Simple PPDB: Japanese

梶原 智之　小町 守
首都大学東京
kajiwara-tomoyuki@ed.tmu.ac.jp, komachi@tmu.ac.jp

1 はじめに

難解なテキストの意味を保持したまま平易に書き換えるテキスト平易化は、言語学習者や子どもをはじめとする多くの読者の文章読解を支援する。テキスト平易化の研究は、語彙的な手法 [1, 2, 3] や統語的な手法 [4, 5]、統計的機械翻訳を用いる手法 [6, 7] など多岐にわたるが、本研究では難解な語句を平易な同義表現に変換する語彙平易化に焦点を当てる。

これまでテキスト平易化は、平易に書かれた大規模コーパス（Simple English Wikipedia）、難解な文と平易な文のパラレルコーパス [7, 8, 9, 10]、難解な語句から平易な語句への言い換え辞書 [11] などの言語資源が豊富に存在する英語を中心に研究が進められてきた。しかし、日本語ではこのようなテキスト平易化のための言語資源が公開されていない。

そこで本研究では、日本語のテキスト平易化のために利用可能な平易な言い換え辞書 “Simple PPDB: Japanese” および大規模な単語難易度辞書を構築し、公開 [1]する。これは、日本語の言い換え辞書である PPDB: Japanese [12] [2]に含まれる言い換え対のうち、難解な単語から平易な単語への言い換え対のみを抽出し、日本語教育語彙表 [3]に由来する 3 段階の単語難易度（初級、中級、上級）および PPDB: Japanese の言い換え確率を付与したもの（表 1）である。

小平ら [13] によって構築された日本語の語彙平易化のための評価用データセットを用いた実験の結果、本研究で構築する Simple PPDB: Japanese はカバレッジが高いため、Accuracy で最高性能を達成した。

表 1: Simple PPDB: Japanese の事例

字引	（上級）→（初級）	辞書	0.878
晩餐	（上級）→（初級）	夕食	0.317
九大	（上級）→（中級）	九州大学	0.875
晩餐	（上級）→（中級）	ディナー	0.176
写真機	（中級）→（初級）	カメラ	0.757
ディナー	（中級）→（初級）	夕食	0.217

2 関連研究

日本語の言い換え辞書としては、基本的意味関係の事例ベース [4]の一部（略記対、異形同義語対、異表記対）、日本語 WordNet 同義語データベース [5]、内容語換言辞書 [6]（SNOW-D2） [14, 15]、日本語言い換えデータベース [2]（PPDB: Japanese） [12] などが構築されている。このうち、日英対訳コーパスから Bilingual Pivoting [16] と呼ばれる手法で構築された PPDB: Japanese は、1,500 万フレーズ対からなる日本語における最大の言い換え辞書である。本研究では、この大規模な言い換え辞書に含まれる言い換え対について、各単語に難易度を付与することにより、難解な単語から平易な単語への言い換え対のみを抽出したテキスト平易化のための言い換え辞書を構築する。

英語では、日本語に先立って Bilingual Pivoting を用いて大規模な言い換え辞書 [7]（PPDB） [17, 18] が構築されており、PPDB から平易な言い換え対のみを抽出したテキスト平易化のための言い換え辞書 [8]（Simple PPDB） [11] も構築されている。Simple PPDB では、1,400 万フレーズ対の PPDB を用いて 450 万フレーズ対の平易な言い換えを収集している。各フレーズ対には、多クラスのロジスティック回帰に基づく「言い換え先フレーズが平易な確率」や、PPDB の「フレーズの言い換え確率」が付与されている。

語彙平易化の研究は、辞書に基づく手法、パラレルコーパスに基づく手法、ノンパラレルコーパスに基づく手法の 3 つに大別できる。パラレルコーパスに基

[1] https://github.com/tmu-nlp/simple-jppdb/
[2] http://ahclab.naist.jp/resource/jppdb/
[3] http://jhlee.sakura.ne.jp/JEV.html
[4] https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html#A-9
[5] http://nlpwww.nict.go.jp/wn-ja/jpn/downloads.html#synonymsdatabase
[6] http://www.jnlp.org/SNOW/D2
[7] http://paraphrase.org/
[8] http://www.seas.upenn.edu/~epavlick/data.html

表 2: 難易度および難易度差推定の Accuracy

手法	単語の難易度	単語対の難易度差
Baseline（頻度 + 閾値）	0.557	0.497
基本素性	0.582	0.508
基本素性 + CBOW（50 次元）	0.681	0.568
基本素性 + CBOW（100 次元）	0.689	0.591
基本素性 + CBOW（200 次元）	0.695	0.601
基本素性 + SGNS（50 次元）	0.701	0.596
基本素性 + SGNS（100 次元）	0.708	**0.607**
基本素性 + SGNS（200 次元）	**0.709**	0.602

