情報検索のエラー分析

難波英嗣
広島市立大学大学院
nanba@hiroshima-cu.ac.jp

酒井哲也
早稲田大学
tetsuya@waseda.jp

1. はじめに

「より良い情報検索システムを構築するために今後何が必要か」を、出力結果のエラー分析を通じて議論し、明確にすることが本タスクの目的である。一般的にこれまでの情報検索研究では、「提案手法の検索精度が、従来手法と比べてどの程度向上するのか」という点が議論されてきた。これに対し、本タスクでは、従来手法と比べてではなく、「現在の検索精度を 100%に近づけるにはどんな問題を解決しなければならないか」を明らかにする。これには、例えば、述語構造解析、含意認識、意味解析などの自然言語処理(NLP)関連の諸技術が関連してこよう。あるいは、外部の知識(例えば、各種 Open Data、オントロジーなど)が必要にもなろう。本タスクは、エラー分析を通じて「現在の自然言語処理技術に足りない技術や知識は何なのかを確認すること」を目的とした Project Next で実施される数多くのタスクのひとつであり、他のタスクを視野に入れた分析もまた期待されている。

本報告書の構成は以下のとおりである。次節では関連研究について、3 節では分析の方針について、4 節では分析結果を報告する。5 節で本報告書をまとめ、6 節で今後の課題について述べる。

2. 関連研究

本節では関連研究として、2003 年に開催された情報検索のエラー分析に関するワークショップ Reliable Information Access (以下、RIA ワークショップ)[Buckley 2003]を挙げる。情報検索システムの有効性は、検索課題によって大きく異なる。このため、個々のシステムが個別に失敗分析を行っても、システム依存の結果しか得られないという問題がある。そこで、情報検索に関わる研究者がシステムを持ち寄り、共通のデータセットを用いてエラー分析をすることにより、情報検索の（システムに依存しない）技術課題を明らかにすることが RIA ワークショップの目的である。RIA ワークショップは、情報検索研究を行っている 12 の研究機関から 28 名が参加し、7 システムを用いて、過去の TREC のトピック 45 件について検索し、TREC の適合性判定結果を用いて評価した結果を分析している。分析の結果を以下のようにまとめている。

- 現在の情報検索システムは個々のトピックに関して、同じような誤り方をしている。
- 現在のシステムの検索誤りは、検索された文書のトピックをうまく捉えられているかどうかにかかっている。
- 情報検索に関する技術を新規に考案するよりも、既存のどの技術をどの検索課題に適用するかを見つける方が重要である。

本研究では以上の 3 点をふまえ、次節で述べる NTCIR のデータセットを用い、日本語文書検索における情報検索の課題を明らかにする。

3. 情報検索タスクエラー分析の方針

3.1 エラー分析の観点

2 節で述べた RIA ワークショップでは、エラー分析を行なう際、以下の 10 種類のカテゴリを定めている。

1. 全体的に成功: 検索システムはうまくいっている

2. 一般的な技術誤り: ステミング、トークン化
   Sample Topic *353 Identify systematic explorations and scientific investigations of Antarctica, current or planned.*
   ほとんどの検索システムは、「Antarctica(南極大陸)」の語幹を「Antarctic(南極の)」としていない。

3. システムはひとつの観点だけを強調し、もうひとつの必須語を外している
   Sample Topic *422: What incidents have there been of stolen or forged art?*
   「stolen or forged art(盗難または偽造された美術品)」について、「art(美術品)」という言葉が強調されていれば検索精度が向上した。

4. システムはひとつの観点だけを強調し、もうひとつの観点を外している
   Sample Topic *355 Identify documents discussing the development and application of*

spaceborne ocean remote sensing[1].
検索システムは「ocean (海洋)」という語を強調すべき。カテゴリ 3 と似ているが、「海洋」の拡張語が検索性能を上げるのに重要。

5. システムはどちらか一方の観点だけを強調しているが、両方必要
Sample Topic *363 What disasters have occurred in tunnels used for transportation?*
「disasters(災害)」または「tunnel(トンネル)」のどちらかではなく、両方が検索時に必要。

