均衡コーパスを用いた日本語語彙平易化データセットの構築

小平 知範　　梶原 智之　　小町 守
首都大学東京
{kodaira-tomonori,kajiwara-tomoyuki}@ed.tmu.ac.jp, komachi@tmu.ac.jp

1 はじめに

語彙平易化は，文中の難解語を検出し，それをより平易な同義語に言い換えるタスクである．語彙平易化により，第二言語学習者や子ども等の日本語にまだ不慣れな読者の文章読解を支援することができる [1]．

英語では，SemEval-2012 において英語の語彙平易化タスク [2] が開催され，語彙平易化の評価のためのデータセットが整備された結果，英語の語彙平易化タスクの研究が加速した．評価用データセットの構築や自動評価尺度の開発は，人手評価のコストや再現性の課題を解決するために重要である．

日本語では，語彙平易化の評価用データセットとして，Kajiwara ら [3] が SNOW E4[1] を構築している．しかし，Kajiwara ら [3] のデータセットには，以下の 3 つの欠点がある．

- 新聞コーパスのみから文を選んでいる．
- 言い換え時に起こる助詞の交替と，言い換えた後の文中に他の難解語が残ることが考慮されていない．
- 難易度ランキングで同順を許しておらず，各作業者の能力を考慮せずに統合している．

本稿では，Kajiwara ら [3] のデータセットの問題の解決に取り組み，新たなデータセットを構築した．主な貢献は，以下の 3 点である．

- 現代日本語書き言葉均衡コーパス [4] から文を抽出した．
- 日本語の用言の多様性をカバーし，難解語を 1 つしか含まない文で構成されている．
- 同順を許し，作業者の信頼度を考慮することで，難易度ランキングの一貫性が改善した．

本稿で構築したデータセットは，GitHub [2] にて公開している．

[1]http://www.jnlp.org/SNOW/E4
[2]http://www.github.com/KodairaTomonori/EvaluationDataset

2 関連研究

SemEval-2012 の英語の語彙平易化タスク [2] で構築された英語の語彙平易化の評価用データセットは，SemEval-2007 の英語の語彙的換言タスク [5] で作成された英語の語彙的換言の評価用データセットを元に作られている．Specia ら [2] は，大学生に言い換えを平易な順に並べ替えてもらっている．McCarthy ら [5] のデータセットに含まれる文は，均衡コーパス [6] から抽出されている．McCarthy ら [5] のデータセットは対象語が難解語でない場合があるため，Specia ら [2] のデータセットの文には難解語が含まれない場合がある．

De Belder ら [7] も同様に，McCarthy ら [5] のデータセットを元に，英語の語彙平易化の評価用データセットを構築した．Specia ら [2] とは異なり，対象語は難解語のみである．彼らは Amazon Mechanical Turk を利用し，各言い換えの難易度ランキングを獲得している．gold-standard ランキングは，作業者の信頼度を元に，各順位に重み付けをし，平均したスコアで作成している．

3 Kajiwara らの日本語の語彙平易化の評価用データセット

Kajiwara ら [3] は，Specia ら [2] の方法に従って日本語で語彙平易化の評価のためのデータセットを構築した．対象語は内容語（名詞，動詞，形容詞，副詞）である．De Belder [7] らに倣い，十分に平易な語や複合語の一部である語は対象語から取り除かれている．

Specia ら [2] と同様に，Kajiwara らも各対象語について 10 種類の文脈中での言い換えとその難易度ランキングを収録している．これらの文は，新聞記事から対象語を含む文を無作為に抽出して集められている．Kajiwara ら [3] はクラウドソーシングを利用して，言い換えと難易度ランキングの各タスクで 5 人ずつの作業者からデータを集めている．言い換えタスクでは，

文	「技を出し合い，気分が**高揚する**のがたまらない」とはいえ，技量で相手を上回りたい気持ちも強い．						
難易度	1. 盛り上がる	2. 高まる 高ぶる	3. 上がる	4. 高揚する	5. 興奮する	6. 熱を帯びる	7. 活性化する

