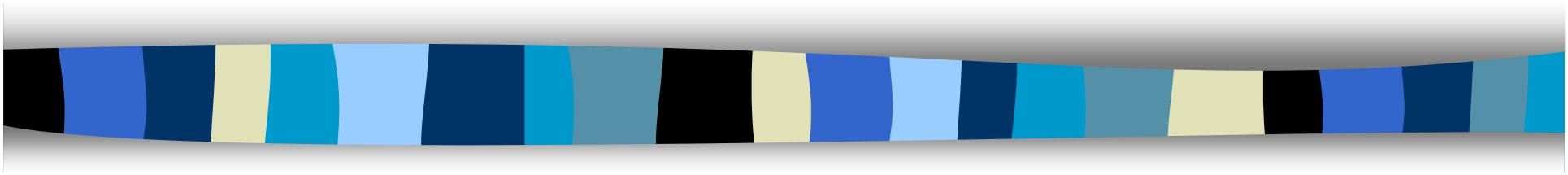
NTCIR GeoTimeタスク



北海道大学　　吉岡真治

タスク概要

- 背景
  - 地理情報を利用した情報アクセスサービスの増加
  - 地理情報に関する情報検索の研究
    - 主に欧米の言語で行われている。
- 地理時間情報を対象とした情報検索
  - 日本語・英語の二言語による情報検索
  - 時間情報を考慮

文書データ

- 文書集合：新聞記事
  - 日本語：毎日新聞
  - 英語： New York Times (NYT), Mainichi News English Edition, Korea Times English edition, Xinhua China News Service English Edition

Collection	Language	Time Period	# Documents	
毎日新聞	J	2002-2005	377,941	NTCIR-8
NY Times	E	2002-2005	315,417	NTCIR-9
毎日新聞	J	1998-2001	419,759	
Korea Times	E	1998-2001	50,129	
Mainichi	E	1998-2001	24,878	NTCIR-9
Xinhua	E	1998-2001	406,792	

1,594,916 total documents
797,700 for Japanese
797,216 for English

検索質問の作成と質問の例

- 参加者が検索質問の作成に参加
  - 基本的な検索システムを用い、日英の文書集合中に適合文書があることを確認
  - 答えに関する情報（例：Wikipediaのページ）の提供
- 質問の例
  - DESCRIPTION：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：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

必要とされる技術

- 地理情報
  - 地理的な包含関係や座標の情報を扱う必要がある。
  - 例:
    「アフリカのどこ」→「ナイジェリア」
    「5° 52′12″N 5° 45′00″E / 5.870° N 5.750° E / 5.870; 5.750」→「ナイジェリアのJesse」
- 時間情報
  - 正規化:「明日」、「来年」、「先月」…
  - ある時点での情報:「2000年1月1日」現在で、最も長い橋

正解判定と評価

- 多段階の判定
  - 完全一致
    - 全ての解答が含まれている
  - 部分一致
    - 一部の解答（地名のみ、時間のみ）
    - 不十分な情報
- 多段階判定を使った評価
  - **Average Precision (AP)**
  - **Q Measure**
  - **normalized Discounted Cumulative Gain (nDCG)**

タスク参加者数と参加チーム(NTCIR-9)

- 日英単言語と日→英、英→日の2方向の言語横断タスクを設定

	英→英	英→日	日→日	日→英
NTCIR-8	6	3	8	1
NTCIR-9	7	2	6	1

BRKLY	**University of California, Berkeley***
GETUA	**University of Alicante, Spain**
INESC	**National Institute of Electroniques & Computer Systems, Lisbon, Portugal ***
HU-KB	**Hokkaido University, Japan***
IRNLP	**Korea Advanced Inst. for Science & Technology**
KOLIS	**Keio University, Library Science**
NAK	**Keio University, Science and Technology**
OKSAT	**Osaka Kyoiku University, Japan**
RMIT	**Royal Melbourne Institute of Technology**
SINAI	**University of Jaén, Spain**
SJTUB	**Shanghai Jiao Tong University, China**
UIOWA	**University of Iowa, USA**