づく手法は、難解な文と平易な文からなるテキスト平易化のためのパラレルコーパスから単語出現頻度比を求める手法 [1] や単語アライメントによって平易な言い換え対を獲得する手法 [2] であり、日本語ではテキスト平易化のためのパラレルコーパスが公開されていないため適用できない。本研究では、国語辞典の見出し語と定義文から自動的に獲得された平易な言い換え辞書 [19] や人手で構築された複数の言い換え辞書 [20] を用いる日本語の語彙平易化手法およびノンパラレルコーパスに基づく手法 [3] と、本研究で構築する Simple PPDB: Japanese を用いる手法を比較する。

3 平易な言い換え辞書の構築

3.1 単語の難易度推定

SVM を用いて単語の難易度を推定する多クラス分類問題を解く。推定する単語の難易度は日本語教育語彙表 [3]の 3 段階の難易度（初級、中級、上級）である。SVM の素性には、単語長、文字種（ひらがな、カタカナ、漢字）、頻度、単語分散表現の 4 つを用いた。各単語の頻度は、Wikipedia [9]の本文を MeCab (0.996) [10] および mecab-ipadic-NEologd [11]によって分かち書きして求めた。単語分散表現は、同様に分かち書きした Wikipedia の本文を用いて word2vec [12]によって学習した。Wikipedia と日本語教育語彙表の両方に出現する 16,447 語に対して 10 分割交差検証によって単語難易度の Accuracy を計算した結果を表 2 に示す。ただし、SVM には scikit-learn (0.18.1) [13]の RBF カーネルを利用し、C と $gamma$ のパラメータはグリッドサーチによって最適な値を選択した。

[9] https://dumps.wikimedia.org/jawiki/20161001/
[10] http://taku910.github.io/mecab/
[11] https://github.com/neologd/mecab-ipadic-neologd
[12] https://code.google.com/archive/p/word2vec/
[13] http://scikit-learn.org/

表 3: 日本語の単語難易度辞書

辞書	収録語数	本研究との重複語数
JLPT [14]	7,759	7,416 (95.6%)
JEV [3]	17,207	16,447 (95.6%)
本研究	571,023	

ここで、Baseline（頻度 + 閾値）とは、単語の出現頻度に 2 つの閾値（閾値$_1$ > 閾値$_2$）を設定し、3 段階の単語難易度を推定するベースライン手法である。すなわち、ある単語の出現頻度が 閾値$_1$ 以上であれば初級、閾値$_1$ 未満かつ 閾値$_2$ 以上であれば中級、閾値$_2$ 未満であれば上級と、各単語の難易度を推定する。SemEval-2012 の English Lexical Simplification タスク [21] などで、単語出現頻度が単語難易度を推定するための有効な尺度であることが知られている。

基本素性は、word2vec の素性を除き、単語長、文字種、頻度の 3 種類の素性のみを用いた提案手法である。テキストの可読性を表すリーダビリティの先行研究では、単語長 [22] や文字種 [23] が有効な尺度であることが知られている。また、CBOW および SGNS は、それぞれ上記の 3 つの素性に加えて word2vec の continuous bag-of-words モデルまたは skip-gram with negative sampling モデルを用いる提案手法である。我々は「難解な単語は難解な文脈で使用されやすく、平易な単語は平易な文脈で使用されやすい」と考え、周辺の単語を考慮できるこれらのモデルを単語難易度の推定に利用する。

表 2 の実験結果から、単語難易度の推定には SGNS モデルを用いる提案手法が有効であることがわかる。そこで我々は、Wikipedia の本文に 5 回以上出現する 571,023 語について、100 次元の SGNS モデルを用いる提案手法で 3 段階の単語難易度を推定し、日本語の単語難易度辞書を構築した。これは、既存の日本語の単語難易度辞書と比較して非常に規模が大きいという特長を持つ（表 3）。

3.2 単語対の難易度差推定

PPDB: Japanese のうち、日本語教育語彙表に出現する単語のみからなる 40,309 単語対を用いて、 3.1 節と同様に単語の難易度を推定した。そして、各単語の難易度をもとに、「言い換え先単語が平易」「言い換え先単語が難解」「言い換え元と言い換え先の単語が同じ難易度」の 3 クラス分類を行ったときの Accuracy を表 2 に示す。

[14] http://www7a.biglobe.ne.jp/nifongo/data/

表 4: 日本語の語彙平易化タスクでの評価

System	Accuracy	Precision	Changed
Kajiwara-15a	0.060	0.114	0.522
Kajiwara-15b	0.127	**0.236**	0.539
Glavaš-15	0.135	0.181	0.746
本研究	**0.181**	0.210	**0.861**

