4P-05

やさしい日本語に書き換えるための単語難易度の推定

黎　明かん†　杉本　徹‡

芝浦工業大学 大学院理工学研究科†　芝浦工業大学 工学部‡

1. 研究背景と目的

我々は、ニュースや案内、お知らせといった日常的な文章をやさしい日本語に自動的に書き換えるシステムの開発を目指している。文をやさしい日本語に書き換える際に、文中の単語の難易度を判断して、難しい単語を易しい単語に置き換える必要がある。そこで本研究では、単語の難易度を推定する方法を研究する。

本研究では、まず日本語学習者にアンケートを実施して500個の日本語単語の難易度データを作成する。次に、単語の難易度に影響を与える要因を調査し、使用頻度や文字の複雑さ、漢字の読み方などの特徴量を付与する。最後に、これらの特徴量と単語難易度の相関分析、および回帰分析を行い、単語難易度の推定における各特徴量の有効性と課題を明らかにする。

2. 単語難易度データの作成

本研究では、日本語能力試験の語彙[1]の中からN1～N5それぞれ100語ずつ計500語を選び、難易度調査の対象語とした。日本語教師5人と日本語を学習中の学生42人の計47人を対象に、この500単語を提示し、それぞれの単語を知っているかどうか答えてもらった。なお、47人のうち42人は中国語を母国語とする者である。

本研究では単語の難易度を、アンケート回答者に占めるその単語を知っていると答えた人の割合と定義し、対象語500語の難易度を求めて単語難易度データとした。表1にデータ例を示す。

表1　単語難易度データの例

単語	難易度
不埒	0.255
危うい	0.638
英語	0.851

3. 研究方法

3.1単語難易度に影響を与える要因

日本語を第二外国語とする学習者にとって、語彙の難しさは文字の複雑さ、音読みと訓読み、学習者の心理特性、文化、母語などの影響を受ける。本研究では、客観的なデータとして求めることができる単語の使用頻度、親密度、文字の複雑さ、漢字が音読みか訓読みかを特徴量として各単語に付与する。

3.1.1 使用頻度

単語の出現頻度として、現代日本語書き言葉均衡コーパス（BCCWJ）の主要6コーパスにおける使用率[2]の平均、およびGoogle日本語Nグラム[3]の1-gram頻度データを採用する。

3.1.2 親密度

単語親密度[4]は、ある単語がどの程度なじみ深く感じられるかを表す指標である。親密度が高い単語ほど、多くの人が知っていると考えられるので、単語の難易度に関係があるといえる。

3.1.3 文字の複雑さ

文字の複雑さは単語の主要な内部特徴の一つで、単語の認知と記憶に大きな影響を与える。

本研究では、日本語の単語の書き方の複雑さを単語を構成するカタカナ、ひらがな、漢字の比率と画数から捉え、文字の複雑さKを定義する。

$K = K_1$カタカナ $+ K_2$ひらがな $+ K_3$漢字

K_1 = カタカナ文字数/単語の文字数
　　* カタカナ画数 * 0.2

Estimation of Word Difficulty for Easy Japanese Rewriting
†Li Minghan, ‡Toru Sugimoto
†Graduate School of Engineering and Science, Shibaura Institute of Technology
‡College of Engineering, Shibaura Institute of Technology

K_2 = ひらがな文字数/単語の文字数
　　* ひらがな画数 * 0.3
K_3 = 漢字文字数/単語の文字数 * 漢字画数 * 0.5

3.1.4 漢字が音読みか訓読みか

漢字の音読みは、もともと中国語における漢字の発音に基づく読み方であり、中国語を母語とする日本語学習者には覚えやすいと考えられる。

本研究では、単語の漢字が音読みか訓読みかを表す特徴量Oを考え、単語に含まれる漢字がすべて音読みの場合O=0、すべて訓読みの場合O=1、音読みと訓読みの両方が含まれる場合O=0.5とする。

3.2 分析の方法

3.2.1 相関分析

本研究で扱う使用頻度（BCCWJ、Google）、親密度、文字の複雑さ、音読みと訓読みの各特徴量が単語の難易度にどの程度影響するかを分析するために、各特徴量と単語難易度の間のピアソンの相関係数を計算する。

3.2.2 回帰分析

本研究では、単語の難易度を目的変数、各特徴量を説明変数として回帰分析を行う。

分析手法として、線形回帰、リッジ回帰、サポートベクトル回帰（linear, RBF）を用いる。

4. 実験結果と考察

3.1節で述べた単語の各特徴量と難易度の間の相関係数を表1に示す。使用頻度は、単語の使用回数そのものの他に対数をとった値も調べた。

表1　単語の特徴量と単語難易度の相関係数

単語の特徴量	相関係数
使用頻度（BCCWJ）	0.183
使用頻度（Google）	0.180
使用頻度（BCCWJ）対数	0.378
使用頻度（Google）対数	0.299
親密度	0.289
文字の複雑さ(K)	-0.049
音読みか訓読みか(O)	0.004
参考：日本語能力試験の級（1～5）	0.678

使用頻度は、使用回数の対数をとった値のほうが単語難易度との相関が大きくなった。そこで使用回数の対数値を他の特徴量と組み合わせて回帰分析を行った。結果を表2に示す。

表2　回帰分析における決定係数

説明変数	線形回帰	リッジ回帰	SVR (linear)	SVR (RBF)
BCCWJ対数 + 親密度	0.176	0.176	0.170	0.223
BCCWJ対数 + 親密度 + K + O	0.179	0.179	0.171	0.266
Google対数 + 親密度	0.116	0.116	0.111	0.389
Google対数 + 親密度 + K + O	0.122	0.122	0.112	0.439

表1から、単語の使用頻度と親密度は単語難易度と相関があることがわかる。

また表2から、単語難易度の推定において上記の特徴量に加えて文字の複雑さや漢字の読み方を考慮することが有効であることがわかる。

5. 結論

本研究では、日本語学習者を対象としたアンケート調査により500個の日本語単語の難易度データを作成し、難易度に影響を与えると考えられる4種類の特徴量を用いた相関分析および回帰分析を行った。

今後は、実験コーパスの範囲と量を拡大し、難易度でラベル付けされた単語の数を拡大することで、より正確な結果を得ることを目指す。

参考文献

[1] 红宝书大全集:新日本语能力考试N1-N5文字词汇详解, 华东理工大学出版社, 2015

[2] 特定領域研究「日本語コーパス」言語政策班, BCCWJ主要コーパス語彙表, 2011 https://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html

[3] グーグル株式会社，Web日本語Nグラム第1版, 言語資源協会, 2007

[4] 天野成昭，近藤公久，日本語の語彙特性，三省堂, 1999