

今後の展望： 情報検索評価プロジェクトの展開

神門 典子 国立情報学研究所 kando@nii.ac.jp

■第1回IREX/NTCIRから学んだこと■

第1回IREX/NTCIRは、多くの方々の協力と参加を得、大規模日本語テストコレクションと、情報・用語抽出の正解付きデータセットを構築し、多様な研究成果が報告され、多くの成果を収めることができた。

- 評議会には、図-1のような効果があるといわれている。IREXとNTCIRには次のような特徴もある。
- (1) 情報検索と、情報抽出・用語抽出の評価を1プロジェクトで
 - (2) 言語の特殊性：日本語、日英の言語横断と対訳パス
 - (3) 文書種類の多様性：新聞記事と学術文書
 - (4) 議論とコンセンサスの尊重
 - (5) IRの正解判定：3段階、検索目的、正解理由の明文化など
 - (6) テストコレクションの妥当性・信頼性の評価

IREXでは新聞の面を用いた検索、NTCIRでは専門用語、カタカナ語、原綴など学術文書で顕著な問題も議論された。小さな点であるが、検索タスクの正解判定は、より自然な判定を意図し、二値ではなく、適合、部分的適合、不適合の3段階とした。正解文書リストの網羅性、適合性判定のゆれのシステム評価への影響についても検討した。複数判定者間の適合性判定の一一致度は40～60%程度であるが、多数の検索課題を使用するので、テストコレクション全体としてはゆれが相殺され、どの判定を用いてもシステム間の順位付けは安定した結果が得られることが明らかにした☆1。国際会議の招待講演・パネルディスカッションへの招待は、NTCIRだけでも、昨年8月以降9回に及び、国際的研究コミュニティの承認という面でも一応の成果を得ることができたと思われる。

IREXとNTCIRは、2000年度から合同し、担当タスクの独自性を尊重しながら、より多面的な評議会議の推進を計画している。

■今後の方向性■

IREXとNTCIRが合同した拡大NTCIRプロジェクトは、研究コミュニティにどのような独自の貢献ができるのだろうか？ 昨年のIREX/NTCIR合同ワークショップにおける「評議会の今後」パネルディスカッションの冒頭で、図-2の提案がなされ⁵⁾、今後のさらなるチャレンジを促す力強いエールを多くの方からいただいた。それを受け、今後は、図-3のように、伝統的情報検索評価と、より新しいチャレンジという、2つの方向が重要であると考える。

■情報検索システムの伝統的評価■

情報検索では、検索実験によって精度・再現率を求める伝統的評価手法が、現状では"Gold Standard"である。これを満たすためには、テストコレクションなどの研究資源の整備、その構築手法・質の評価などに関する研究が必要である。この面では、日本語という言語の特殊性と言語横断検索に重点を置くことで独自性を発揮していきたい。

日本語検索では、テキストから索引単位を切り出すための語分割が1つの問題である。昨年の情報検索タスクでは、各チームがそれぞれ独自の語分割法と検索アルゴリズムを用いており、最終的な結果に何がどのように寄与しているかという相互比較が複雑であった。アルゴリズムと語分割のそれぞれについてのより深い検討が必要である。

■新たなチャレンジ1—WWW検索■

チャレンジの1つは、1, 3, 4編でも指摘されているよ

☆1 これは、1980年代にGerard Saltonが指摘し、TRECの実験で実証されている（Voorhees, E. M.: Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness, In Proceedings of ACM-SIGIR '98, pp.315-323）が、日本語テストコレクションでも確認することができた。

大規模テストコレクション・評価用データの提供：種類も長さも多様な文書。実用システムに匹敵
技術移転の促進：実用規模データでの評価実験を可能に
研究者のフォーラム。“Who is Who”環境の創生：研究者が互いに知り合い、同じ基盤で議論、意見交換を進める場を提供。単純な優劣の評価ではなく、各システムの特徴を明らかにするためベンチマーク。互いに学びあう場
研究データの蓄積：検索結果データは、データフュージョンなどの研究を促進
研究の動機づけ：特定テーマの研究を提案
研究のモデルの提示、評価手法の標準化