表 2 から、やはり SGNS モデルを用いる提案手法が有効であることがわかる。英語の Simple PPDB [11] でも同様の 3 クラス分類が実施されており、本研究と同等の 0.604 の Accuracy が報告されている。そこで我々は、PPDB: Japanese のうち、Wikipedia の本文に 5 回以上出現する単語のみからなる 512,284 単語対について、100 次元の SGNS モデルを用いる提案手法で単語対の難易度差を推定した。そして、言い換え先が言い換え元よりも難解な単語対を除き、340,952 単語対の平易な言い換え対を抽出することで日本語の平易な言い換え辞書 “Simple PPDB: Japanese” を構築した。なお、言い換え先が言い換え元よりも平易な対は 133,274 単語対含まれている。それぞれの単語対には、「言い換え元単語の難易度」「言い換え先単語の難易度」「PPDB: Japanese の言い換え確率」の情報を付与した（表 1）。

4 語彙平易化タスクでの評価

小平ら [13] の日本語の語彙平易化のための評価用データセットを用いて、Simple PPDB: Japanese の語彙平易化タスクでの有用性を評価する。これは、現代日本語書き言葉均衡コーパス [15]（BCCWJ）から抽出された 2,010 文に 1 語ずつ難解語が含まれており、5 人のアノテータによって各難解語の平易な言い換えが平均 4.3 語ずつ付与されたデータセットである。表 4 に、日本語の語彙平易化タスクでの評価の結果を示す。

各手法を概説する。Kajiwara-15a は、国語辞典の見出し語と定義文から自動的に獲得された平易な言い換え辞書を用いる日本語の先行研究 [19] である。Kajiwara-15b は、人手で構築された複数の言い換え辞書を用いる日本語の先行研究 [20] である。Glavaš-15 は、単語分散表現のコサイン類似度によって類義語を集め、頻度や言語モデルなどによってリランキングする英語の先行研究 [3] である。本研究は、3.2 節で構築した Simple PPDB: Japanese を用いて平易な言い換えを集める提案手法である。平易な言い換え候補が複数存在する場合は、言語モデル確率によって最適な候補を選択する。言語モデルには、KenLM [24] を用いて Wikipedia [9]から 5-gram 言語モデルを構築した。

[15] http://pj.ninjal.ac.jp/corpus_center/bccwj/

表 5: 語彙平易化の例

Kajiwara-15a	Kajiwara-15b	Glavaš-15	本研究
こうして企業の【筆頭】{トップ, 先頭, 頂点} に立つ人間は、社内で最年長の人間ということになる。			
最初	先頭	中心	トップ
そしてこの調査は【疑わしい】{疑問がある, 怪しい}。			
—	変だと思う	興味深い	怪しい
なるほど、立場が上の人が、下の者にたいして、相手を尊重して【謙虚な】{おとなしい, 控えめな} 態度で接するのはよいことだ。			
—	—	誠実な	彼な

評価には、英語の語彙平易化タスク [25] と同様に、Accuracy、Precision、Changed Proportion の 3 つの尺度を用いた。Changed Proportion とは、システムが何らかの変換（正しい変換でなくても構わない）を行った割合を表す。

表 4 の実験結果から、本研究で構築した Simple PPDB: Japanese はカバレッジが高いため、Accuracy で最高性能を達成できたことがわかる。日英対訳コーパスから Bilingual Pivoting によって自動的に構築された PPDB: Japanese は大規模である反面、誤った言い換え対も含んでいる。そのため Precision では、人手で構築された言い換え辞書を用いる Kajiwara-15b には及ばなかった。本研究では既存の大規模な言い換え対から平易な言い換え対を抽出する手法を提案したが、日本語の言い換え対を大規模かつ高精度に収集することは、今後の課題である。

表 5 に、語彙平易化の例を示す。例えば 1 文目であれば、【筆頭】が難解語であり、この文脈での平易な言い換えは平易な順に {トップ, 先頭, 頂点} である。Simple PPDB: Japanese を用いると、「筆頭」に対して「トップ, 頭, 長」などの平易な言い換え候補を得ることができ、言語モデルを用いたリランキングによって「トップ」が選択される。2 文目の例に注目すると、Glavaš-15 は似た文脈で用いられる非同義語を出力している。これは単語分散表現のコサイン類似度を用いて候補を収集する手法の特徴であり、言い換え辞書を用いる他の手法ではこの誤りは発生しにくい。3 文目の例に注目すると、本研究では同義でも類義でもなく、出現文脈も似ていないと思われる出力を行っている。これは Bilingual Pivoting における単語アライメント誤りであると考えられる [16]。

