Vol.2012-NL-206 No.6
Vol.2012-SLP-91 No.6
2012/5/10

「やさしい日本語」作成支援のための日本語の難易度自動推定の検討

張 萌[1,a)]　伊藤 彰則[1,b)]　佐藤 和之[2,c)]

概要：近年のグローバル化社会の進展に伴い，日本にいる外国人が増えつつある中で，日本人が普段使っている日本語よりも簡単で，日本語に不慣れな外国人にも理解が容易な「やさしい日本語」が注目されている．「やさしい日本語」で作成した文は，日本語に不慣れな外国人にとって有効であることが従来研究より示されているが，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ため，「やさしい日本語」の作成は容易ではない．そこで我々は，外国人の感覚に合った日本語の難易度自動推定について検討した．まず，難易度自動推定の枠組みをモデル化し，日本語の難易度に関連すると考えられる様々な特徴量を調査した．そして，実際に外国人に日本語の難易度の評価を行ってもらうことで得たデータを利用し，各特徴量と自動推定の枠組みの有効性の評価実験を行った．leave-one-out クロスバリデーションで評価を行ったところ，外国人の主観評価値と自動推定値の相関が約 0.66 を得た．

キーワード：やさしい日本語，日本語の難易度，自動推定，線形回帰モデル

Automatic Assessment of Easiness of Japanese for Writing Aid of “Easy Japanese”

Zhang Meng[1,a)]　Ito Akinori[1,b)]　Sato Kazuyuki[2,c)]

Abstract: As foreign population increasing for travel and short-term academic exchange in Japan, it is necessary to prompt easy Japanese language for them. Especially, when they are encountering unexpected disasters, such as earthquake and tsunami. It was proved that Easy Japanese is effective by previous studies. However, there is a problem for writing messages in Easy Japanese: it is difficult for a native Japanese speaker to understand what kind of Japanese sentences are easy or difficult for non-native speaker of Japanese. To solve this problem, we develop a method to assess the easiness of Japanese sentences This method is intended to be used as a writing aid of Easy Japanese (EJ). We examined many features about the easiness of Japanese, and used linear regression model to combine the features. As a result of evaluation experiment by a leave-one-out cross validation, we obtained correlation coefficient of 0.66 between the predicted scores and the easiness scores given by human subjects.

Keywords: Easy Japanese, Easiness of Japanese, Assessment of readability, Linear regression

1. はじめに

近年のグローバル化社会の進展に伴い，日本に住む外国人が増加しつつある中で，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].

日本語に不慣れな外国人は，日本人が普段利用する日本

[1] 東北大学大学院　工学研究科
School Graduate School of Engineering, Tohoku University
[2] 弘前大学　人文学部
Faculty of Literature, Hirosaki Unuversity
[a)] zhangm@spcom.eceitohoku.ac.jp
[b)] aito@spcom.eceitohoku.ac.jp
[c)] kazykis@cc.hirosaki-u.ac.jp

表 1 やさしい日本語の例
Table 1 Example of Easy Japanese

普通の日本語	やさしい日本語
火の元を確認してください。	ガスを消してください。
仙台市は断水や停電となり、市民の生活は麻痺しています。	仙台市は、水と電気が使えません。
直ちに高台に避難してください。	すぐに高いところに逃げてください。

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，公共の情報を受ける際に，不利な立場に置かれているといえる．平時であれば，情報提供者が複数の言語で情報を提供することによって，ある程度の配慮が可能であるが，大地震や津波，台風などの災害時にはそのような余裕はなく，ほとんどの情報が普通の日本語で提供されている．記憶にも新しい東日本大震災からも分かるように，正確な情報伝達が生死を分ける場合もあり，外国人にも理解可能な情報伝達方法は必須である．

そこで，日本に住んでいる外国人にも理解可能なやさしい日本語が有効となる．やさしい日本語は，日本語に不慣れな外国人のために考えられた日本語の表現方法であり，やさしい日本語を利用して情報を作成することで，1 つの言語表現で日本人にも外国人にも情報伝達を行えることが期待できる．

しかしながら問題点として，やさしい日本語を利用して文を作成することは容易ではないことが挙げられる．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っており，日本人が「やさしい」と考えて作成した文が，外国人には「難しい」といったことが起こり得る．この問題を解決するために，やさしい日本語の作成支援技術として，我々は外国人の感覚に合った日本語の難易度を自動推定する方法について検討する．