図-1 評価会議の効果

・キーワードから質問へ：情報要求を表現。多様な形式、あいまい性、文脈、対話型。言語非依存性（言語横断など）
・文書から回答へ：言語非依存性（言語横断、非言語情報 Yes-No）、自然言語処理技術など
・アクセスから処理へ：データマイニング、テキストマイニング、価値の付加、共有・協働など
・受動から能動へ：真の対話性、会話主導型検索、談話分析、文書種類に応じたテンプレートなど
・データからオーソリティへ：質、社会的信頼性

評価は？ 適合性、満足度、完結性、信頼性...
(Leong⁹⁹) より翻訳・整理)

図-2 情報検索システム研究の今後の方向

伝統的情報検索評価	独自のチャレンジ
言語の特殊性： 日本語・アジア言語	検索から利用支援へ：NLPとIRの融合
固有の問題 ・語分割	情報抽出、要約、翻訳、回答抽出、Q&A、複数文書対比・要約、テキストマイニング...
・日本語に適したアルゴリズムなど	
言語横断・多言語情報： 国際協力	現実に即した評価：
リソースの共有	対話型
リソースの構築	多様な文書：Web文書、特許、画像、音声...
	特にWWW検索
評価手法の議論	研究者のフォーラム コンセンサスの形成

図-3 NTCIRプロジェクトの今後

うに、より現実的な評価という問題である。現在、情報検索が直面している問題は、何であろうか。Web文書、社内文書、メモ、医療文書、特許など検索が必要な文書は多様である。多様な利用者像、対話型、多言語アクセスなどへの対応の強化が必要であろう。

なかでも、WWW検索は、多様な利用者、大量の文書と同時アクセス、更新頻度の高さ、リンク、マルチメディア、多様な文書タイプ、多言語など、多くの面で従来の情報検索システムと異なり、これを、どのように評価していくかは大きなチャレンジである。テキストだけでなく、HTMLタグ、ヘッダ、最終更新日、被参照回数、リンク、被リンク数などの何を用いるか。高速化、索引の圧縮などの効率面をどう評価するか。検索の成否の判定対象は、文書か、サイトか、リンクの提供か、あるいは回答そのものなのか。トピックとして適合していればよいのか、それとも利用者の検索目的や状況に応じた満足度、情報の入手可能性、信頼性、有用性なども視野に入れるべきか、などの多くの選択肢がある。

何をどのように評価したいのかということによって、どのような文書データを収集し、どのような利用者からどのような課題を収集するかが決まる。何を評価したいかについて、多くの方々からの意見と議論をいただければ幸いである。

他方、現実に即した文書や検索課題の収集には、著作権など微妙な問題もある。関連諸機関からの理解と協力を期待したい。

■新たなチャレンジ2—NLPとIRの融合■

もう1つのチャレンジは、自然言語処理と情報検索の融合である。自然言語処理の情報検索への応用が試みられてきたが、英語では、汎用的検索モデルの頑強さに比べ、根本的な検索性能改善を得られなかったといわれている。

情報検索は、情報ではなく、情報を含む可能性が高い文

書を検索する。しかし、利用者が対話型で直接操作する現実のシステムでは、今後は、文書の速読支援、要約、翻訳、回答や情報そのものの抽出、質問回答、複数文書から抽出した回答の対比、複数文書間の関係の分析（テキストマイニング）など、文書から情報を取り出し、分析し、その利用を支援する機能がますます重要になる。自然言語処理と情報検索の融合は、日本語の文書や検索質問からの語句の切出しなどに貢献できる可能性もあるが、それ以上に、このような情報利用支援の側面で大きな成果が期待される。