利用する外部リソース

- 地名情報に関して、文書中からの抽出を行うために外部リソースを利用
  - Wikipedia, DBpedia, Geonames, Alexandria Digital Library and Yahoo! PlaceMaker
- NLP技術の応用
  - Semantic role labeling
  - 固有名抽出
- 質問応答の結果の利用
  - NTCIR-8：質問応答システムの結果を検索語拡張としてフィードバック
  - NTCIR-9：手作業で作成した結果を検索語拡張としてフィードバック

スコアリング手法

- 基本的な検索エンジンの結果からリランキング
  - Semantic roleの利用
    - Semantic roleに応じた様々な結果をランクアグリゲーションで統合
  - 固有名詞に注目
    - 質問に関連する固有名詞を含まない文書のスコアを修正
  - 解が含まれることが期待される文書に注目（地理時間表現が多い）
  - まとめの記事の方が、より、多くの情報を含んでいる（後の記事の方に重みを置く）

トピックの難易度

- 言語依存の部分
  - 背景知識の違い
    - コンゴ民主共和国がアフリカであることが日本人にとって常識であれば、アフリカのコンゴ民主共和国とは記述されない。
- 表現形式の問題
  - 質問が表現形式を変えることによって難易度が変わる可能性がある。
    - 一つのイベントを答える場合と、類似するイベントを列挙する問題
    - 表記のバリエーション：記事中での表現と質問での表現の対応がとれている場合
      →疑似適合文書フィードバックがうまく働く

結果(NTCIR-9:日本語)

RUN	AP	RUN	Q	RUN	nDCG
OKSAT-JA-JA-03-D*	0.6449‡	OKSAT-JA-JA-03-D*	0.6666	OKSAT-JA-JA-03-D*	0.8547
HU-KB-JA-JA-03-D	0.4490	HU-KB-JA-JA-03-D	0.4804	HU-KB-JA-JA-03-D	0.6630
KOLIS-JA-JA-05-D	0.4227	KOLIS-JA-JA-05-D	0.4540	KOLIS-JA-JA-05-D	0.6294
RMIT-JA-JA-01-D	0.3779	RMIT-JA-JA-01-D	0.4119	RMIT-JA-JA-01-D	0.6152
NAK-JA-JA-01-D	0.2928	NAK-JA-JA-01-D	0.3190	NAK-JA-JA-01-D	0.4936

•manual run (human interaction in query formulation)
‡ statistically significant difference (α=0.01) from the value of the run in the next row