文の難易度を推定する研究は従来から研究されており，文の長さや単語の長さから，文の難しさを算出する方法などが検討されている [2], [3]．日本語文の難易度推定も行われているが，従来は基本的に日本人が感じる難易度を対象としている [4], [5]．よって，外国人の感覚に合った日本語の難易度は新たな課題といえる．そこで本稿では, まず日本語の難易度推定の枠組みをモデル化し，「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], 日本語の難易度に関連すると考えられる様々な特徴について検討する．そして，実際に外国人に日本語の難易度の評価を行ってもらうことで得たデータを利用し，各特徴量とモデルによる自動推定の枠組みの有効性を評価したので報告する．

2. やさしい日本語

2.1 やさしい日本語の概要

「やさしい日本語」とは, 普通の日本語よりも簡単で, 外国人にも理解が容易な日本語のことである．その発端は，1995 年の阪神淡路大震災である．阪神淡路大震災では，緊急情報のほとんどが一般的な日本語で書かれており，多くの外国人が困難を強いられた．これを受けて，1999 年に佐藤らにより「やさしい日本語」が提案された [1]. やさしい日本語は英語における Basic English[6] と似た考え方であり，母国語が異なる様々な外国人が統一的に理解できる日本語の記述方法である．やさしい日本語の実現は，多くの外国人の助けとなることが期待できる．

やさしい日本語は，日本語能力検定試験 3 級を合格した人が理解可能なレベルの表現を基本的に想定している．つまり，やさしい日本語の文は，日本語能力検定試験 3，4 級程度の語彙を使うことが望ましい．文法的にも可能な限り単文が望ましく，「〜わけではない」「〜ではないでしょうか」のようなあいまいな表現を避け，可能な限り直接的に表現する．やさしい日本語と普通の日本語を対比させた例を次の表 1 に示す．このように，文を簡易かつ直接的に表現することで，日本語に不慣れな外国人にも理解可能となる．実際に，大地震を想定した公開実験などから，やさしい日本語は普通の日本語と比べて有効であることが示されている [7].

2.2 やさしい日本語の作成支援

やさしい日本語の作成支援として，我々は「やんしす」(YAsasii Nihongo SIen System) というシステムを作成した [8].「やんしす」は，ユーザの入力文の単語や表現に対して，難しい部分を指摘するものであり，実際に多くの機関やボランティアの方に利用されている．

「やんしす」は,「やさしい日本語」の作成に有効と考えられる基準を人手で作成し，その基準を元にしたルールベースで難しい部分を指摘している．しかしながら，本当に外国人の感覚に合った指摘ができているかは分からず，また,「やんしす」は単語やフレーズレベルでしか難しい部分を指摘できない．文全体として外国人が難しいと感じるかどうかを指摘することが，やさしい日本語の作成支援としては理想的である．よって，やさしい日本語の作成支

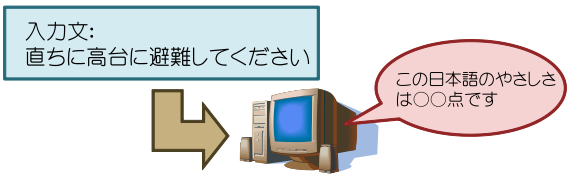

図 1 日本語の難易度自動推定システム
Fig. 1 System for automatic estimation

援の質を高めるためには，ユーザの入力文に対して，外国人の感覚に合った形で日本語の難易度を求める技術が必要となる．

3. 日本語の難易度自動推定

3.1 日本語の難易度自動推定システム

日本語の難易度自動推定技術は，ユーザが任意の日本語文章を入力した際に，日本語の難易度のスコアを表示するシステムとしての実現を想定している．想定するシステムの形を図 1 に示す．このように，文としてスコアを与えることで，ユーザは作成した文章を外国人が理解できるかどうかを判断できる．

3.2 日本語の難易度のモデル化

自動推定を行うために，日本語の難易度のモデル化を行う．我々は，線形回帰モデルにより，日本語の難易度をモデル化する．

ある日本語文 s に対し，外国人によって付与された難易度のスコアを $E(s)$ とする時，これを

$$E(s) = \sum_{k=1}^{K} w_k f_k(s) + w_0 + \sigma(s) = \boldsymbol{W}^\top \boldsymbol{f}(s) + \sigma(s) \quad (1)$$