図 1: Kajiwara ら [3] のデータセットの一部．難易度は左から平易な順に並んでいる．gold-standard ランキングは平均のスコアで並べられているため，同順が存在する．

文脈中で対象語と置換可能な語または句を収集している．難易度ランキングタスクでは，対象語とその言い換えを，平易な順に重複なく並び替えている．

言い換えタスクにおける作業者間の一致率は 66.4%，難易度ランキングタスクにおけるスピアマン順位相関係数は 0.332 であった．これは Specia ら [2] の難易度ランキングタスクでのスピアマン順位相関係数より，6.4 ポイント低い．このタスクでは，作業者間の評価の揺れが大きく，クラウドソーシングを使ってランキングを集めると質が低くなることがわかる．

図 1 に示す SNOW E4 のデータセットの一部を例にとり問題点を述べる．

データセットの分野が制限されている． Kajiwara ら [3] はニュース記事から文を抽出しているため，文体と語彙が偏ってしまう．英語の語彙平易化データセット [2, 7] では，英語の均衡コーパス [6] を使用しているため，このような問題はない．

用言の語彙的換言の多様性を制限している． 日本語では動詞の言い換えが助詞の交替を伴う場合があるが，Kajiwara ら [3] は対象語のみを置換するため，助詞の交替を伴う語彙的換言が得られない．図 1 において，対象語「高揚する」の前後の助詞「が」「の」を一緒に言い換えると「の高まり」などが得られるが，彼らのデータセットには含まれない．

対象語を平易化しても文中に難解語が残る． 図 1 の文は，対象語「高揚する」を平易にしても難解語「技量」が残っており，語彙平易化によって全体が平易な文を生成できない．対象語を平易にすることで，全体が平易な文が得られるような問題設定にすべきである．日英すべての先行研究 [2, 3, 7] の語彙平易化データセットがこの問題を持っている．

難易度ランキングに同順を許可していない． 各作業者は言い換えリストの難易度を考える際に，先行研究 [2, 3] では全ての単語に異なる難易度を付与しなければならない．例えば，図 1 の「高まる」と「高ぶる」の難易度が同程度であるとしても，各作業者は必ずこの 2 語に異なる難易度を付与しなければならない．これは言い換えリストに同程度の難易度の語が存在した場合にランキングの一貫性を損なう．De Belder ら [7] は難易度ランキングで同順を許し，Specia ら [2] より高い作業者間一致率を報告している．

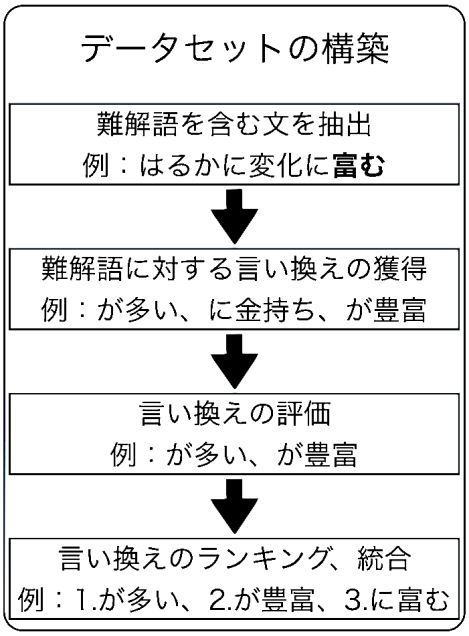

図 2: データセット構築の流れ

複数人のランキングの統合方法が単純である． 先行研究 [2, 3] では，ランキングの統合には平均を用いているが，平均では不真面目な作業者のランキングに統合ランキングが引っ張られる問題がある．例えば，作業者 4 人がある言い換えに対して順位を 1 とつけても，不真面目な 1 人の作業者が順位を 6 とした場合，作業者間の順位の平均は 2 となってしまう．

4 均衡コーパスを用いた日本語の語彙平易化データセットの構築

本研究では，日本語の語彙平易化の評価のためのデータセットを改良した．データセット構築の主な流れを図 2 に示す．最初に難解語を含む文をコーパスから抽出し，次に難解語に対する言い換えを獲得し，そこで得られた言い換えが正しいかを評価し，最後に言い換えを平易な順に並べ替えてもらい gold-standard ランキングを作成する．