6. システムは不適切な観点を強調し、トピックの核心は外している
Sample Topic *347 The spotted owl episode in America highlighted U.S. efforts to prevent the extinction of wildlife species. What is not well known is the effort of other countries to prevent the demise of species native to their countries. What other countries have begun efforts to prevent such declines?*
検索システムが「spotted owl(ニシアメリカフクロウ)」と「U.S. efforts(米国の試み)」という重要でない語を強調してしまっている。

7. 一般語に対する外部拡張が必要
Sample Topic *448 Identify instances in which weather was a main or contributing factor in the loss of a ship at sea.*
「weather(天気)」という一般名詞を拡張する必要がある。

8. QA クエリ解析と関係性が必要
Sample Topic *414 How much sugar does Cuba export and which countries import it?*
質問語と数量表現の関係に関する概念が必要。

9. システムは人手による支援が必要な難しい観点を外している
Sample Topic *413 What are new methods of producing steel?*
「new methods(新しい手法)」という表現(の処理)が難しい。

10. 2 つの観点間の近接関係が必要
Sample Topic *438 What countries are experiencing an increase in tourism?*
第一の焦点として「increase(増加)」、第二の焦点として「increase(増加)」と「tourism(観光事業)」という 2 つの観点が近くに現れること。

RIA ワークショップで情報検索システムの分析が行われて 10 年になるが、現在、これらのカテゴリがどの程度当てはまるのか、また、日本語を対象とした場合に問題がないかを明らかにする。

3.2 エラー分析に用いるデータ

NTCIR 事務局から入手可能なテストコレクションのうち、以下のものを本研究の分析対象に用いる。

- Web 文書
  - NTCIR-9, 10 INTENT タスク
  - NTCIR-3, 4, 5 Web 検索タスク
- 新聞記事
  - NTCIR-3, 4, 5, 6 言語横断検索(CLIR)タスク(うち日本語検索課題に対し日本語新聞記事を検索するもの)

これらを選択した理由のひとつには、比較的長期間にわたって実施されたタスクであることが挙げられる。NTCIR 事務局からは、検索課題だけでなく、タスクに参加した各システムの出力結果も入手することができる。それらの結果を分析することで、時間とともに情報検索システムエラーの種類が変わる場合、その現象を観測できる可能性があるからである。

もうひとつの理由は、検索対象文書の性質に依存する問題に対処するためである。Web 文書を対象にした検索システムを構築する場合、検索対象の文書本文だけでなく、アンカー文字列、文書間のリンクなども利用するのが一般的であり、分析の際､Web 文書固有の性質を考慮する必要がある。他方、新聞記事の場合は、Web 文書のような固有の構造は存在しないが、文体が統一されており、表記ゆれがほとんど存在しないなど、別の性質がある。そこで、Web と新聞の両者を対象とすることで、文書依存の問題に対処する。

これらの理由に加え、さらに、Web 検索では再現率よりも精度が重視されるのに対し、新聞記事の検索では再現率と精度の両方が重視されるため、エラー分析でもその点を考慮する必要がある。

4. 分析

本節では、エラー分析の現状について報告する。

4.1 データ

3.2 節で述べたデータのうち、今回は NTCIR-5 Web 検索タスクと NTCIR-6 CLIR タスクを取り上げ、そのトップシステムの結果を分析した。分析に用いたデータの詳細を表 1 に示す。

[1] 人工衛星による海洋リモートセンシング

表 1　分析に用いたテストコレクションの概要

	NTCIR-5 Web 検索タスク	NTCIR-6 CLIR タスク
トピック(クエリ)数	269	50
検索対象文書数	約 1 億ページ (1.36TB)	858,400 件(毎日/読売 2000-2001)
評価尺度	DCG & WRR	MAP 他

