応般

3.8 Project Next IR
―情報検索の失敗分析―

難波 英嗣（広島市立大学）　酒井 哲也（早稲田大学）　神門 典子（国立情報学研究所）

Project Next IR の目的

「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，失敗分析を通じて議論し，明確にすることが本タスク（以下，Project Next IR）の目的である．一般的にこれまでの情報検索研究では，「提案手法の検索有効性が，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，Project Next IR では，従来手法と比べてではなく，「現在の検索システムをより良くす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．本タスクは，失敗分析を通じて「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し，次の研究課題を明らかにすること」を目的とした Project Next NLP で実施される数多くのタスクの 1 つである．このため，ほかのタスクを視野に入れた分析が期待されているが，情報検索には自然言語処理とは異なる技術的課題も多数あるため，それらも視野に入れた分析を目指す．

情報検索における失敗の原因

情報検索における失敗とは何であろうか？　この質問に答える前に，まず情報検索のタスク設定について説明しておく．たとえば，Google などの Web 検索システムを使って，「新宿にある美味しいフレンチレストラン」を探す場合を考えてみよう．検索者は，「新宿」や「フレンチレストラン」，また人によっては「美味しい」などの語を検索フォームに入力して検索する．得られた検索結果を順に見ていき，目的のレストランを見つけることになる．情報検索研究のタスク設定では，「新宿にある美味しいフレンチレストラン」などの検索質問（検索者が何を探そうとしているのかを明確に記述したもの）や必要に応じてその背後の検索意図や正解判定基準などを記述した「検索課題」，「検索対象の文書」，さらに，検索対象文書の中でどの文書が正解かを人手で判定したもの（正解文書リスト）を用意しておく．今，ある検索システムの良し悪しを測るために，実際に検索質問をそのシステムに入力して検索を実行する．その実行結果を正解文書リストと比較し，再現率や精度などの評価尺度を用いて数値化したものが，その検索システムの検索有効性ということになる．

情報検索の失敗には 2 種類ある．1 つは，人手で正解と判定された文書が検索システムで検索できなかった場合であり，もう 1 つは，検索システムが検索課題と関係のない文書を返した場合である．前者はたとえば，正解文書中に「フレンチレストラン」ではなく「フランス料理店」と書かれている場合である．もし検索システムが「フレンチレストランはフランス料理店と同じ意味である」ことを知らなければ，この正解文書は検索できない．後者は，「フレンチレストラン」と「新宿」の両方が出現するものの，新宿のフレンチレストランとはまったく関係のない文書を返す場合がその例として挙げられる．

Project Next IR における失敗分析

情報検索の失敗分析を多数のシステムを横断して系統的に行い，現在の技術的課題を明らかにするという試みは，実は Reliable Information Access[1] と呼ばれるワークショップ（以下，RIA）で，2003 年に行われている．この RIA の取り組みからすでに 10 年経過しているが，この間に，技術面でも，タスクの多様性の面でも広がりを見せてきているため，RIA の成果も踏まえた上で，改めて情報検索の失敗分析を行う Project Next IR を 2014 年 7 月

より開始した．このプロジェクトでは，評価型会議NTCIR☆1のデータのうち，Web文書を対象とした検索課題と新聞記事を対象とした検索課題を分析対象として取り上げた．これらのデータを用いた分析により明らかになった主な失敗カテゴリを以下に示す．

索引語の言語単位

索引語を単語とするか，あるいは複合語とするかに関するものがある．「鳥取県の二十世紀梨を知りたい．鳥取県農協の公式サイトを適合とする」という検索課題において，「鳥取県が『二十世紀梨記念館』開設へ」という文書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．

複数の検索語

「Exciteの英和辞典を使いたい」という課題で，Exciteのほかのサービスが検索されたり，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」の両方が文書に含まれている必要がある．

検索課題と文書の主題の不一致

「ヒトES細胞の紹介記事」を探す課題でサルのES細胞に関する記事が誤って検索された事例があった．ただし，この事例はヒトES細胞について何も論じていないが「ヒトES細胞」という文字列が出現するため，不適合と検索システムが判断するのは困難である．

外部知識が必要な検索語

「ティーンエイジャーの社会問題を扱った記事」を探す課題では，「ティーンエイジャー」が13歳から19歳の若者を指すという知識が必要である．

語の多義性

「EAGLESというロックバンドの公式サイト」を検索するという課題で，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．

不適切な検索語

ユーザが入力する検索語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，正解文書を返せない．たとえば，「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」を探す課題が該当する．FP資格は，一般的にはCFP資格と表現される．

なお，これらの分析結果に関する詳細は参考文献2)を参照されたい．

☆1 http://ntcir.nii.ac.jp

Project Next IRの今後

本稿では，2014年より開始した情報検索の失敗分析タスクProject Next IRを紹介した．2015年4月からは，より大規模な分析を実施するため，以下のメンバで活動を開始している．

- 江口 浩二（神戸大学）
- 神門 典子（国立情報学研究所）
- 欅 惇志（東京工業大学）
- 酒井 哲也（早稲田大学）
- 清水 敏之（京都大学）
- 難波 英嗣（広島市立大学）
- 波多野 賢治（同志社大学）
- 平手 勇宇（楽天株式会社）
- 藤井 敦（東京工業大学）

2014年度の分析経験と手法を踏まえ，2015年度は，state-of-the-artな複数のシステムの失敗分析を系統的に行い，現在の情報検索における技術的課題と制約を明らかにし，今後，検索技術をより良くするための具体的課題を明らかにすることを目標としている．

参考文献

1) Buckley, C. and Harman, D. : Reliable Information Access Final Workshop Report, Proceedings of the Reliable Information Access Workshop (RIA). NRRC, pp.1-30 (2003).
2) 難波英嗣，酒井哲也：情報検索のエラー分析，言語処理学会第21回年次大会 自然言語処理におけるエラー分析ワークショップ（2015）．

（2015年9月30日受付）

難波 英嗣（正会員）nanba@hiroshima-cu.ac.jp
博士（情報科学）．北陸先端科学技術大学院大学．東京工業大学助手などを経て，現在広島市立大学大学院情報科学研究科准教授．2011年本会論文誌データベース（TOD）優秀論文賞．

酒井 哲也（正会員）tetsuya@waseda.jp
博士（工学）．早稲田大学．早稲田大学情報企画部副部長・基幹理工学部情報理工学科教授．国立情報学研究所客員教授．本会IFAT研主査・論文誌TOD共同編集長など歴任，論文賞（2回）・山下記念研究賞など受賞．Information Retrieval Journal（Springer）共同編集長．

神門 典子（正会員）kando@nii.ac.jp
博士（図書館・情報学）．慶應義塾大学．国立情報学研究所教授．ACM TALIP，IP&M（Elsevier）Associate Editorを歴任．NTCIR共同ジェネラルチェア，SIGIR 2017共同ジェネラルチェア．