難解語に対する言い換えの獲得，評価，難易度ランキングにおいてはクラウドソーシング（ランサーズ [3]）

[3] http://www.lancers.jp/

を用いた．各タスクにおいて，過去の作業承認率が95%以上である作業者に作業を依頼した．

4.1 難解語を含む文の抽出

本研究では日本語教育語彙表 [8][4] の上級の語を難解語と定義した．また，内容語を変換の対象とし，名詞，動詞，形容詞，形容動詞，副詞，サ変名詞，サ変動詞を扱った．Kajiwara ら [3] とは異なり，我々は動詞と形容詞の両方で，活用を扱った．

難解語を含む文は現代日本語書き言葉均衡コーパス (BCCWJ) [4] からランダムに抽出した．この際，難解語を 1 語だけ含む文を対象とし，形態素数は 7〜35 に制限した．なお，複合名詞や複合動詞は，その一部を置き換えると意味を保持できない場合が多いので，複合語の一部である語は除外した．先行研究 [3] で，語彙的換言とその難易度が文脈に依存して変化することが検証されているため，本研究でも 1 つの難解語に対して複数の文脈の中での言い換えと難易度順位を得た．先行研究 [2, 3] に倣い，難解語 1 語に対して 10 種類の文脈を扱った．これらの条件を満たすように品詞ごとに 30 種類の難解語を無作為抽出し，合計 2,100 文（30 語× 10 文× 7 品詞）の難解文を選定した．このうち，1,800 文はクラウドソーシング，300 文は研究室の学部生に 4.2-4.4 節の作業をしてもらった．

4.2 難解語に対する言い換えの獲得

前節で選定した難解文に含まれる各難解語について，平易化の候補となる語彙的換言をクラウドソーシングを用いて収集した． 各文の難解語に対して 5 人の作業者が，文の意味が変わらないような言い換えを語や句で列挙した．この時，難解語の前後の助詞を含めて言い換えた．

作業者間の一致率は，すべての作業者のペア ($p_1, p_2 \in P$) について一致率が計算される（式 1）．この式によって計算される作業者間の作業の一致率は，19.4 %であった．言い換えの品質を改善するために，ここで得られた言い換えを「言い換え候補」と考えて次節で不適切な言い換えを削除する．

$$\frac{1}{|P|}\sum_{p_1,p_2\in P}\frac{p_1 \cap p_2}{p_1 \cup p_2} \quad (1)$$

4.3 言い換えの評価

前節で得た言い換え候補から，クラウドソーシングを用いて不適切な言い換えを削除した． 各文の難解語に対する言い換え候補から，その難解語と置き換えても，文の意味が変わらない適切な言い換えを 5 人の作業者が選択した．

この時の作業者間の一致率は 66.9%であった．このスコアは，Kajiwara ら [3] のスコアと同程度である．

言い換えが正しいと過半数の作業者が判断したものを，正しい言い換えとした．すべての言い換えが不適切であると評価された 9 種類の難解語を含む 90 文を削除し，次節以降では残りの 2,010 文を使用した．

4.4 言い換えの難易度ランキング

前節で得たすべての言い換えおよび難解語を，クラウドソーシングを用いて所与の文脈中で平易な順に並び替えた． 各文の難解語と得られた正しい言い換えを 5 人の作業者が平易な順に並び替える．この時，同程度の難易度である語同士には，同順を許した．しかし，不真面目な作業者がすべてに同順を付与する可能性があるため最大 4 つまでの同順を許可した．

この作業において，スピアマン順位相関係数は，0.522 であった．このスコアは，Kajiwara ら [3] のスコアより 19.0 ポイント高い．

4.5 ランキングの統合

作業者のランキングを，De Belder ら [7] の外れ値を持つ作業者にペナルティを課す最尤推定 [9] を用いて統合した．この手法は，作業者の信頼度およびランキングの真の順序を推定する．以下の 2 つの方法で，信頼度を使用した．