また、各タスクのトップシステムの概要は以下のとおりである。

- NTCIR-5 Web 検索タスク
  トップシステム(TNT-3) [Fujii 2005]
  検索モデルとして BM25 を採用。本文の他にアンカー文字列も利用[2]。さらに、「Excite 翻訳」と「エキサイト翻訳」のような日本語文書中の英単語に関する表記ゆれに対応するため、翻字技術も利用。
- NTCIR-6 CLIR タスク
  トップシステム (TSB-J-J-D-02) [Sakai 2007]
  検索モデルとして BM25 を採用。擬似適合性フィードバックも利用。

各システムが出力した結果のうち Web、CLIR から各 10 トピック、上位 10 件[3]の計 200 事例をエラー分析の対象とした。

4.2 分析結果

3.1 節で紹介した RIA のカテゴリのうち、RIA-1 以外のカテゴリで、4.1 節のデータを用いた場合でも同様の誤りが確認できたものの例を以下、RIA-2～RIA-10 に示す。さらに、今回の分析で新たに判明したエラーを NEW-1～NEW-2 に示す。また、各事例のトピック番号は、NTCIR-5 Web 検索タスクの課題 1003 の場合「Web-1003」NTCIR-6 CLIR タスクの課題 16 の場合「CLIR-16」と表記する。

(RIA-2) 一般的な技術誤り: ステミング、トークン化[4]

このカテゴリに含まれる誤りとして、形態素解析の誤りよりもむしろ索引語の言語単位をどうするか(単語 or 複合語)に関するものがあった。「2 ちゃんねるのサイトを探したい」(Web-1004)というトピックにおいて、「電波 2 ちゃんねる」という 2 ちゃんねる検索サイト(文書番号:0156422_0000001)が誤って検索されている。これは、「電波 2 ちゃんねる」という固有名詞の中にトピックの重要単語である「2 ちゃんねる」という語が含まれているためである。同様の事例として、「鳥取県の 21 世紀梨[5]を知りたい。鳥取県農協の公式サイトを適合とする。」(Web-1003)というトピックにおいて、「鳥取県が『二十世紀梨記念館』開設へ」というニュース記事(文書番号:0238963_0002136)が検索されているケースが挙げられる。これは、固有名詞「二十世紀梨記念館」の中に「二十世紀梨」という語が含まれていることが原因である。

(RIA-3) システムはひとつの観点だけを強調し、もうひとつの必須語を外している

ANA（全日空）のオンラインチケットサービスサイトを探すトピック(Web-1006)で、オンラインチケット予約のサイトだが ANA のオンラインチケット予約ではないものが誤って検索されている。同様の事例として、「環境ホルモンによって引き起こされる病気や脅威に関する文書」(CLIR-14)というタスクで環境ホルモンの水質調査に関する記事が誤って検索される事例があった。

(RIA-4) システムはひとつの観点だけを強調し、もうひとつの観点を外している

「EAGLES というロックバンドの公式サイト」を検索するトピック(Web-1012)で、上智大学アメリカン・フットボール部 EAGLES が誤って検索された。正解文書には「ロックバンド」や「ロック」という語は含まれていないが、「ミュージック」「視聴」「邦楽」「アーティスト」「曲名」などの音楽関

[2] このチームは、PageRank を使った場合についても結果を提出しているが、本タスクでは、PageRank を使うとかえって検索精度が低下していた。

[3] ただし、上位 10 件の中に A 判定の正解文書が含まれていないトピックに限定して分析している。

[4] 分析をはじめるまでは、新聞記事と比べ、Web 検索では表記ゆれに関する問題が多いのではないかと思っていたが、Web 検索ではアンカー文字列を使うと、表記ゆれに対する問題が案外目立たない。例えば、「2 ちゃんねるのサイトを探したい」(Web-1004)というトピックの場合、正解ページのアンカー文字列に「２ちゃんねる」「2ch」「２チャンネル」「http://www.2ch.net」「某掲示板」など、「2 ちゃんねる」のあらゆる異表記が含まれている。