[16] 言語モデルによる誤りではなく、候補が 1 つのみであった。

5 おわりに

本研究では、日本語のテキスト平易化のために利用可能な平易な言い換え辞書 “Simple PPDB: Japanese” および大規模な単語難易度辞書を構築し、公開[1]した。単語難易度辞書には、「単語、単語の難易度」の 2 項目について 57 万組を収録した。平易な言い換え辞書には、「難解な単語、平易な単語、難解な単語の難易度、平易な単語の難易度、言い換え確率」の 5 項目について 34 万組を収録した。上級の表現から初級の表現へ言い換えるなど、この言語資源を利用することで平易な言い換えを容易に実現できる。

内的評価では、言い換え対の難易度推定の Accuracy について、英語の先行研究と同等の性能を達成することができた。外的評価では、語彙平易化タスクの Accuracy および Changed Proportion について、最高性能を達成することができた。

今後は、句への拡張や、言い換え確率および難易度推定の精度を改善し、この言語資源を更新していく。

参考文献

[1] Or Biran, Samuel Brody, and Noemie Elhadad. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proc. of ACL 2011*, pp. 496 – 501, 2011.

[2] Colby Horn, Cathryn Manduca, and David Kauchak. Learning a Lexical Simplifier Using Wikipedia. In *Proc. of ACL 2014*, pp. 458 – 463, 2014.

[3] Goran Glavaš and Sanja Štajner. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proc. of ACL-IJCNLP 2015*, pp. 63 – 68, 2015.

[4] Dan Feblowitz and David Kauchak. Sentence Simplification as Tree Transduction. In *Proc. of PITR 2013*, pp. 1 – 10, 2013.

[5] Gustavo Paetzold and Lucia Specia. Text Simplification as Tree Transduction. In *Proc. of STIL 2013*, pp. 116 – 125, 2013.

[6] Sander Wubben, Antal van den Bosch, and Emiel Krahmer. Sentence Simplification by Monolingual Machine Translation. In *Proc. of ACL 2012*, pp. 1015 – 1024, 2012.

[7] Tomoyuki Kajiwara and Mamoru Komachi. Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings. In *Proc. of COLING 2016*, pp. 1147 – 1158, 2016.

[8] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proc. of COLING 2010*, pp. 1353 – 1361, 2010.

[9] William Coster and David Kauchak. Simple English Wikipedia: A New Text Simplification Task. In *Proc. of ACL 2011*, pp. 665 – 669, 2011.

[10] William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proc. of NAACL 2015*, pp. 211–217, 2015.

[11] Ellie Pavlick and Chris Callison-Burch. Simple PPDB: A Paraphrase Database for Simplification. In *Proc. of ACL 2016*, pp. 143 – 148, 2016.

[12] Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Building a Free, General-Domain Paraphrase Database for Japanese. In *Proc. of O-COCOSDA 2014*, pp. 129 – 133, 2014.

[13] Tomonori Kodaira, Tomoyuki Kajiwara, and Mamoru Komachi. Controlled and Balanced Dataset for Japanese Lexical Simplification. In *Proc. of ACL 2016 SRW*, pp. 1 – 7, 2016.

[14] 山本和英, 吉倉孝太郎. 用言等換言辞書を人手で作りました. 言語処理学会第 19 回年次大会発表論文集, pp. 276 – 279, 2013.

[15] 山形祐輝, 山本和英. 普通名詞換言辞書の構築. 言語処理学会第 20 回年次大会発表論文集, pp. 7 – 10, 2014.

[16] Colin Bannard and Chris Callison-Burch. Paraphrasing with Bilingual Parallel Corpora. In *Proc. of ACL 2005*, pp. 597 – 604, 2005.

[17] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The Paraphrase Database. In *Proc. of NAACL 2013*, pp. 758 – 764, 2013.

[18] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proc. of ACL-IJCNLP 2015*, pp. 425 – 430, 2015.

[19] 梶原智之, 山本和英. 語釈文を用いた小学生のための語彙平易化. 情報処理学会論文誌, Vol. 56, No. 3, pp. 983 – 992, 2015.

[20] Tomoyuki Kajiwara and Kazuhide Yamamoto. Evaluation Dataset and System for Japanese Lexical Simplification. In *Proc. of ACL-IJCNLP 2015 SRW*, pp. 35 – 40, 2015.

[21] Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. SemEval-2012 Task 1: English Lexical Simplification. In *Proc. of SemEval 2012*, pp. 347 – 355, 2012.

[22] Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, Vol. 32, pp. 221 – 233, 1948.

[23] 柴崎秀子, 玉岡賀津雄. 国語科教科書を基にした小・中学校の文章難易学年判定式の構築. 日本教育工学会論文誌, Vol. 33, No. 4, pp. 449 – 458, 2010.

[24] Kenneth Heafield. KenLM: Faster and Smaller Language Model Queries. In *Proc. of WMT 2011*, pp. 187–197, 2011.

[25] Gustavo Paetzold and Lucia Specia. Benchmarking Lexical Simplification Systems. In *Proc. of LREC 2016*, pp. 3074 – 3080, 2016.