結果（NTCIR-9:英語）

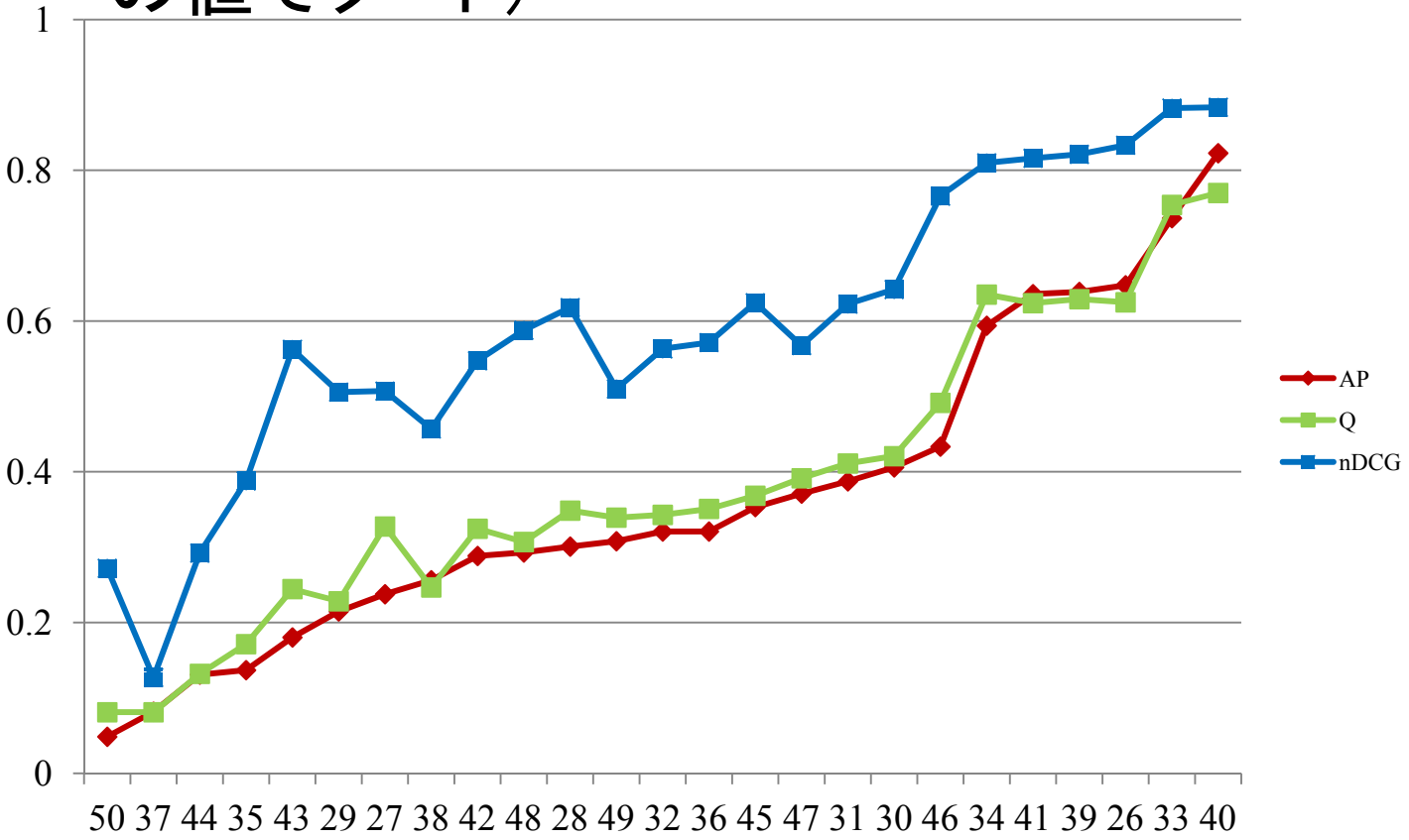
RUN	AP	RUN	Q	RUN	nDCG
OKSAT-EN-EN-03-D*	0.5376†	OKSAT-EN-EN-03-D*	0.5562	OKSAT-EN-EN-03-D*	0.7523
SINAIUJAEN-EN-EN-01-D	0.4341	SINAIUJAEN-EN-EN-01-D	0.4564	SINAIUJAEN-EN-EN-01-D	0.6587
UIOWA-EN-EN-01-D	0.4164	UIOWA-EN-EN-01-D	0.4372	UIOWA-EN-EN-01-D	0.6425
BRKLY-EN-EN-01-D	0.4066	BRKLY-EN-EN-01-D	0.4246	BRKLY-EN-EN-01-D	0.6012
INESCID-EN-EN-01-D	0.3260	INESCID-EN-EN-01-D	0.3497	INESCID-EN-EN-01-D	0.5791

•manual run (human interaction in query formulation)

† statistically significant difference (α=0.05) from the value of the run in the next row

