4ZJ-4

日本語のやさしさの自動推定のための特徴量に関する基礎的検討

張　萌†，伊藤　彰則†，佐藤　和之‡
†東北大学大学院工学研究科　　‡弘前大学人文学部

1　はじめに

近年のグローバル化社会の進展に伴い，普段日本人が利用する日本語よりも簡単で，外国人にも理解が容易な「やさしい日本語」に注目が集まっている [1]．日本語が不慣れな外国人は，日本人が普段利用する日本語から正しい情報を読み取ることができないことも多く，特に記憶にも新しい東日本大震災などの災害時では，正確な情報伝達が生死を分ける場合もある．したがって外国人にも理解可能な日本語文章の作成が求めれている．

しかしながら，やさしい日本語の文章を作成することは容易ではない．なぜなら，外国人がどのような日本語をやさしいと感じるかを，日本人は分からないからである．つまり，日本人の難しいという感覚と，外国人の難しいという感覚は異なる．

そこで我々は，外国人にとっての日本語のやさしさを自動推定するシステムを検討する．日本語のやさしさを自動で推定できれば，日本人が「やさしい日本語」を作成する際の支援を行うことが可能である．本稿では,「やさしい日本語」の作成を支援するための作成ルールを参考にして [1]，日本語のやさしさに関連すると考えられる基準が自動推定のための特徴量として有効かどうかを調査したので報告する．

2　日本語のやさしさ自動推定システム

日本語のやさしさを自動推定するシステムは，ユーザが任意の日本語文章を入力した際に，日本語のやさしさのスコアを表示するシステムを想定している [2]．これにより,「やさしい日本語」を作成する際の支援を行うことが可能である．

本稿では，単純な線形回帰モデルにより，日本語のやさしさをモデル化する．ある日本語文章 s の日本語のやさしさのスコアを $E(s)$ とした場合，日本語のやさしさの推定は次の (1) 式に従う．

$$E(s) = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{F}(s) + w_0 \tag{1}$$

$\boldsymbol{F}(s)$ は日本語文章 s の特徴ベクトル，$\boldsymbol{W}$ および w_0 はモデルパラメータである．学習データからモデルパラメータをあらかじめ推定しておくことで，任意の s に対してスコアを得ることが可能となる．本稿では，主に $\boldsymbol{F}(s)$ の構成要素について焦点を当てる．

Basic Study of Feature for Automatic Estimation of Japanese Easiness
†Zhang MENG, †Akinori ITO　‡Kazuyuki SATO
†Graduate School of Engineering, Tohoku University
‡Faculty of Literature, Hirosaki University

3　日本語のやさしさに関連する特徴

3.1　やさしい日本語の作成基準

やさしい日本語の文章を作成する場合に，以下の基準に従って文章を作成することが有効と考えられている．

- 文章の構造を簡単にする．
- 難しい日本語の単語を使わない．
- 外来語を使わない．

本稿ではこの基準が，自動推定のため特徴量として有効かどうかを検討する．

3.2　文章構造に関する特徴量抽出

本稿では文章構造に関する特徴として，文の長さ，動詞の数，名詞の数，文節の数について検討する．

文の長さ (文を構成する単語の数) については，短いほどやさしい日本語であると考えられる．我々は，入力文章に対して形態素解析を行い，解析後の総形態素数を特徴量として利用する．同様に，名詞および動詞に関しても，形態素解析でそれぞれ名詞，動詞タグがついた単語の数を，特徴量として利用する．

文節数についても，少ないほどやさしい日本語であると考えられる．我々は，入力文章に対して係り受け解析を行い，解析後の文節数を特徴量とする．

3.3　単語レベルに関する特徴量抽出

単語レベルを特徴量化するために，日本語能力検定試験の語彙のレベルを利用する．約 8000 の単語に対して，1 級から 4 級までのレベルが与えられているので，入力文章に含まれる助詞，助動詞を除く各単語に対してレベルを求め，その平均値を特徴量とする．なおレベルが与えられていない単語に関しては，0 級を与えることにする．よって，単語レベルの特徴量の値が大きいほど，やさしい日本語であることが期待される．

