めるだけでなくそれを知識としてまとめ上げる部分までを観察しつつ，そのような広い行為の観察に伴う揺れやノイズをできるだけ少なくする，
- 情報アクセス環境を用いる利用者の振る舞いに加えて，その中の構成要素間のやりとりについても統一的かつ詳細なデータを取得する，

ことを目指している．

提出された核部を含んだ情報アクセス環境が共通の課題を用いた被験者実験によって評価される．その課題は，与えられた情報アクセス環境を用いて，与えられたトピックに関して適当と判断される出来事や事実を収集しレポートにまとめるというもので，「アジアでの航空機墜落事故」等，指定された出来事の特徴を収集しまとめるというイベント収集課題と，「ガソリンを巡る状況」等，統計量の変化（動向）とその原因や影響を要約するというトレンド要約課題の 2 種類が実施された．それぞれの課題は練習用のトピック 1 とと 4 つのトピックからなる．文書集合は毎日新聞の 1998-2001 年の記事を利用した．出来事や事態に関する情報が複数記事にまたがり，それらをまとめる等，集約と統合という側面が現れることや．10 年前の記事ということで，予想しなかった事実が見つかったり，新しい解釈が生まれる等があり，被験者の振る舞いに変化がおこることを期待した．

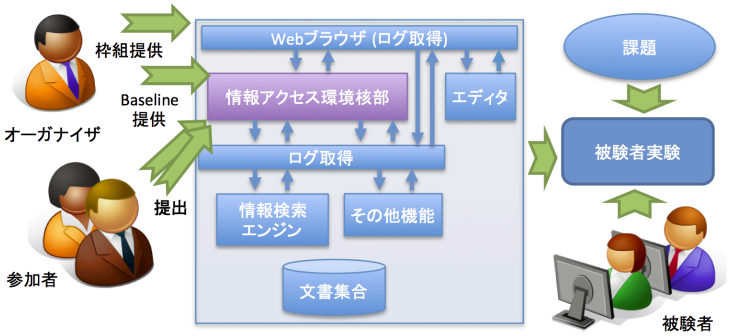



**図 10** VisEx の枠組み

今回の実施では，4 チームが参加し，ベースラインと合わせて 8 ユニット（システム×課題）の実験を実施した．各参加チームのシステムの特徴を以下に示す．

- 観点の異なるキーワードを 2 次元の並びで受け取り，キーワードの交差する格子に検索結果を配置する．
- 統計情報のグラフをインタフェースとし，グラフ上の関心のある点を指定することで関連する記事が検索できる．
- 既読記事にマークを付与することができ，新たに検索された記事をそれらと区別できる．
- 出版日と関連する地名による絞り込み検索と指定した記事と類似したものを探す類似検索を許す．

実験を通じて，情報アクセスの様々な側面を表わす以下のデータが取得できた．

- 被験者によって作成されたレポート（情報アクセスの成果物）
- 被験者の行為やシステムの動作を記録した操作ログ（情報アクセスの動的なプロセス）
- 被験者へのアンケート結果（情報アクセスを行った主体が持った印象）

これらのデータが統一的に取得できる点が有益で，その分析を通じて，参加チームは，特に参加システムとベースラインとを比較することで，多くの有益なフィードバックを得たことを報告している．一方で，得られた知見の一般化はなかなか困難で，より詳細な分析と洗練された追加実験が必要であるとの感触を得ている．特に，タスクの設計では，課題をもう少し複雑なものとすること，実験デザインとして，被験者間のばらつきを吸収する仕組みを導入することが重要である．

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

## 9. Math

### 9.1 タスクの背景と目的

数式は，汎用的な知識の表現方法であり，科学的な情報の伝達において重要な役割を果たす．たとえば科学文献中の数式は，単に数値計算だけではなく，定性的な性質を論じたり，自然言語による説明のあいまい性を補足してモデルや手法を厳密に定義したりするためにも用いられる．しかし，その重要性にもかかわらず，情報検索の分野では数式コンテンツのアクセスに関する研究はあまり行われて来なかった．その大きな原因として，多くの電子文書では数式が画像や文字の位置情報として貼り付けられており，数式の構造情報が機械可読な形で提供されていないことがあげられる．そこで NTCIR-10 Math Task では，評価用データセットの構築やタスクの設計を通して，数学的コンテンツへのアクセス技術に関する研究の推進やコミュニティの形成に寄与するとともに，将来的には電子出版における数式電子化の促進に結び付けることを目指す．

### 9.2 タスクの概要

本タスクでは，以下の２つのサブタスクを設定して，データセット構築および課題設計を行う予定である．タブタスク A はアドホック型の数式検索タスクであり，与えられた文書集合の中からユーザの検索要求に適合する数式または数学的記述を見つける．サブタスク B は数学的概念を対象とした情報抽出タスクであり，文書中の数学記号や数式の定義表現を抽出する．検索対象はウェブ上で公開されている科学技術文書で，サブタスク A では 10,000 文書，サブタスク B では 100 文書程度のデータセットを想定している．文書中の数式は原則として，W3C が定める標準である MathML の記法にしたがって表現するものとし，latex ソースが利用可能であるものについては，あわせて latex ソースへの URL も提供する．タスクの詳細については順次，`http://ntcir-math.nii.ac.jp/` でアナウンスする予定である．オーガナイザは Akiko Aizawa (NII)，Michael Kohlhase (Jacobs University Bremen)，Iadh Ounis (University of Glasgow) の３名，アドバイザは Noriko Kando (NII) である．参加者からのフィードバックも歓迎している．

