

## 固有表現抽出を用いた時事問題生成システム

小泉 昌也

昨今、公務員試験や就職活動、企業における昇進のための社内試験などにおいて、近年の社会の動きなどを問う時事問題が出題される場面が多くある。この時事問題は、その性質上、時間の流れに比例して極めて流動性が高く、書籍・ウェブサイトで提供されるものは即時性の面で疑問が残る。また、専門的な業界内の時事について対策を行いたい場合は、既存の書籍・ウェブサイトでは網羅性が低く不十分だと考えられる。そこで研究は、利用者が自ら必要とする分野に関連する文章を入力し自由に問題を生成することができるシステムを構築し、即時性を持つ時事問題や時事に限らずより多くの分野に関する問題を社会に提供することを目的とする。

本研究で構築したシステムでは、まず利用者が必要とする分野の文章を新聞や教科書の文章を抜粋して入力すると、元の文章を形態素分析して固有名詞の抽出を行う。更にそれを、TF-IDF を用いて重要度順にソートし、重要度が高いと判断された特定の単語の中から1つを正答と設定し、その箇所を穴埋めする問題として表示する。穴埋め式問題とすることで、これまでの研究に比べて利用者の事前準備の負担を減らし、より効率の良いシステムとすることを目指している。

本システムの有用性を検証するため、4つの実験を行った。正確性に関する指標である適合率と、網羅性に関する指標である再現率の調和平均によって求められるF尺度を用いて、①TF-IDF計算におけるジャンル分けの有効性の確認、②抽出する固有名詞の個数とF尺度の関係性、③固有名詞の抽出実験を行った。そして「朝日キーワード就職2019 最新時事用語&一般常識」を用いて、④生成した問題の妥当性を評価した。

実験の結果、TF-IDF計算におけるジャンル分けはある程度有効であり、システムに入力したテキストにおける赤字部分の1.1倍の個数を抽出した際に最もF尺度が高くなることがわかった。またジャンルごとに抽出した固有名詞の単語欠損率と単語一致率には明確な差が生じることがわかった。その要因には数詞の多さ、地名などの出現室の高さによる重要度順のずれなどが考えられる。

本研究では、時事問題対策における既存コンテンツの即時性・網羅性の低さを向上させるため、利用者がテキストを指定してより効率的かつ有用な時事問題を生成するシステムを構築した。本システムを用いることで、テキストから固有名詞を抽出するというアプローチにおいては高い適合率を得ることができ、問題を生成することができたと言える。今後の課題は、抽出した単語の重要度の測定の精度が挙げられる。ジャンルごとに様々な性質があることがわかったため、その性質に合わせた重要度の基準設定が求められると考えられる。

(指導教員 辻 慶太)