The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

再現が容易な単語の平易化判定手法

A Simple Reproducible Judgment Method for Lexical Simplification based on Word Difficulty

高田　理功 *1　　梶原　智之 *2　　奥村　紀之 *1
Takada Riku　　Kajiwara Tomoyuki　　Okumura Noriyuki

*1明石工業高等専門学校　　*2大阪大学
National Institute of Technology, Akashi College　　Osaka University

A lexical simplification is one of the important techniques for leaners of second languages. In this paper, we propose a simple reproducible judgment method for lexical simplification without using probabilistic model. In addition, our proposed method can simplify vocabularies to easier vocabularies at a significant advantage to previous researches.

1. はじめに

グローバル化が進む中で，日本語学習者が日本人だけでなく外国人を含むようになっている．しかし，日本語学習者のうち，日本語初学者である子供や外国人は難解な単語の意味を理解することができない．そのため，日本語初学者は，文が平易な単語で構成されていると読みやすい．それを実現する技術として，自動的に難解な単語を平易な単語に言い換える語彙平易化という技術がある．語彙平易化は言い換え元と言い換え先の単語のどちらが簡単か判定する平易化判定を行う必要がある．そのため，本稿は平易化の基盤である単語の平易化判定を行う．

先行研究として，梶原ら [梶原 17] がある．梶原らの手法は素性に単語分散表現の 1 つである Word2Vec[Mikolov 13] を用いている．しかし，単語分散表現は学習結果が初期値の乱数に依存するため再現が困難であり，かつ平易化判定時にどの素性が有用か判断することができない．

そこで我々は Word2Vec がどの程度初期値の乱数に依存するか調査する．また，文字数，文字種，頻度のように結果の解釈が容易な素性のみを用いる平易化判定手法を構築する．

本稿の結果，我々の平易化判定手法は，先行研究よりも accuracy が 6.5 ポイント改善し，かつ再現が容易な手法と分かった．また素性として，日本語 Wikipedia 中の文字単位の頻度と BCCWJ 中の単語単位の頻度が有用であることが分かった．

2. 関連研究

語彙平易化では，入出力の単語が以下の 2 つを満たすように変換される．

[**同義性**] 言い換え先の単語が言い換え元の単語の意味を含有する
[**平易化**] 言い換え先の単語が言い換え元の単語より平易である

同義性に関する研究として，PPDB[Ganitkevitch 13] がある．PPDB は言い換え辞書を自動生成する研究であり，英語をある言語に翻訳した後，再度翻訳して英語に戻すことで，言い換え元の意味と同義なフレーズを取得できるとし構築された．また日本語で同様に自動作成された言い換え辞書として PPDB: Japanese*1[Mizukami 14] がある．PPDB: Japanese は行ごとに言い換え元，言い換え先，言い換え確率などが書かれている言い換え辞書である．PPDB: Japanese は言い換え確率の閾値が異なる 5 つのサイズが公開されており，本稿ではその中で最も言い換え対が多く収録されている 10best を利用している．

平易性に関する研究として，Simple PPDB[Pavlick 16] がある．Pavlick らは，PPDB を改良した PPDB: 2.0[Pavlick 15] を元に平易化判定を行い，平易化用の大規模な言い換え辞書 Simple PPDB を構築した．Pavlick らの平易化判定は，フレーズから抽出できる素性を利用した，ロジスティック回帰モデルを用いるモデルである．

これと同様に日本語の平易化用の大規模な言い換え辞書を構築した例として，Simple PPDB: Japanese*2[梶原 17] がある．梶原らは以下の 2 つの手順で平易化判定を実現した．

[**難易度推定**] 言い換え元の単語と言い換え先の単語の難易度を推定する
[**平易化判定**] 言い換え元の単語と言い換え先の単語の難易度差を算出する

難易度推定には RBF カーネルを採用した SVM を用いており，素性には日本語 Wikipedia*3 中に出現する単語の頻度, 単語長, 文字種 (ひらがな，カタカナ，漢字), 日本語 Wikipedia 上の Word2Vec[Mikolov 13] による単語分散表現を用いている．最も高い accuracy を示した Word2Vec は，100 次元の skip-gram with negative sampling モデルであった．

3. 単語分散表現の再現の容易性

2 節のように，梶原らは単語分散表現を用いている．しかし単語分散表現は，学習結果が初期値の乱数に依存するという問題が存在する．どの程度学習結果に初期値の乱数が依存するか調査するため，以下の手順で実験を行った．