のようにモデル化する．ここで，$\boldsymbol{W}$ はモデルパラメータ，$\boldsymbol{f}(s)$ は日本語文 s の特徴ベクトル，$\sigma(s)$ は予測誤差である．また，K は用いる特徴量の次元である．モデルパラメータ $\boldsymbol{W}$ は (2) 式の $N+1$ 次元のベクトルである．

$$\boldsymbol{W} = \begin{bmatrix} w_0 & w_1 & w_2 & \cdots & w_K \end{bmatrix}^\top \quad (2)$$

同様に，特徴ベクトル $\boldsymbol{f}(s)$ も (3) 式の $N+1$ 次元のベクトルで表される．

$$\boldsymbol{f}(s) = \begin{bmatrix} 1 & f_1(s) & f_2(s) & \cdots & f_K(s) \end{bmatrix}^\top \quad (3)$$

モデルパラメータ $\boldsymbol{W}$ は，誤差 $\sigma(s)$ 学習データについての二乗和が最小になるように推定される．本稿では，リッジ回帰によりモデルパラメータを推定する [9]．リッジ回帰では，通常の 2 乗誤差最小基準に L2 ノルムによるペナルティ項を加えて目的関数を構成する．これにより，学習データに対するオーバーフィッティングを防ぐことが可能である．文章と日本語の難しさの組が N 個与えられた場合，次の (4) 式，(5) 式のようにベクトルを定義する．

$$\boldsymbol{F} = \begin{bmatrix} 1 & f_1(s_1) & f_2(s_1) & \cdots & f_K(s_1) \\ 1 & f_1(s_2) & f_2(s_2) & \cdots & f_K(s_2) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & f_1(s_N) & f_2(s_N) & \cdots & f_K(s_N) \end{bmatrix} \quad (4)$$

$$\boldsymbol{E} = \begin{bmatrix} E(s_1) & E(s_2) & \cdots & E(s_K) \end{bmatrix}^\top \quad (5)$$

この時，リッジ回帰によるモデルパラメータの推定は次の (6) 式に従う．

$$\hat{\boldsymbol{W}} = (\boldsymbol{F}^\top \boldsymbol{F} + k\boldsymbol{I})^{-1} \boldsymbol{F}^\top \boldsymbol{E} \quad (6)$$

$k(>0)$ はリッジパラメータであり，k によりペナルティ項の大きさを調整できる．$k=0$ の時は，2 乗誤差最小基準による推定と等価である．なお，$\boldsymbol{I}$ は単位行列である．

以上の枠組みにより，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，任意の文 s に対してスコアを得ることが可能となる．

4. 日本語の難易度に関連する特徴

4.1 やさしい日本語の作成基準

次に，$\boldsymbol{f}(s)$ を構成する特徴量について検討を行う．特徴量の検討に際して，我々はやさしい日本語の作成ルールを参考にする [1]．やさしい日本語を作成する場合に, 以下の基準に従って日本語文を作成することが有効と考えられている．

(1) 文の構造を簡単にする
(2) 難しい日本語の単語を使わない
(3) 外来語を使わない [10]

この他に，文に含まれる平仮名や漢字の文字シンボルの割合も関係があると考えられる．本稿では，この四つの基準に関係があると考えられる特徴量について検討を行った．

4.2 文の構造に関する特徴量抽出

本稿では文構造に関する特徴量として, 文の長さ，各品詞の数，各品詞の割合，文節数，係り受けの距離，係り受けの回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，各品詞の数に関しては，形態素解析でそれぞれの品詞タグがついた単語の数を利用する．品詞の割合は，品詞の数を文の長さで割ったものである．具体的には，本稿では名詞と動詞について検討した．

文節数についても文の長さと同様に，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して利用

表 2 日本語能力検定試験における語彙のレベル
Table 2 Vocabulary of Japanese Language Proficiency Test

語彙レベル	種類数
1	3025
2	3771
3	718
4	791

表 3 やさしい日本語の評価基準
Table 3 Evaluation words and the values for the Easy Japanese

評価基準	評価値
完全に分かる	2
ちょっと理解できる	1
全然分からない	0

する．係り受け解析は，文節間の「修飾する」及び「修飾される」の関係のことである．この修飾関係が離れている場合は，理解が困難である．よって，2 単語の係り受けの関係を距離として算出し，文全体の最大値および平均値を特徴量として利用する．また，1 つの単語が複数の単語から修飾されている場合は，意味的に曖昧性が生まれる．よって、1 単語が修飾されている回数の最大値も特徴量として利用する．