他方、情報抽出などのテキスト処理技術の評価は、従来、トピックと文書種類が限定された小規模な文書集合を用いてきた。今後は、情報検索が扱うような大規模で多様な文書集合への適用が課題となる。それには、まず、特定トピックの文書を検索し、あるいは、トピックと文書種類でフィルタリングしてから、詳細なテキスト処理をするというような検索と抽出技術の連携も重要になるだろう。IREXとNTCIRは、当初から情報検索と自然言語処理の評価を1つのプロジェクトで取り上げてきた。IREXとNTCIRの合同は、さらにこの方向を強めると期待される。

■評価手法の議論■

伝統的評価と新たなチャレンジを推進するための基盤は、研究者のフォーラムと、そこでの議論を通じたコンセンサスの形成である。

自然言語処理と情報検索の融合は、異なる評価の文化を持った集団の融合でもある¹⁾。たとえば、情報検索では、評価の基準はレレバנסス (relevance) ^{☆2}すなわち、検索された文書が、その利用者のその時点での情報要求に適合しているかという主観的な判断である。テストコレクションの適合性判定でも、最終判定はその検索課題を作った人(=利用者)である。評価は、要素技術の評価も含め、再現率と精度という検索有効性への効果を調べる。評価値は絶対値ではなく、同じコレクションを使用したシステム間

☆2「レレバансス」も、トピックとして客観的に捉えられるものから、心理的、状況に依存したもの、システムとのやりとりの中でシフトするものなど多様な観点がある。検索の成否の判定に影響を及ぼす要因をどこまで考慮するか判断を必要とする。他方、これらの要因を排除して、検索システムの最もコアとなる客観的なトピックとしての適合性によって検索性能を評価することもシステム開発の過程では必要とする立場もある。

```

<REC>
<ACCN>kaken-j-0924516300</ACCN>
<YEAR>1992</YEAR>
<SBJ1 TYPE="kanji">802: 情報学</SBJ1>
<PJNM TYPE="kanji">文献 の 論理_構造 に 基づく 全文_データベース_検索_システム の 開発_研究</PJNM>
<ABST TYPE="kanji"><ABST.P>本_研究 は 、 学術_文献 など の 文書 の 全文 を 収容 する 全文_データベース に ついて 、 それら の 文書 の 論理_構造 に 即した 検索 を 可能 と する システム を 研究 、 開発 しよう と する もの で ある 。 3_年次 に わたって 下記 の 項目 に ついて 研究 および 開発 を 行なった 。
</ABST.P><ABST.P> 1 . 全文_データベース に 対する 検索_要求 の 詳細_分析 を 行ない 、 SGML の 文書型_定義 に 基づいて 検索_表示_要求 を 効率的 に 記述 する ため の 表記_形式_DQL ( Document Query Language ) の 詳細_設計 を 行なった 。 SQL を 拡張 し 、 文書_構造 を 捉う ため の 記述 を 可能 に した 。 </ABST.P><ABST.P> 2 . 文献 の 文書_構造 を 圖形的 に 表示 し 、 要素 を ポイントティングデバイス で 指定 して 検索_条件_表示_指示 を 行なう ユーザ系 の ソフトウェア を 設計 し 、 ワークステーション 上 で グラフィカルユーザインターフェース

```

語分割は、語と語構成要素の2段階である。" "は語の区切り、"_"は語構成要素の区切りを示す。分割には平和情報センターのHappiness Ver.3.5を使用した。

図-4 語分割データの例

で、何%上回ったかという相対的評価として意味を持つ。それに対し、自然言語処理では、客観的な正解を求め、そのために定義についての議論やコンセンサスが特に重要である。要素技術の評価は単独の性能評価も可能であり、最終的な評価値は絶対値としての意味を持ち得るものである。

■第2回NTCIRワークショップ■

今年のNTCIRワークショップには次のタスクがある。

- (1) 中国語検索タスク：英・中言語横断検索を含む
- (2) 日本語・英語検索タスク：日英の単言語・言語横断検索
- (3) テキスト自動要約タスク：日本語の要約作成、情報検索タスクによる評価