3.4 外来語数の特徴量抽出

外来語については，形態素解析後の各形態素に対して、外来語かどうかを判断し，その総外来語数を特徴量とする．外来語かどうかの判断については，全ての文字がカタカナの場合，その形態素が外来語であるとみなすこととした．

4 評価実験

4.1 実験データ

実験データを収集するために，我々はNPO法人多文化共生マネージャー全国協議会の情報 [3] から，東日本大震災において外国人のために書かれた文章400文を抽出した．この400文の各文章に対して，日本語能力検定試験のレベルが3，4級程度の中国人留学生30人に，日本語のやさしさの評価を行ってもらった．日本語のやさしさの評価基準を次の表1に示す．

表 1: やさしい日本語の評価基準

評価基準	評価値
完全に理解できる	2
ちょっと理解できる	1
全然理解できない	0

これにより各文章に対して，30人からそれぞれ評価値が付与された．ここでは，30人の評価値の平均値を日本語のやさしさのスコアとする．このスコアが高いほど，やさしい日本語であると考えられる．

4.2 各特徴量の有効性と自動推定の評価

各特徴量の有効性を評価するために，実験データ全400文に対して，各特徴量と日本語のやさしさのスコアの相関を求めた．特徴量抽出において，形態素解析にはmecab-0.99[4]，係り受け解析にはcabocha-0.60[5]を使用した．その結果を以下の表2に示す．

表 2: 各特徴量と日本語のやさしさのスコアの相関

特徴量	相関値
文の長さ	-0.30
動詞の数	-0.39
名詞の数	-0.09
文節の数	-0.34
単語レベル	0.23
外来語の数	-0.26

表2の結果から，日本語のやさしさに関連すると考えられる基準は，名詞の数を除いて有効であることが分かった．

次に先ほど検討した6つの特徴量を利用して特徴ベクトルを構成し，実際に自動推定のモデルを構築して実験を行う．今回は，実験データを4分割し，3個を学習データとしてモデルパラメータ $\boldsymbol{W}$ を求めるのに利用し，残り1個を評価データとする4-fold cross-validationで実験を行った．その際，被験者による評価値とモデルによる推定値の相関を評価した．評価の結果から，0.54という相関値を達成した．その時の散布図を次の図1に示す．

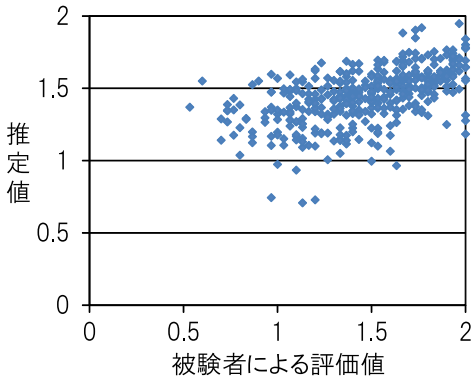

図 1: 被験者による評価値と推定値の散布図

この結果から，日本語のやさしさに関連すると考えられる基準を組み合わせることで，ある程度自動推定可能であるということが分かった．

5 まとめ

本稿では，日本語のやさしさに関連すると考えられる基準が自動推定のための特徴量として有効かどうかを調査した．その結果，各基準は有効であり，組み合わせることで，ある程度自動推定が可能であることが分かった．今後は，推定性能をさらに高める特徴について検討する予定である．

参考文献

[1] 「やさしい日本語」研究会編, “『やさしい日本語』が外国人の命を救う ”,「やさしい日本語」研究会, 2007.

[2] 伊藤彰則 他, “「やさしい日本語」作成支援システムの試作 ”, 電気関係学会東北支部連合大会, pp.299, 2008.

[3] http://tabumane.jimdo.com/

[4] T.Kudo, et al, “ Applying Conditional Random Fields to Japanese Morphological Analysis ”, In Proc. EMNLP, pp.230-237, 2004.

[5] T.Kudo , et al, “ Japanese Dependency Analysis using Cascaded Chunking ”, In Proc. COLING,pp.63-69, 2002.