トピックの難易度（英語）

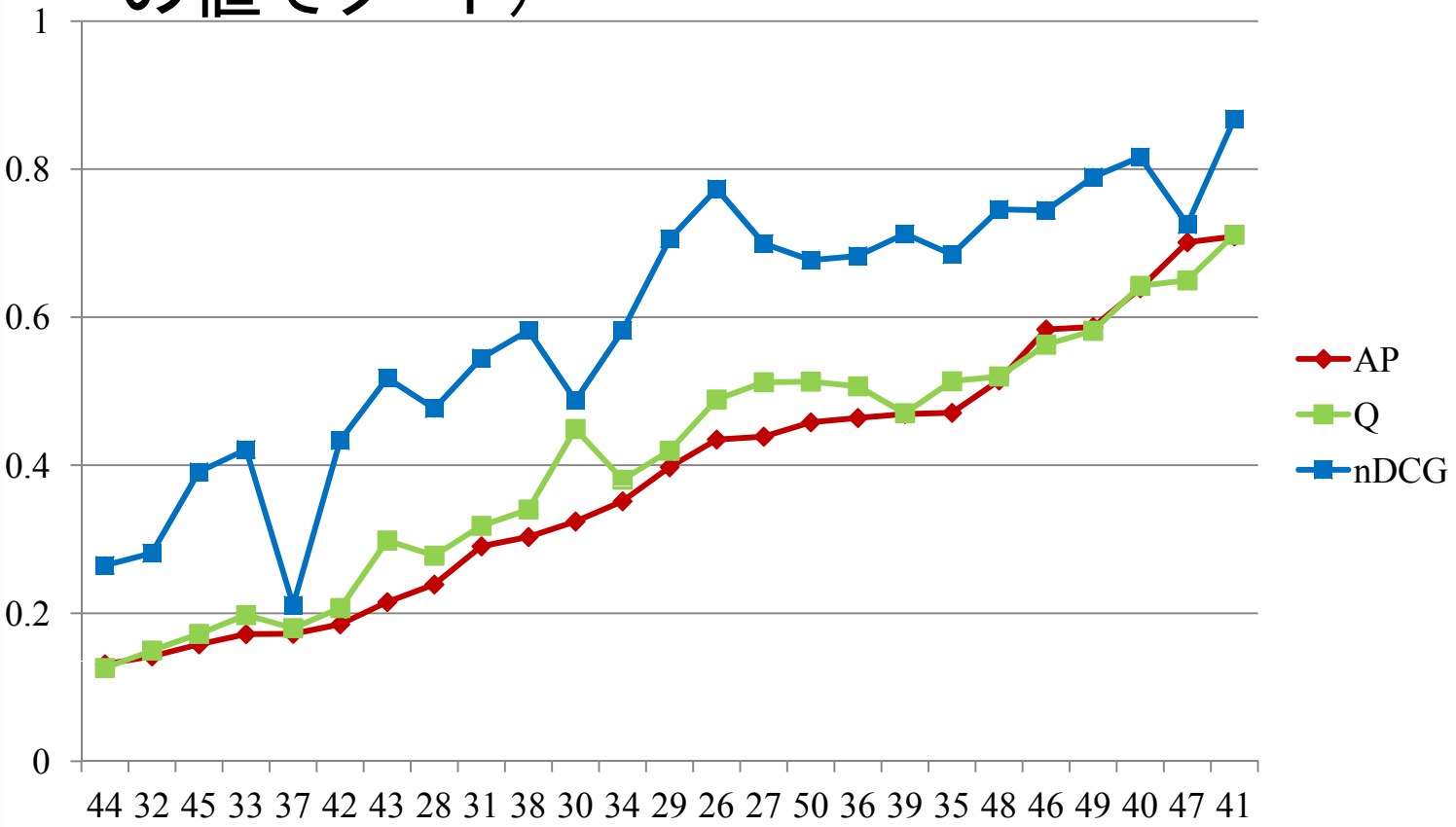
- 37のランについてのAP, Q and nDCG平均値（APの値でソート）



最も難しかったトピック 50:
いつ、どこで、中米自由貿易協定（CAFTA)は署名されましたか？

トピックの難易度（日本語）

- 31のランについてのAP, Q and nDCG平均値（APの値でソート）



最も難しかったトピック 44:
南アメリカで起きた死者が出た地震について、いつ、どこで起きたかを述べよ。

日英の差が大きな課題

- 日本語が難しかったトピック
  - 33：砒素の毒で4人が死亡し、多くの病人が出たのは、いつ、どこでですか？（日：22番、英：2番）
    - 原因：「砒素」が新聞中では、「ヒ素」と表記
  - 50:南アメリカで起きた死者が出た地震について、いつ、どこで起きたかを述べよ。（日：25番、英：10番）
    - 昔のチリ地震の津波に関する記事が多く存在
- 英語が難しかったトピック
  - 35:500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？（日：7番英：22番）
    - アフリカのパイプライン事故の記事は、日本ではあまり報道されず、適合となる重大事故のみが報道されていたため。

まとめ

- 地理時間情報に関する情報検索
  - 地理情報については、様々なオープンな地理情報の活用方法などが提案
    - ランクアグリゲーションなどのリランキングのテクニックとの融合
  - 時間情報については、まだ十分な提案がなされているとはいえない。
- 今後の課題
  - 難易度に与える影響の分析
  - TwitterやBlogなどを対象とする場合には、テキスト外の地理時間情報の利用法についても検討する必要がある。