本稿では，梶原らと同様に，MeCab*4 と mecab-ipadic-NEologd[佐藤 17] によって単語分割された日本語 Wikipedia の本文を用いて，100 次元の skip-gram with negative sampling モデルを学習した．同一のテキストデータから構築され

連絡先: 高田 理功，明石工業高等専門学校，兵庫県明石市魚住町西岡 679-3,e1319@s.akashi.ac.jp

*1 http://ahclab-server.naist.jp/resource/jppdb/
*2 https://github.com/tmu-nlp/simple-jppdb
*3 https://dumps.wikimedia.org/jawiki/20161001/
*4 http://taku910.github.io/mecab/

る複数の単語分散表現間の差について調査するため，初期値以外のパラメータを統一して単語分散表現の学習を二度（単語分散表現 A および B）行った．

3.1 コサイン類似度

単語分散表現 A と単語分散表現 B がどの程度一致するかを確認するため，日本語 Wikipedia から無作為抽出した 1 万種類の単語について，単語分散表現 A-B 間のコサイン類似度を測定した．単語分散表現の学習およびコサイン類似度の測定を 5 回実施し，その平均値を表 1 に示す．

表 1: 再現による各単語のコサイン類似度

実験回数	コサイン類似度
1 回目	-0.004
2 回目	-0.001
3 回目	-0.005
4 回目	-0.008
5 回目	-0.002
平均	-0.004

実験の結果，コサイン類似度は平均-0.004 であり，ベクトルがほとんど一致していないことがわかる．そのため，初期値の乱数を指定しない場合，単語分散表現を再現することは困難であると言える．

3.2 梶原らの再現実験

梶原らの手法を単語分散表現 A，単語分散表現 B それぞれを用いて再現した．単語分散表現 A および B それぞれを用いて再現実験を行うことで，分散表現によりどの程度予測に差が生じるのか確認することができる．単語分散表現 A および B を用いて再現実験した結果の accuracy を表 2 に示す．なお，単語の難易度推定の評価は日本語教育語彙表 [*5] のうち，単語分散表現 A および B に存在する 16,421 単語でテストを行っている．また，平易化判定の評価は PPDB: Japanese の 10best うち上記の難易度予測が可能であった単語 1 語ずつで構成されている 47,423 単語対でテストした．

表 2: 梶原らの再現実験の accuracy

手法	難易度推定	平易化判定
梶原ら [梶原 17]	0.709	0.607
単語分散表現 A	0.652	0.649
単語分散表現 B	0.651	0.651

再現実験によると，難易度推定，平易化判定ともに梶原らの手法とは 5 ポイント程度 accuracy が異なる．また，単語分散表現 A-B 間は accuracy に差がない．

次に，単語分散表現 A と単語分散表現 B の，難易度推定を行った予測単語のうち，予測が一致しない単語数を調査した．結果，単語分散表現 A-B 間で難易度予測が異なった単語数は，715 単語存在した．これは，16,421 単語中の約 4.4%に該当する．平易化判定も同様に予測が一致しない単語対数を調査したところ，判定結果が異なる対は 1,954 単語対存在した．これは，47,423 単語中の約 4.1%に該当する．よって accuracy に差が生じるとは限らないが，実験ごとに単語の予測難易度の一部が変化することがわかる．

*5 http://jhlee.sakura.ne.jp/JEV.html

4. 提案手法

先行研究である梶原らの研究では素性の 1 つに Word2Vec が用いられている．また，Word2Vec は 3 節で述べたように，学習の初期値に乱数を利用するため，実験の再現が困難である．そのため本稿では，実験の再現が容易なモデルを提案する．

本稿で提案するモデルは梶原らの手法 [梶原 17] を踏襲し，以下の手順で単語の平易化判定を行う．

[**難易度推定**] 言い換え元の単語と言い換え先の単語の難易度を推定する

[**平易化判定**] 言い換え元の単語と言い換え先の単語の難易度差を算出する

本稿では推定する難易度は 3 段階と 6 段階の 2 種類を用意した．3 段階難易度は初級，中級，上級であり，6 段階難易度の場合は 3 段階の各難易度を前半と後半に分ける．難易度推定の際は数値に対応付けるため，平易なほど小さい値をとるようにし，難解なほど大きい値を取るようにする．3 段階難易度の場合，難易度が初級のものを難易度 1 とし，上級のものを難易度 3 とした．6 段階難易度の場合，難易度が初級前半のものを難易度 1 とし，上級後半のものを難易度 6 とした．単語の平易化判定では，言い換え元の単語の難易度が，言い換え先の単語の難易度より大きい場合を平易，同じ場合を同等，小さい場合を難解とする．

