用語抽出技術を応用した穴埋め問題自動作成システム

須藤槙

本研究では主に就職活動生が時事問題、一般常識問題対策ができるよう、任意の新聞記事や重要時事まとめサイトの文章を入力すると自動で穴埋め問題を出力することができるシステム作りを目指し、その構築を行った。

本システムは形態素解析器である MeCab を利用し、時事問題に出題される可能性が高い固有名詞と、その他一般常識問題で出題される可能性が高い名詞を抽出できるよう、ユーザ辞書を作成したものである。基本的にユーザ辞書は Wikipedia のタイトルリストから、不要だと思われる語をプログラムと目視によって削除する形で作成した。単語のコスト等に関しては自動推定機能を用いた。

このシステムの評価に関しては、①実際の時事問題を入力し、本システムが抽出した語が問題の解答語句とどれだけ一致しているか（＝設問一致率)、また、設問になる可能性が高いと判断した語（＝重要語句）をどれだけ抽出できているか（＝重要語句一致率)、②新聞記事の本文を入力し、設問になる可能性が高いと判断した語をどれだけ抽出できているか、③日本史の教科書の本文を入力し、日本史用語集に収録されている語をどれだけ抽出できているかという 3 つの方法と基準に従い、4 種類の実験を行った。

上記①の実験では本研究で作成したユーザ辞書を用いた場合は総合設問一致率が 64.3%、総合重要語句一致率が 67.2%であり、ユーザ辞書を使わない場合（設問一致率が 17. 6%、重要語句一致率が 33.5%）と比べるとその一致率は格段に上昇した。よって、このような文章を用いて本システムを使えば就職活動生の勉強に役立てられると考えられる。②の実験では重要語句一致率は高かったが、同時に重要ではない語の抽出率も高くなってしまった。しかしこれは時事問題で出題される可能性が低い内容について書かれた記事も入力してしまったことも原因であるため、新聞記事を用いて本システムを使用する場合は、自分の求めている情報をはっきりさせて入力する必要がある。③の実験では、①の実験の一致率とさほど変わらない結果が出てきた。抽出した語のうち用語集に収録されていない語を詳しく見てみると、固有名詞ではあるが歴史的には重要でない語がほとんどであり、問題を回答する際にさほど悪影響が出るものではなかった。全ての実験を通して一致率を下げていた一番の要因は、複数の単語から成る固有名詞を必要以上に分割してしまうというところにあった。

今回の実験を通して判明した課題点としては、①表記の揺れや省略名称などの別解を提示することができていない、②スペースを含んだ固有名詞を抽出することができていないということが挙げられる。②の問題に対する解決策はまだ見つかっていないが、①の問題に対しては、同じものを指す語がリスト化されている辞書のようなものを作成し、それに含まれる語が抽出された場合はそのリストを元に別解を表示できるようにしたら良いのではないかと考えられる。

（指導教員　辻慶太）