[5] 正しくは「20 世紀梨」であることは、トピックにも記述されている。

連の用語が拡張語として得られれば、検索できる[6]。

(RIA-5) システムはどちらか一方の観点だけを強調しているが、両方必要
「ExCite の英和辞典を使いたい。」という課題(Web-1013)で、Excite の他のサービス(Woman excite や Excite blog)が検索されたり、Excite 以外の英和辞典サイトが検索されたりする事例があった。「Excite」と「英和辞典」の両方が必要。

(RIA-6) システムは不適切な観点を強調し、トピックの核心は外している
「ロックバンド EAGLES の公式(official)サイト」を探す課題(Web-1012)で、「公式」や「official」を含んでいるが EAGLES を含んでいない文書が検索されている事例があった。

(RIA-7) 一般語に対する外部拡張が必要
「ティーンエージャーの社会問題を扱った記事」を探すトピック(CLIR-18)で、「ティーンエージャー」が 13 歳から 19 歳の若者を指すという知識が必要である[7]。システムが検索した文書 JY-20010602J1TYEUH0400020 は、ある人物の引きこもりに関するものであるが、その人物の年齢が 21 歳であるため、非適合文書となっている。

(RIA-8) QA クエリ解析と関係性が必要
いわゆるファクトイド型質問応答のように、「いくら」や「どのくらい」などを問うトピックであるが、今回の分析トピックの中に該当するものはなかった。

(RIA-9) システムは人手による支援が必要な難しい観点を外している
「違法盗聴によるプライバシーの侵害、特に国家レベルではなく個人レベルのプライバシー」を扱っている文書を探す課題(CLIR-24)で、エシュロンに関する記事(記事番号:JA-010528214)を検索していた。エシュロンが国家レベルのプライバシーを扱ったものであるという判断は容易ではないと思われる。ただ、本来ならば、多くのシステムの検索結果を分析する必要があるが、今回は 1 システムの結果しか分析していないので、このカテゴリの該当事例が見つかっていない。

(RIA-10) 2 つの観点間の近接関係が必要
「ヒト ES 細胞の紹介記事」(CLIR-3)を探すタスクでサルの ES 細胞に関する記事が誤って検索された事例(文書 JY-20001002J1OYEFF0300010)があった。ただし、この事例中にはヒト ES 細胞に関する記述も一部あるため、非適合文書であるとシステムが判断するのはかなり難しいと思われる。

(NextNLP-1) 不適切なクエリ
ユーザが入力するクエリが不正確であったり、一般的な表現を用いなかったりした場合に、正解文書ができない場合がある。例えば、RIA-2 の例で挙げた例「鳥取県の 21 世紀梨を知りたい。」(Web-1003)や、「FP(ファイナンシャル・プランニング)資格試験の情報のページ」(Web-1014)などが該当する。「FP 資格」は、一般的には「CFP 資格」と表現されることが多い。

(NextNLP-2) 否定表現
「西暦 2000 年問題による故障(バグ)に対処する方法や管理に関連する諸産業における問題」(CLIR-20)について述べた記事を探すタスクで、特定の地域で問題が発生しなかったという記事が誤って数多く検索された。これは「異常なし」の「なし」といった否定表現を考慮していないためである。

4.3 考察

誤りカテゴリについて

分析した誤り事例数は十分ではないものの、その多くは RIA で提案されたカテゴリに分類できることが分かった。RIA 以外のカテゴリとして、今回新たに NextNLP-1 と NextNLP-2 を提案した。分析の順序から考えれば、まず、正確な語がクエリとして使われる場合について十分議論した後に、次に、NextNLP-1 のような事例を検討すべきである。しかし、そもそもユーザは、分からないこと、知らないことを調べたいから検索するのであり、クエリに不正確な語が含まれるのは、むしろ自然な現象であると言っても過言ではない。

一方、NextNLP-2 については、2000 年問題以外のトピックで当てはまる事例を見つけていないため、このカテゴリがどの程度一般的なものであるかを見定めるには、さらに多くの検索誤りを分析する必要がある。

検索誤りの改善方法について

各カテゴリに分類した事例に対処するにはどのような技術を用いれば良いか？については、今後検討しなければならない点が多い。例えば、RIA-4 の例にある EAGLES の公式サイトを検索する場合、ロックバンドという語の拡張語を用いれば、