単語の難易度推定モデルは，梶原らと同じ RBF カーネルによる SVM を用いる．また多クラス分類への拡張は one-versus-rest 法で行った．SVM は sklearn(0.19.1) の SVC[*6] を用いた．ハイパーパラメータである C と gamma はグリッドサーチを用いて最適解を求めた．

4.1 素性

素性は単語長，文字種 (ひらがな，カタカナ，漢字)，日本語 Wikipedia 中の単語単位の頻度，日本語 Wikipedia 中の文字単位の頻度，BCCWJ 中の単語単位の頻度を用いた．以下で各素性について説明する．

単語長とは単語の文字列としての長さであり，1 次元の素性である．単語長は，Flesch Reading Ease[Flesch 48] に単語長の平均が用いられているように，可読性を測る有用な素性になると考えられる．

文字種は以下の 2 パターンを用意した．どちらの場合も，素性の次元数は 3 次元である．

[**文字種の有無**] 単語に含まれるひらがな，カタカナ，漢字の有無 (1，0)

[**文字種の文字数**] 単語に含まれるひらがな，カタカナ，漢字の文字数

日本語 Wikipedia 中の単語単位の頻度とは，MeCab と mecab-ipadic-NEologd によって日本語 Wikipedia を単語分割した際の各単語の対数頻度で，素性の次元数は 1 次元である．頻度情報は，SemEval-2012 の English Lexical Simplification タスクで Baseline として用いられたが，優れたスコアを示したことなどから，有用な素性になると考えられる．Pavlick らの研究 [Pavlick 16]，梶原らの研究 [梶原 17] においても Wikipedia 中の単語単位の頻度を素性として利用している．なお，対数頻度において出現回数が 0 回のとき，対数頻度を 0 としている．

日本語 Wikipedia 中の文字単位の頻度は，単語に含まれる各文字単位の日本語 Wikipedia 中の頻度である．文字単位の

*6 http://scikit-learn.org/0.19/modules/generated/sklearn.svm.SVC.html

頻度は，平易な単語に用いられやすいひらがなやカタカナは頻度が大きくなると仮定する．一方漢字の頻度は低くなり，特に難解語に用いられる漢字ほど頻度が低くなると仮定する．

文字単位の頻度は，「子犬」という単語に対して，「子」「犬」それぞれに頻度が存在する．このように既出の各素性と異なり，単語長に依存して素性の数が変化する．しかし SVM の入力素性は固定長にしか対応しない．素性の数を単語長に依存させないようにするため，素性は文字単位の頻度の合計，平均，最大，最小の 4 パターンを用意し，対数化した後素性として加えた．よって素性の次元数は 4 次元である．

BCCWJ 中の単語単位の頻度は，現代日本語書き言葉均衡コーパス (BCCWJ)[*7][Maekawa 10] 中の単語単位の対数頻度である．BCCWJ は書籍全般，雑誌全般，ブログなどの様々なジャンルにまたがって無作為にサンプリングされた均衡コーパスである．BCCWJ 自体は限定公開であるため，BCCWJ に含まれる各ジャンルごとの頻度や全体における頻度などをまとめた『現代日本語書き言葉均衡コーパス』語彙表 ver1.0[*8] を用いて頻度を取得した．語彙表は 2 種類の単語区切りが公開されており，本稿では固有表現に対応する長単位を選択した．

語彙表に掲載されている情報のうち素性として用いるのは，ジャンルごとの頻度の対数とした．可変長と固定長があるものは合計された頻度のものを用いる．BCCWJ 中の単語単位の頻度は，13 種類のジャンルごとの頻度と，全体における頻度の合計 14 次元の素性となる．

4.2 素性の組み合わせ

本稿で用いる素性は単語長，文字種の有無と文字数，日本語 Wikipedia 中の単語単位の頻度，日本語 Wikipedia 中の文字単位の頻度，BCCWJ 中の単語単位の頻度である．しかし，これらの全てが有用な素性であるとは限らない．そこで SVM に入力する素性の構成を複数パターン用意した．素性の組み合わせとして用意した baseline1 および 2 と，提案手法 1 から提案手法 8 までを以下に箇条書きで紹介する．なお基本素性 1 は Wikipedia 中の頻度，単語長，文字種の有無で，基本素性 2 は Wikipedia 中の頻度，単語長，文字種の文字数である．