### 9.3 タスクの展望

数式は抽象的な知識表現であるため，数学的知識のアクセスでは非言語オブジェクトを周辺テキストの言語要素と対応づけることが必要であり，数式および言語の意味解析の有効性が期待される．ここで，コンピュータを用いた数学知識処理の研究分野では，伝統的に数学文献の電子化や検索に関する研究が行われており，その中で数式 OCR，ユニバーサルアクセス，数学の基本オントロジー定義などの技術が培われている．一方，情報検索や自然言語処理の研究分野においては数式に特化したタスク設定は新しい試みであり，NTCIR Math Track はパイロットタスクとして，双方のコミュニティの橋渡しと融合を目指す．なお，関連するワークショップとして，オーガナイザの１人である Michael Kohlhase らが主催する Math IR Symposium が，2012 年 7 月に Conferences on Intelligent Computer Mathematics (CICM 2012) にて開催される予定である．

Vol.2012-IFAT-106 No.5
Vol.2012-DD-85 No.5
2012/3/26

## 10. ま と め

本稿では，NTCIR-9 の簡単な総括を行うとともに，NTCIR-10 の新たな取り組みについて紹介した．是非ご興味のあるタスクに参加いただき，国内外の多くの研究機関とともに情報アクセス研究を盛り上げていただきたい．タスク参加要領，その他の最新情報は随時 NTCIR-10 ホームページ[*1]に掲載するのでご覧いただきたい．

**謝辞** NTCIR 参加者および関係者，特に，NTCIR-9 のタスクオーガナイザを務められた Fredric Gey, Ray R. Larson, Jorge Machado, Ruihua Song, Min Zhang, Makoto P. Kato, Yiqun Liu, Miho Sugimoto, Qinglei Wang, Naoki Orii, Kiyoaki Aikawa, Tatsuya Kawahara, Tomoko Matsui, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Yusuke Miyao, Shuming Shi, Koichi Takeda, Kelly Y. Itakura, Mitsunori Matsushita, Bin Lu, Ka Po Chow, Eiichiro Sumita, Benjamin K. Tsou 諸氏に感謝する．

## 参 考 文 献

1) 秋葉 友良: 音声ドキュメント検索の現状と課題，情報処理学会研究報告，Vol.2010-SLP-82, No.10 (2010).
2) Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T. and Matsui, T.: Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop, NTCIR-9, pp.223-235 (2011) http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-SpokenDoc-AkibaT.pdf
3) Gey, F., Larson, R. R., Machado, J. and Yoshioka, M.: NTCIR9-GeoTime Overview - Evaluating Geographic and Temporal Search: Round 2, NTCIR-9, pp.9-17 (2011) http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-GEOTIME-GeyF.pdf
4) *GIR '10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, New York, NY, USA, ACM. (2010)
5) Goto, I., Lu, B., Chow, K. P., Sumita, E. and Tsou, B. K.: Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, NTCIR-9, pp.559-578 (2011). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-PATENTMT-GotoI.pdf
6) Kato, T., Matsushita, M. and Joho, H.: Overview of the VisEx task at NTCIR-9, NTCIR-9, pp.524-532 (2011). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-VisEx-KatoT.pdf
7) Maekawa, K., Koiso, H., Furui, S. and Isahara, H.: Spontaneous Speech Corpus of Japanese, *Proceedings of International Conference on Language Resources and Evaluation*, pp.947–952 (2000).
8) 酒井 哲也: 曖昧なクエリと (不) 明快なクエリ: NTCIR-10 INTENT-2 と 1CLICK-2 タスクへの誘い, 情報処理学会研究報告 2012-IFAT-106 / 2012-DD-85 (2012).
9) Sakai, T., Joho, H.: Overview of NTCIR-9, NTCIR-9 Proceedings, pp.1-7 (2011). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-SakaiT.pdf
10) Sakai, T., Kato, M.P. and Song, Y.-I.: Overview of NTCIR-9 1CLICK, NTCIR-9 Proceedings, pp.180-201 (2011). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-1CLICK-SakaiT.pdf
11) Shima, H., Kanayama, H. Lee, C.-W., Lin, C.-J., Mitamura, T., Miyao, Y., Shi, S. and Takeda, K.: Overview of NTCIR-9 RITE: Recognizing Inference in TExt, NTCIR-9, pp.291-301 (2011) http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-RITE-ShimaH.pdf
12) Song, R., Zhang, M., Sakai, T., Kato, M.P., Liu, Y., Sugimoto, M., Wang, Q. and Orii, N.: Overview of the NTCIR-9 INTENT Task, NTCIR-9 Proceedings, pp.82-105 (2011). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-INTENT-SongR.pdf
13) Tang., L.-X., Geva, S., Trotman, A., Xu, Y. and Itakura, K. Y.: Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery, NTCIR-9 Proceedings, pp.437-463 (2011) http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-CROSSLINK-TangL.pdf

*1 http://research.nii.ac.jp/ntcir/ntcir-10/index.html