情報検索のための自然言語処理ツール群の開発

関口 宏司(1210905)

北陸先端科学技術大学院大学 情報科学研究科

2014 年 2 月

キーワード：情報検索, 形態素解析, 固有表現抽出

今日、情報検索はますます不可欠なツールとしてコンピュータユーザの日常に溶け込んでいる。しかしながら、現在の情報検索が抱える問題も大きい。情報検索システムがユーザの期待に応えられていない点は、(1)ユーザが求める情報を返せない、(2)ユーザが求めていない情報を返してしまう、という 2 点に集約される。前者の問題は「検索漏れ」、後者の問題は「検索誤り」と呼ばれる。検索漏れがあると、ユーザは必要な情報が得られるまで検索質問を少しずつ変えながら繰り返し検索を実行しなければならない。また検索誤りがあると、情報検索システムから返ってきた膨大な検索結果の中からユーザが求める文書を探す作業が困難になる。一般に検索漏れと検索誤りの問題はトレードオフの関係にある。検索漏れに対処するために検索システムの出力を大きくすると検索誤りが増えてしまい、検索誤りに対処するために検索システムの出力を小さくすると検索漏れが増えてしまうためである。そこで本課題研究では、両問題を同時に解決するのではなく、段階的に対処する方法を提案する。提案方法では、まず検索漏れを小さくし、次いで絞り込み検索によって漸次的に検索誤りを小さくする。

検索漏れを小さくするために、自然言語処理の既存研究を応用した「原型語とその省略語の自動抽出プログラム」「漢字送りがな表記揺れ知識の自動抽出プ

ログラム」「N-best パス探索形態素解析プログラム」という 3 つのツールを開発した。

「原型語とその省略語の自動抽出プログラム」は、原型語と省略語候補の類似度計算に基づく酒井らの研究を辞書に応用したものである。開発したプログラムを Wikipedia に適用したところ、再現率は 90%を超えたが精度は 70%を下回った。しかし、省略語だけでなく関連語も正解とするように条件を緩めると、再現率はあまり変わらず、精度は 80%を超えた。「漢字送りがな表記揺れ知識の自動抽出プログラム」は、単語辞書から表記揺れがある漢字送りがなの組を自動抽出するものである。形態素数 392,126 件の単語辞書 IPAdic に本プログラムを適用したところ、9,449 件の漢字送りがな表記揺れ知識が抽出できた。「N-best パス探索形態素解析プログラム」は、日本語における単語分割の曖昧性の問題に対処した形態素解析プログラムである。本プログラムは日本語テキストを形態素解析し、コスト最小のパス上の全トークンを出力した後、2 位以下 N 位までのパス上の名詞のみを出力する。2 位以下の重複するトークンや検索される可能性の低い名詞以外の語を出力しないことで情報検索における処理負荷を軽減できる。また、2 位以下の解からも名詞を出力することで、単語区切りの不一致による検索漏れを防ぐ効果が期待できる。本形態素解析プログラムは、永田の方法に基づき入力文の N-best パス解出力を行う。永田の方法では、最初に与えられた入力文を 1 文字ずつ前向きに解析し、単語ラティスを作成する。次に作成した単語ラティスからコストが小さい上位 N 個のパスを探索してパス上のトークンを順に出力する。単語辞書にはトライ構造をコンパクトにメモリに格納できるダブル配列を採用して実装した。さらに、「原型語とその省略語の自動抽出プログラム」および「漢字送りがな表記揺れ知識の自動抽出プログラム」の出力を利用できるようにした。実験により N-best パス解が出力され、「原型語とその省略語の自動抽出プログラム」および「漢字送りがな表記揺れ知識の自動抽出プログラム」の出力が適用されたことが確かめられた。このことから、この出力を使って転置インデックスを作成すれば、情報検索における再現率の向上が期待できることがわかった。

以上の検索漏れ対策により検索の精度は低下する。そこで、絞り込み検索により漸次的に検索誤りを小さくする。情報検索における絞り込み検索とは、ユーザが最初に実行したクエリに加えて新しい検索条件を追加して再検索することを指す。「絞り込み」検索のため、追加する検索条件は既存の検索条件に AND

でつながれる。この操作により、精度を漸次的に向上させていくことが可能である。絞り込み検索を行うためには、検索対象文書レコードが、絞り込みを行うための構造を持っていなければならない。新聞記事などをはじめ、多くの文書はこのような構造を持っていないので絞り込み検索を容易に実行することができない。しかし、固有表現抽出技術を適用すると、絞り込み検索に適した構造を持たせることができる。固有表現抽出は品詞タグ付けなどと同様、系列ラベリング問題の一種である。固有表現タグ付きコーパスから、入力文における個々のトークンに対する固有表現タグを決定するモデルを機械学習した。機械学習アルゴリズムは条件付き確率場（Conditional Random Field; CRF）を採用した。訓練データとなる固有表現タグ付きコーパスとして、関根の固有表現タグ付きコーパスを用いた。情報検索をビジネスの場面で利用することを想定し、粒度の細かい関根の拡張固有表現階層のタグを 9 個の新たなタグにまとめ直した。関根の固有表現タグ付きコーパスから 12,000 文を抽出し、3 分割の交差検定によって本課題研究で実装した固有表現抽出ツールを評価したところ、固有表現タグの種類により F 値に差が出ることがわかった。そこで固有表現の品詞の出現頻度を調べたところ、成績の低い固有表現タグでは一般性の高い品詞の出現頻度が 1 位となっていた。逆に成績の高い固有表現タグでは固有表現の意味に沿った特徴的な品詞の出現頻度が高かった。また、別の有用な素性を追加することで F 値が改善されることを考察した。