逆文献頻度と文字の難易度に基づく単語レベル判定システムの構築

北村達也（甲南大学） 川村よし子（東京国際大学）

1. はじめに

本研究の目的は，単語レベル判定のための信頼できる基準を作り，その基準に基づいて単語レベル判定システムを構築することである。そして，日本語教師及び学習者がインターネットを介してこのシステムを利用し，教材のレベル判定に役立てられるようにする。我々はすでに日本語能力試験の出題基準，単語親密度，単語頻度を基準にした単語レベル判定システムを構築し，「リーディング・チュウ太（http://language.tiu.ac.jp/）」の一機能として公開している（川村, 1999, 2008）。本発表では，逆文献頻度（Inverse document frequency, IDF）（Jones, 1972）と単語中の文字の総画数が単語レベルの判定基準として使える否かを検討した。

2. 逆文献頻度に基づく単語レベル判定

逆文献頻度は文書集合における単語の重要度に関する指標である。文書集合において，ある単語が多くの文書に現れれば，その単語の逆文献頻度は小さくなる。$df(w_i,D)$を文書集合$D=(d_1,d_2,\ldots,d_N)$において単語w_iが出現する文書数とすると，その語の逆文献頻度は次式で与えられる。

$$idf(w_i,D)=\log\frac{N}{df(w_i,D)}$$

対数をとるのは，全体の文書数が変化しても逆文献頻度が大きく変化しないようにするためである。同じ単語であっても，逆文献頻度を計算する文書集合が異なれば逆文献頻度の値は異なる。

本研究では，2006 年毎日新聞データベースと国立情報学研究所により提供されている「Yahoo!知恵袋」データから逆文献頻度を求めた。後者はインターネット上の掲示板の文章であるため，話しことばに近い文体が使われている。小さい逆文献頻度を持つ単語ほど重要度が高いと仮定し，上位 12,000 語のリストを 2,000 語ずつ 6 段階の単語レベルに分

割した。毎日新聞データベースと Yahoo!知恵袋から求めた 2 種類の基準を用いて単語レベル判定システムを構築し，我々が過去に開発したシステム（川村, 1999, 2008）と比較した。その結果，新聞記事から求めた逆文献頻度は中上級教材向けの語彙の選定に適しており, Yahoo!知恵袋から求めた逆文献頻度は初級教材の語彙選定に適していることが明らかになった。これは，新聞記事と中上級教材は書きことばが中心であること，Yahoo!知恵袋と初級教材に話しことばに近いものが多く含まれていることに対応している（北村，川村, 2009，川村, 2009）。

3. 文字の難易度に基づく単語レベル判定

本研究では，難しい漢字は画数が多く，かつ単語の「文字面の」レベルが単語に含まれる文字数と漢字の難易度で規定できると仮定する。つまり，単語中の文字の総画数が多い単語ほどレベルの高い単語と考える。そして，単語レベルを単語中の文字の総画数，すなわち当該単語に含まれる文字の画数の総和で判定することを試みた。

漢字の画数のデータは Breen により公開されている kanjidic2 から得た。なお，現時点では平仮名及び片仮名の画数は 0 または 1 とし, いずれかを利用者に選択させることを想定している。この基準を用いて新聞記事に現れる単語のレベル判定を試みたところ，この基準は我々の単語難易度に関する感覚と近い点があるものの，単語の意味や用法，文体レベル等を考慮していないことによる不具合があることが明らかになった。

4. おわりに

我々の今後の課題は，新聞記事と Yahoo!知恵袋により求めた逆文献頻度を融合する方法，さらに単語中の文字の総画数と融合する方法を検討することである。これよって，よりよい単語レベル判定基準の構築が期待できる。なお，逆文献頻度に基づく単語レベル判定システムは，「リーディング・チュウ太」の web ページにて公開する予定である。

謝辞 本研究の一部は，（独）科学技術振興機構 平成 21 年度シーズ発掘試験（11-143）及び（独）日本学術振興会 平成 21 年度科学研究費補助金 基盤研究（B）（21320095）により実施された。