4.3 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，(旧) 日本語能力検定試験の語彙のレベルを利用する．これは約 8000 の単語に対して，単語レベルが付与されたものである．日本語の語彙レベルを次の表 2 に示す．

これを利用して，入力文から単語レベルに関する特徴量を抽出する．まず，入力文章に対して，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．さらに，それぞれのレベルの単語数，1 文に含まれるそれぞれのレベルの割合も特徴量として検討する．レベル 1 の単語が多ければ難しく，レベル 4 の単語が多ければ簡単になると考えられる．

4.4 外来語に関する特徴量抽出

外来語は，実際に欧米で使われる意味や発音とは異なることが多い．また，外来語は日本語を勉強する外国人にとっては，特に難しいことが示されている [11]．本稿では，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，文に含まれる外来語の数および，外来語の割合を特徴量として検討する．外来語かどうかの判断については，全ての文字シンボルがカタカナの形態素を，外来語であるとみなすこととした．

4.5 文字シンボルに関する特徴量抽出

日本語文には大きく，漢字，ひらがな，カタカナの 3 種類の特徴的な文字シンボルが利用される．これらは，日本語のやさしさに関係があると考えられる．よって，文章に含まれる漢字，ひらがな，カタカナ，漢字，それぞれの割合を特徴量として検討する．

5. 評価実験

5.1 実験データ

本稿で検討したことの有効性を確かめるために実験を行う．実験データとして，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 文を抽出して利用した [12]．この 400 文の各文章に対して，中国人留学生 30 人に，日本語の難易度の評価を行ってもらった．日本語の難易度の評価基準を次の表 1 に示す．

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難易度の主観評価値とする．この主観評価値が高いほど，やさしい日本語であると考えられる．今回利用した実験データの例と，その主観評価値を表 2 に示す．

5.2 各特徴量の有効性の評価

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難易度の主観評価値との相関を求めた．特徴量抽出において，形態素解析には mecab-0.99[13]，係り受け解析には cabocha-0.60[14] を使用した．その結果を図 2 に示す．

図 2 の結果から，有効な特徴量と有効ではない特徴量があることが分かる．名詞の数，レベル 1 の単語数，レベル 3 の単語の割合，レベル 4 の単語数，漢字の数，の 5 つの特徴量を除き，有意水準 5 %で有意であった．今回は，有意であった特徴のみを自動推定のための特徴ベクトルの構成要素として利用する．

考察としては，基本的には，文の長さや平均レベル，レベル 4 の単語の割合，外来語の数など，考えた通りの傾向が出ていることが分かる．今回最も相関があったのはひらがなの数であった．これは漢字の割合と大きく関係がある．つまり，漢字の割合が多くひらがなが少ないほど，文の難易度が低いという結果である．これは，今回評価を行ってもらった対象が，中国人留学生であったことに起因する．中国人は漢字圏であり，漢字だけから意味を推測することが可能である．逆にひらがなやカタカナは中国人にとっては難しいということがいえる．

今回の結果から，文の構造，単語レベル，外来語，それぞれに関する特徴は，母国語非依存の一般的な特徴と考えられるが，文字シンボルに関する特徴は，母国語依存であ

表 4 実験に用いた文とその主観評価値の例
Table 4 Examples of Sentences for an experiment and that of subjective evaluation

例文	主観評価値
避難所にいる人は病気になりやすいです。	1.8
もし警察や市役所の人に在留期間について聞かれたときは下のＵＲＬのページをその人に見せてください。	1.5
食費、宿泊費、交通費は自己負担です。	1.1
妊娠している人と赤ちゃんのために大切なことですから、勇気を出して相談してください。	1.06
そして妊娠中は、おりものの量が増えます。	0.6
ティッシュは耳栓代わりに、タオルはアイマスク代わりに使えます。	0.53