[**baseline1**] Wikipedia 中の頻度
[**baseline2**] 基本素性 1
[**提案手法 1**] 基本素性 2
[**提案手法 2**] 基本素性 1，文字単位の頻度
[**提案手法 3**] 基本素性 2，文字単位の頻度
[**提案手法 4**] Wikipedia 中の頻度，BCCWJ 中の頻度
[**提案手法 5**] 基本素性 1，BCCWJ 中の頻度
[**提案手法 6**] 基本素性 2，BCCWJ 中の頻度
[**提案手法 7**] 基本素性 1，文字単位の頻度，BCCWJ 中の頻度
[**提案手法 8**] 基本素性 2，文字単位の頻度，BCCWJ 中の頻度

baseline1 は単語の対数頻度のみを素性として用いた．基本素性 1 は梶原らの論文 [梶原 17] で提案されている基本素性と同じ素性である．梶原らの論文で有用な素性として発表されているため，これを基本素性 1 とした．また，基本素性 1 のみを素性とする手法を baseline2 とした．

基本素性 2 は梶原らの論文で用いられている文字種の真理値の代わりに文字種の文字数を素性としたものである．文字種の文字数を素性とすることで，単語長の情報を含んだ素性となり，accuracy の向上を期待する．

*7 http://pj.ninjal.ac.jp/corpus_center/bccwj/index.html
*8 http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html

基本素性 1 と基本素性 2 に対して，Wikipedia 中の文字単位の頻度，BCCWJ 中の単語単位の頻度をそれぞれ加えたものと加えないものを用意した．この素性の構成パターンにより，基本素性 1，基本素性 2，日本語 Wikipedia 中の文字単位の頻度，BCCWJ 中の単語単位の頻度それぞれが有用な素性か判断することができる．

5. 実験

提案手法で述べたモデルを用いて，実際に単語の難易度推定と平易化判定を行い，提案手法を評価する．また，再現が容易か調査した．

5.1 単語の難易度推定

4.2 節に示した各素性の組み合わせで，SVM を用いた単語の難易度推定を行う．本稿では教師データに日本語教育語彙表を用いた．日本語教育語彙表は日本語教育用の語彙である 17,920 単語の表記，読み，語彙の難易度などの情報を収録した表である．各語彙の難易度は 5 人の匿名教師により初級前半，初級後半，中級前半，中級後半，上級前半，上級後半の 6 段階難易度で設定されている．

推定する単語の難易度は 6 段階難易度と初級，中級，上級の 3 段階難易度の 2 種類である．評価は 10 分割交差検定を用いて行った．

日本語教育語彙表を用いた baseline1 および 2 と提案手法 1 から提案手法 8 の単語の難易度推定を行った．10 分割交差検定での難易度推定の accuracy を表 3 に示す．

表 3: 難易度推定の accuracy

手法	3 段階難易度	6 段階難易度
梶原ら [梶原 17]	**0.709**	-
baseline1	0.557	0.411
baseline2	0.572	0.420
提案手法 1	0.575	0.419
提案手法 2	0.587	0.432
提案手法 3	0.585	0.432
提案手法 4	0.624	0.458
提案手法 5	0.633	0.465
提案手法 6	0.636	0.470
提案手法 7	0.626	0.415
提案手法 8	**0.644**	**0.478**

表 3 からわかるように，日本語 Wikipedia 中の文字単位の頻度，BCCWJ 中の単語単位の頻度を素性として用いるのは有用である．特に提案手法 3 と提案手法 8 に着目すると，BCCWJ 中の単語単位の頻度を用いることで，BCCWJ 中の単語単位の頻度を含まない同様の手法より 5.9 ポイント accuracy が上がることを確認できる．提案手法 1 から提案手法 6 までを見ると基本素性 1 と基本素性 2 ではどちらを用いても accuracy が大きくは変わらないが，提案手法 7 と提案手法 8 には accuracy に大きな差が生じた．

最も優れた accuracy を示した手法は提案手法 8 で，3 段階難易度の場合 baseline1 より 8.7 ポイント，6 段階難易度の場合 baseline1 より 6.7 ポイント高い accuracy となっている．先行研究の梶原らの手法と比べると提案手法 8 は accuracy が 6.5 ポイント低い．しかし，Wikipedia 中に含まれない単語でも対数頻度を 0 として予測するため，より予測できる単語の総数は増加している．