(1) はニュース記事、(2) は学会発表論文や研究報告書などの要旨約70万件（訓練用の半数以上、評価用の約4割は日英対訳）、(3) は多様な種別の新聞記事を使用する。

参加申込は9月中旬の検索結果提出の前であれば、受付可能である。成果報告会は、2001年2月に東京で開催する。詳細はNTCIRホームページ²⁾をご参照いただきたい。

中国語語検索は、言語横断検索のための国際協力に向けたステップとして国立台湾大学の協力により実現した。その他、アジア諸国的情報検索グループ、ヨーロッパのCLEF (Cross-Language Evaluation Forum)³⁾、TREC (Text REtrieval Conference)⁴⁾とも密接な連携を保っている。

テキスト自動要約は、1950年代から研究が開始されたが、その評価は難しいとされてきた。この数年、研究が活発化している。このタスクは、Text Summarization Challenge (TSC) と称し、昨年8月から、要約の種類、定義、評価手法について活発な議論を重ねている。

日本語・英語検索では、日本語特有の問題を検討するため、図-4のような日本語の語分割データを提供する。この目的は、(1) 海外からの参加促進、(2) 語分割法の影響を最小限にした検索アルゴリズムの比較、語分割法の検索有効性への効果の比較など、システム間の相互比較を深めることである。語分割データの利用は必須ではないが、各自の問題設定に応じた積極的活用を期待している。昨年のNTCIRに統一して学術文書を使用し、昨年使用した対訳文書から自動生成した対訳辞書の新文書集合での評価、専門用語の処理などもポイントとなる。

■国際的動向■

主な情報検索の評価会議には、TREC⁴⁾、CLEF³⁾、TDT (Topic Detection and Tracking)⁶⁾などがある。TRECは、2000年からは、米国DARPAのTIDES (Translingual Information Detection, Extraction and Summarization Program)⁷⁾という研究推進プログラムの下で行われる。

TIDESは、名前から明らかのように、言語横断とトピック同定、抽出、要約などのテキストから情報を取り出し、利用を支援する技術に関心がある。言語横断検索は、TRECでは欧米諸言語間については、英語単言語の90%以上に達し得ることが明らかになり、英語とアジア諸言語など、言語の構造や起源が異なる言語間へと関心が発展している。2000年のTREC-9でも英語・中国語を取り上げる。2000年から開始されたCLEFは、欧州諸言語の検索のより深い議論を目的とし、国際協力体制で検索課題作成と正解判定を行う。

TRECでは、新聞記事などを用いた通常の検索は、TREC-8まで終了し、TREC-9からはWeb文書を用いたWebトラックが中心となる。Q&Aは、昨年のTREC-8では最も議論が盛り上がったトラックであった。自然言語処理と情報検索の融合したタスクとして興味深い。

情報検索システムは、現在、図-2に示したように伝統的な文書の検索から、情報や回答の提供へ、そして、評価基準もより現実に即したもののが求められている。

新しい技術には新しい評価の枠組みが必要である。評価会議では、結果の相互比較だけでなく、評価手法や今後の方針について、議論の場を提供し、コンセンサスを築くことが重要である。多くの方々からのご示唆をお待ちしています。

参考文献

- 1) Voorhees, E. M.: Construction of Q&A Test Collection, ACM-SIGIR 2000 (to appear).
- 2) NTCIR: <http://www.rd.nacsis.ac.jp/~ntcadm/>
- 3) CLEF: <http://www.iei.pi.cnr.it/DELOS/CLEF/>
- 4) TREC: <http://trec.nist.gov/>
- 5) Leong, M.-K.: Trends and Their Ramifications for the Evaluation of IR, Presented at the NTCIR/IREX Joint Workshop (Sep. 1, 1999).
- 6) TDT: [http://www.nist.gov/speech_and_selecting/benchmark_tests_and_then_TDT-3"](http://www.nist.gov/speech_and_selecting/benchmark_tests_and_then_TDT-3)
- 7) TIDES: <http://www.darpa.mil/ito/research/tides/index.html>
(平成12年7月6日受付)