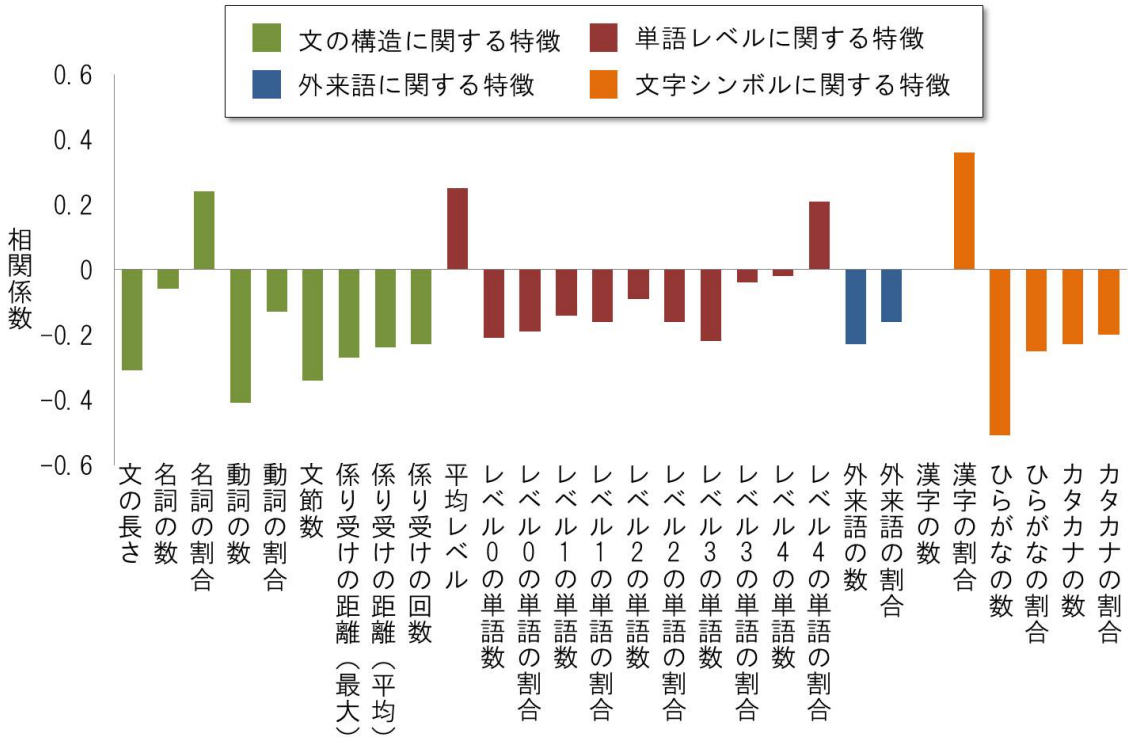

図 2 各特徴量と日本語の難易度のスコアの相関
Fig. 2 Correlation coefficeint between linguistic features and subjective evaluation

ることが示唆された．

5.3 自動推定の評価

前述の検討において，統計的に有意であった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，closed, 及び open で評価実験を行った．

closed 条件として，全 400 文の実験データを学習データと評価データ両方に用いる．全 400 文を用いて，モデルパラメータ $\boldsymbol{W}$ を推定し，同様のデータに対して推定値を求め，その際の主観評価値とモデルによる推定値の相関を評価した．なおこの時のリッジパラメータは $k = 0$ とし，モデルパラメータを推定した．評価の結果から，0.70 という相関値を達成した．その時の散布図を次の図 3 に示す．

次に open 条件として，実験データ D を 400 分割し，399 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，leave-one-out クロスバリデーションにより実験を行った．その際，被験者による評価値とモデルにより推定されたスコアの相関を評価した．まず，リッジ回帰によるモデルパラメータの推定がオーバーフィッティングを軽減に有効かどうかを調査するために，リッジパラメータ k を変化させて実験を行った．その結果を次の図 4 に示す．

この結果から，k の調整により，$k = 0$ の場合よりも相関が上がることが分かる．つまり，2 乗誤差最小基準の場合よりもリッジ回帰による推定が有効であることが分かる．$k = 0.2$ の時に最も相関が高くなり，0.66 を得た．この時の散布図を次の図 5 に示す．

この結果から，日本語の難易度に関連すると考えられる基準を組み合わせることで，日本語の難易度をある程度自

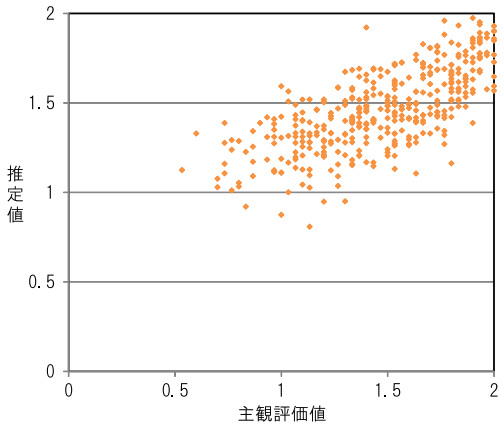

図 3 推定値と主観評価値の散布図 (closed)