5.2 単語の平易化判定

5.1 節の難易度推定の結果を利用して単語の平易化判定を行った．本稿では平易化判定のための単語対を PPDB: Japanese の 10best データから取得した．PPDB: Japanese の 10best データのうち，言い換え先と言い換え元が日本語教育語彙表に載っていて，かつ単語 1 語で構成されている単語対を評価に利用する．以上の手順で取得できた単語対は 48,646 対存在する．平易化が可能か判定するため，単語の難易度差を言い換え先より平易，同等，難解の 3 段階で算出した．なお難易度は 3 段階難易度により平易化判定したもの，6 段階難易度により平易化判定したものの 2 種類を用意した．評価データは日本語教育語彙表による難易度差を利用する．表 4 に，以上の方法で単語の平易化判定を評価した結果を示す．

表 4: 平易化判定の accuracy

手法	3 段階難易度	6 段階難易度
梶原ら [梶原 17]	**0.607**	-
baseline1	0.614	0.597
baseline2	0.625	0.600
提案手法 1	0.624	0.597
提案手法 2	0.625	0.578
提案手法 3	0.625	0.576
提案手法 4	0.661	0.652
提案手法 5	0.668	0.658
提案手法 6	0.668	0.658
提案手法 7	0.670	0.625
提案手法 8	**0.672**	**0.662**

単語の難易度推定と同じく，最も優れた accuracy が得られたのは提案手法 8 であった．3 段階難易度の場合 baseline1 より 5.6 ポイント，6 段階難易度の場合 6.5 ポイント向上したことがわかる．梶原らの手法は 3 段階難易度による平易化判定で accuracy が 0.607 であるため，提案手法 8 は梶原らの手法と比べて accuracy が 6.5 ポイント改善していることがわかる．ただし本稿では，対数頻度 0 の単語の難易度を予測することでより多くの単語対に難易度を付与している．そのため，baseline を含め先行研究と単純には比較できない．

また，BCCWJ 中の頻度を素性とした判定が，BCCWJ 中の頻度を用いない手法より 4 ポイント程度高い accuracy を示すことがわかる．基本素性 1 と基本素性 2 では最大でも 0.3 ポイントしか accuracy に差が生じないため，文字種の文字数を素性としても accuracy は向上しない．

5.3 再現の容易性の評価

我々は accuracy が高いだけでなく，再現性が高いモデルの構築を目指している．そのため，最も accuracy の高かった提案手法 8 を用いて，全く同じ実験を 5 回繰り返し再現の容易性を確かめた．

結果，5 回とも難易度推定において全ての単語の予測難易度が一致した．また，平易化判定においても難易度推定と同様に判定が一致している．本稿で用いた素性は一切の確率的な変動が存在しないため，容易に再現が可能であることがわかる．

6. おわりに

本稿では難易度推定と平易化判定共に，baseline より accuracy を 5.8 ポイント以上向上させ，かつ再現が容易な手法を提案した．また素性として，日本語 Wikipedia 中の文字単位の頻度，BCCWJ 中の単語単位の頻度が有用であることを示した．先行研究と比較すると，平易化判定は梶原らの手法 [梶原 17] より accuracy を 6.5 ポイント改善した．梶原らの構築した平易化用の辞書の改善など，本研究が今後活躍することを期待する．

参考文献

[Flesch 48] Flesch, R.: A new readability yardstick, *Journal of Applied Psychology*, Vol. 32, pp. 221–233 (1948)

[Ganitkevitch 13] Ganitkevitch, J., Durme, B. V., and Callison-Burch, C.: PPDB: The Paraphrase Database, in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 758–764 (2013)

[Maekawa 10] Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H., and Den, Y.: Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese, in *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pp. 1483–1486 (2010)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, pp. 3111–3119 (2013)

[Mizukami 14] Mizukami, M., Neubig, G., Sakti, S., Toda, T., and Nakamura, S.: Building a Free, General-Domain Paraphrase Database for Japanese, in *Proceedings of the 17th Oriental COCOSDA Conference*, pp. 129–133 (2014)

[Pavlick 15] Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., and Callison-Burch, C.: PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2, pp. 425–430 (2015)

[Pavlick 16] Pavlick, E. and Callison-Burch, C.: Simple PPDB: A paraphrase database for simplification, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2, pp. 143–148 (2016)

[梶原 17] 梶原 智之, 小町 守：Simple PPDB: Japanese, 言語処理学会第 23 回年次大会, pp. 529–532 (2017)

[佐藤 17] 佐藤 敏紀, 橋本 泰一, 奥村 学：単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討, 言語処理学会第 23 回年次大会, pp. 875–878 (2017)