3 情報検索研究者にとってのテストコレクションと評価ワークショップ

3.1 大学研究者から見たテストコレクションと評価ワークショップ

山本 幹雄　筑波大学　myama@is.tsukuba.ac.jp
梅村 恭司　豊橋技術科学大学　umemura@tutics.tut.ac.jp

■チーム紹介■

我々のチームは，筑波大学・豊橋技術科学大学・AT&T Labs-Researchの混成チームであった．メンバは計算機（人工）言語と音声・自然言語の研究者であり，Suffix Arrayと呼ばれるデータ構造を使って任意の文字列の各種統計量を大規模コーパスから得ることを共通の興味としていた．Suffix Arrayが任意の長さの文字列を扱えるということは，単語という概念が曖昧な日本語の処理に向いており，特に最適な処理単位がはっきりしない日本語情報検索に最適ではなかろうかという議論を前々からしていた．このようなわけで，我々のチームは情報検索に興味があることには間違いはないのだが，情報検索の専門家の集団とはいえなかった．

■参加の経緯■

非専門家のグループであるため，実際に情報検索システムを作成してアイデアを実現・評価するには躊躇があった．特に，システムの性能評価にはかなりの規模の評価データを利用する必要があると聞かされ，また現在の情報検索における標準的な評価手法もフォローしていない点がネックとなっていた．特に大学を中心とした非専門家のグループに，大規模な評価データを作成する余裕はない．そんな折，ちょうどNTCIRとIREXのコンテストの計画が発表された．データも提供され，しかも評価までしてくれるという．我々にとってまさに願ってもないものであった．このような経緯で，計画が発表されると，すぐさま両方のコンテストへの参加を決定し，それから情報検索の勉強を始めた．お恥ずかしい話であるが，11pt平均精度（再現率0～100%を10等分した11点について，それぞれの精度を平均した評価指標）の計算で，再現率が0%の点がなぜ含まれるのかも分からず，調べるのにかなり苦労するレベルであった（回答は文献1）で見つけた）．

■参加システムと結果[2),3)]■

ここでは，自然言語文で書かれた検索質問文を入力とした情報検索タスクにエントリーしたシステムについて簡単に紹介する．情報検索システムでは，検索質問と文書間の類似度を定義し，検索質問と類似度の高い文書から順に，検索結果として返すことになる．この過程は，一般のパターン認識と似ている．たとえば音声認識では，入力音声をシステムが貯えている音韻や単語のモデルと比較し，類似度の最も高い音韻や単語をその出力とする．ただし，音声は時間軸方向への変化が個々の発声ごとに大きいため，入力音声とあるモデルの類似度は，時間軸をさまざまに伸縮させてマッチさせた中で最大の類似度と定義される．

この「入力をモデル（文書）ごとに変形する」というアイデアを利用したのが，我々のシステムの特徴である．

音声における時間軸は，テキストにおいては文字列の並びの方向となる．筑波大学のシステムでは，検索質問文とある文書間の類似度は，検索質問文のあらゆる（文字列）分割を考慮し，その中で最大となる類似度と定義した[2)]．

また，豊橋技術科学大学のシステムはさらに音声認識に近く，検索質問文と文書間をDPマッチングすることによって，単語の順序まで考慮した類似度となっている[3]．

筑波大学のシステムは，副作用として，システムがよしとした検索質問の単語分割結果を得ることができる．この分割結果は，分割の基準として用いる類似度の違いによって面白い特徴があることが分かった．情報検索の標準的な類似度を使うと，性能は上がるが，分割結果は人間が見て単語らしくない単語境界上の文字列となる傾向がある[2]．類似度を若干変えると，分割結果は人間が見ても比較的まともな単語・句となるが，検索性能があまり上がらない[4]．これは，現在の単語中心の情報検索単位に疑問を投げかける結果かもしれない．

この結果は，我々のシステムが両コンテストで上位であれば，興味深かったのであるが，残念ながら，IREXで中位，NTCIRでは低位の成績であるため，あまり説得性がない．しかし，言い訳を1つ．我々の方式は，構造化や付加情報のない単なるテキストデータを対象としたときに特徴が発揮されると考え，利用できるキーワード情報を扱わないシステムであった．方式の特徴からキーワードを等価的に自動抽出しているいるので，キーワードを用いないでも大丈夫と考えたところが間違いだった．コンテスト後にすべての与えられている情報を使用して検索実験をやり直すと，11pt平均精度で7～8%程度向上する（あくまでも後智恵ではあるが）．

■今後に期待すること■

ある研究グループが，自らの分野の手法を用いて，異分野へ進出する際に苦労するのは評価方法の違いであろう．アウトサイダーがインサイダーに対して性能改善を納得させるのは一般に難しい．これはどの分野についてもいえると思うが，特に，情報検索は歴史が長い割に問題設定がシンプルであるため，性能を非常に重視し，その評価手法に落ち度があってはいけないようである．情報検索の国際会議でrejectされた論文の批判を見ると明らかである．曰く，ベースライン・システム（比較のための基準システム）の性能が低すぎる．曰く，十分に一般的な結論であるかどうかがはっきりしない．ベースラインも精一杯作り込んだつもりではあるが，こう言われると精進するしかない．また，「この改善は一般的か」と言われてしまうと，これは使っている評価データそのものの信頼性を問われているのに近く，アウトサイダーにはお手上げである．

コンテストに最も期待するのはこの点の改善である．使われたテストコレクションが，情報検索システム評価のための1つの標準テストコレクションとして国内外に認知されることを期待する．これを使った実験結果に対しては少なくとも「評価データが信頼できない」などとは言われたくないものである．また，同じテストコレクションを用いたコンテスト結果をreferenceとして示せば，ベースラインとして誰もが納得してくれる状況ができるとありがたい．今回の2つのコンテストは，英文での参加募集もあり（実は我々のチームで最初に参加しようと言い出したのは，米国の研究者であった），特にNTCIRは国際会議の形態をとったため，NTCIRのテストコレクションは国際的な認知という意味ですでにかなり改善されたことと思う．今後もさらに，認知度をあげてほしい．

また，今回初めて情報検索のシステムを構築してみて閉口したのは，情報検索の類似尺度のバリエーションの多さであった．類似尺度は重要な検討課題の1つであろうが，類似尺度を主要なテーマとしない研究まで，類似尺度の違いによって比較が困難になっているのは残念である．できれば，コンテストを通じて日本語情報検索における比較的よい類似尺度の候補を明らかにしていただいて，コンテストの標準類似尺度を提示していただけるよいのではないか．

いずれにせよ，素人は素人なりに，今回の2つの情報検索コンテストを大いに楽しませていただいた．今回は，IREXが先行する形で，自然言語処理の分野の研究者が気楽に参加できる雰囲気があったのはたいへんよかった．情報検索はすでに長い歴史があり，基本技術はかなり確立しているが，おそらく，現在のインターネットの検索エンジンで満足している人は少ないであろう．多くの研究者がさらなる発展に興味を持っている分野である．今後の発展のためにも，特に多様性という観点から，さらに敷居を下げて，さまざまな分野の多くのチームが気楽に参加できる雰囲気ができるとよいと思う．

最後になったが，関係各位のご努力に感謝するとともに，この種のコンテストの今後の順調な発展をお祈りしたい．