[6] この例は RIA-7 にも分類できるかもしれない。
[7] トピック中にもティーンエージャーの説明はあるが、11 歳から 19 歳と説明自体が誤っている。

正解文書が検索できる可能性がある。しかし、どのような条件で、どのように、またどの程度検索語を拡張するかは明らかではない。

「LaTeX の奥村先生について知りたい」(Web-1025)というトピックの場合、LaTeX の解説文書が検索されたり、LaTeX とは関係のない同姓の人物について言及した文書が検索されたりした。奥村先生という人名は多義語であり、LaTeX という語が、どの奥村先生であるのかを限定している。これは、上述の EAGLES とロックバンドの関係と同じである。しかし、EAGLES の例と同様に、LaTeX の拡張語を用いても、さらに TeX 関連の解説文書が検索されてしまうのは容易に推測できる。

多義語ではないが、LaTeX−奥村先生と類似した関係を持つトピックとして、「J リーガーの中澤佑二について知りたい」(Web-1022)というものがある。このトピックの場合も、LaTeX の奥村先生の事例と同様、J リーガーのみを含んだ文書を検索しており、それが検索精度を低下させる原因のひとつになっていた。この場合は J リーガーと中澤佑二という 2 つの検索語で検索するよりも、中澤佑二という 1 語で検索した場合の方が、おそらく検索精度は向上するであろう。しかし、将来的に同姓同名の（J リーガーでない）中澤佑二が現れる可能性はあり、この場合は、J リーガーという語が検索語として必須になる。人名等の固有名が多義語であるかどうかは、実際に検索して結果を見るまでは分からないこともある。従って、まずは中澤佑二だけで検索し、検索結果から多義語である可能性が高ければ、自動的に J リーガーを追加して再検索をする、などの改善方法が考えられる。

5. おわりに

本報告書では、NTCIR-5 Web 検索タスクおよび NTCIR-6 CLIR タスクのトップシステムの出力結果を用い、エラー分析を行った。分析の結果、10 年前に実施された RIA ワークショップで提案されたカテゴリの多くが確認されたが、新たな現象として多義語の問題と否定表現の問題も見つかった。

6. 今後の課題

今回は、分析に用いるデータを入手し、分析環境を構築するまでにかなりの時間を要したため、分析そのものに十分な時間を割り当てることができなかった。分析結果の一般性を高めるためにも、より多くのシステムの出力結果を対象にする予定である。また、Web 検索に関しては最近行われた INTENT タスクと NTCIR-3〜5 の Web 検索タスクとの違い、新聞記事検索に関しては 4 回にわたって行われた CLIR タスクの結果の比較を予定している。

参考文献

[Buckley 2003] Buckley, C. and Harman, D. “Reliable Information Access Final Workshop Report.” Proceedings of the Reliable Information Access Workshop (RIA). NRRC, pp.1-30 (2003)

[Fujii 2005] Fujii, A., Itou, K., Akiba, T., and Ishikawa, T. “Exploiting Anchor Text for the Navigational Web Retrieval at NTCIR-5.” Proceedings of NTCIR-5 (2005)

[Kishida 2005] Kishida, K., Chen, K.-H., Lee, S., Kuriyama, K., Kando, N., Chen, H.-H., and Myaeng, S.H. “Overview of CLIR Task at the Fifth NTCIR Workshop.” Proceedings of NTCIR-5 (2005)

[Oyama 2005] Oyama, K., Takaku, M., Ishikawa, H., Aizawa, A., and Yamana, H. “Overview of the NTCIR-5 WEB Navigational Retrieval Subtask 2 (Navi-2).” Proceedings of NTCIR-5 (2005)

[Sakai 2007] Sakai, T., Koyama, M., and Izuha, T. “Toshiba BRIDJE at NTCIR-6 CLIR: The Head/Lead Method and Graded Relevance Feedback.” Proceedings of NTCIR-6 (2007)