Fig. 3 Comparison of System output and subjective evaluations(closed)

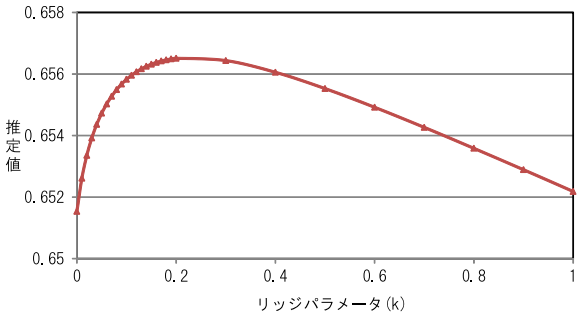

図 4 リッジパラメータと相関の関係

Fig. 4 Relations between ridge parameter and correlation

動推定可能であるということが分かった.

6. まとめ

本稿では，やさしい日本語の作成支援の質を高めるために，外国人の感覚に合った形で日本語の難易度を自動推定する技術について検討した．まず日本語の難易度推定の枠組みをモデル化し，「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，日本語の難易度に関連すると考えられる様々な特徴について検討した．そして，実際に外国人に日本語の難易度の評価を行ってもらうことで得たデータを利用し，各特徴量とモデルによる自動推定の枠組みの有効性を評価した．その結果，どの特徴量が有効であるかが分かった．また，各特徴量を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．また，今回は中国人だけを対象として実験をしたが，他の国の人も対象とすることで，特徴量の母国語依存性についても検討する予定である．

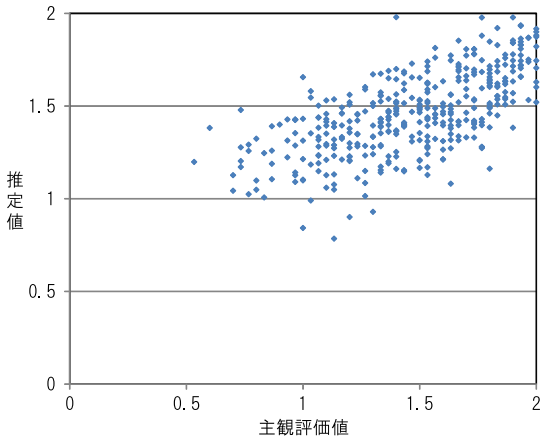

図 5 推定値と主観評価値の散布図 (open)

Fig. 5 Comparison of System output and subjective evaluations(open)

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] R. Flesch, “A new readability yardstick”, Journal of Ap-plied Psychology, Vol 32, No. 3, pp.221-233, 1948.
[3] P. B. Mosenthal and I. S. Kirsch, “A new measure for assessing document complexity: The PMOSE/IKIRSCH document readability formula”, Journal of Adolescent & Adult Literacy, Vol. 41, No. 8, pp.638-657, 1998.
[4] S. Sato, S. Matsuyoshi and Y. Kondoh, “Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus,” Proc. LREC, pp. 654-660, 2008.
[5] 長谷川優，山村毅,“マハラノビス距離を用いた日本語文章の難易度判定システムの提案”, 電気情報通信学会論文誌 D, Vol.J94-D, No.9, pp.1589-1592, 2011.
[6] C.K.Ogden, “Basic English as an international second language”, Harcourt, Brace&World,1968.
[7] http://human.cc.hirosaki-u.ac.jp/kokugo/EJ5yuukousei.htm
[8] 伊藤彰則，鹿嶋彰，前田理佳子，水野義道，御園生保子，米田正人，佐藤和之,“「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.
[9] A.E.Hoerl, R.W.Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems”, Technometrics 12, pp.55-67, 1970.
[10] F. E. Daulton, “English Loanwords in Japanese . The Built-In Lexicon”, The Internet TESL Journal, Vol. V, No.1, 1999.
[11] E. Lovely, “Learner ’s Strategies for Transliterating English Loanwords into Katakana”, New Voices, Vol. 4, pp.100-122, 2011.
[12] Earthquake Information, `http://eqinfojp.net/?page_id=66`
[13] K. Yamamoto, Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis”, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004.
[14] T. Kudo, Y Matsumoto,“ Japanese Dependency Analysis using Cascaded Chunking ”, Proceedings of the 6th Conference in Natural Language Learning 2002 (COLING-2002), pp. 63-69, 2002.