1 つ目に，作業者が言い換えにつけた順位に信頼度を掛けたものの平均でランキングを統合する（**順位統合**）．だが，信頼度スコアは実数値のため，同程度の難易度を表すのが難しい．ゆえに，平均したスコアの差が 0.155 以下の場合，同順とみなした．この値は予備実験で，スピアマン順位相関係数が最大となるように調整した．

2 つ目に，外れ値を持つ作業者を除外するために，信頼度を使用する（**除外**）．具体的には，信頼度が平均より 0.05 以上下回る作業者を除外した．これにより，140 人中 9 人の作業者が除外された．その結果，131 人の作業者のランキングで gold-standard ランキングを構築した．

5 結果

5.1 データセットの特性

表 1 に各データセットの規模を示す．名詞にはサ変名詞，動詞にはサ変動詞，形容詞には形容動詞が含ま

[4] http://jhlee.sakura.ne.jp/JEV.html

表 1: データセットの比較

Dataset	均衡	言語	文数	名詞 (%)	動詞 (%)	形容詞 (%)	副詞 (%)	クラウドソーシング
De Belder ら [7]	yes	英語	430	100 (23.3)	60 (14.0)	160 (37.2)	110 (25.6)	yes
Specia ら [2]	yes	英語	2,010	580 (28.9)	520 (25.9)	560 (27.9)	350 (17.6)	no
Kajiwara ら [3]	no	日本語	2,330	630 (27.0)	720 (30.9)	500 (21.5)	480 (20.6)	yes
本研究	yes	日本語	2,010	570 (28.3)	570 (28.3)	580 (28.8)	290 (14.4)	yes

表 2: 均衡コーパスから抽出した文の詳細

分野	PB	PM	PN	LB	OW	OT	OP	OB	OC	OY	OV	OL	OM	all
文数	0	64	628	6	161	90	170	700	1	0	6	9	175	2010
平均獲得言い換え数	0	4.12	4.36	5.17	4.41	4.22	3.9	4.28	4	0	5.5	4.11	4.45	4.3

文	最も安上りにサーファーを**装う**方法は，ガラムというインドネシア産のタバコを，これ見よがしに吸うことです．						
難易度	1. のふりをする	2. に見せかける	3. の真似をする の振りをする	4. を真似る	5. に成りすます	6. を装う	7. を偽る

図 3: 本稿データセットの例．難易度は左から平易な順に並んでいる．

表 3: 統合ランキングの相関係数

	ベースライン	除外	除外 + 順位統合
平均	0.541	**0.580**	N/A
最尤推定	0.528	0.567	**0.581**

れる．本研究で構築したデータセットの規模は，先行研究 [2, 3] と同程度である．

5.2 獲得した言い換え

抽出した文の詳細を表 2 に示す．本稿で構築したデータセットでは，得られた言い換えの総数が 8,636 個，一文あたり平均 4.30 個獲得できた．

本稿で構築したデータセットの例を図 3 に示す．この中で助詞を含めた言い換えは，75 の文脈，全体の 3.7 %の文脈で得られた．図 3 を見ると，対象語のみを言い換えた場合には 2 種類の言い換えしか得られないが，助詞を含めて言い換えることによって新たに 5 種類の言い換えが獲得できている．

5.3 ランキングの相関

統合ランキングを評価するために，スピアマン順位相関係数を計算する．ベースラインは平均のスコア [3]，と最尤推定 [9] を用いた統合ランキングとする．提案手法は，Matsui ら [9] の手法を用いて計算した信頼度スコアを元に，外れ値を持つ作業者を除外した．

表 3 に統合ランキングの結果を示す．提案手法は，先行研究 [2, 3] より優れている．外れ値を持つ作業者を除外することで，特に相関の向上がみられた．

6 終わりに

本研究では，日本語の語彙平易化の評価用データセットの改良を行った．今後はこのデータセットを用いて語彙平易化の自動評価を行い，日本語学習者による主観評価との相関を調査する．

参考文献

[1] Jan De Belder and Marie-Francine Moens. Text simplification for children. In *SIGIR Workshop on Accessible Search Systems*, pp. 19–26, 2010.

[2] Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. Semeval-2012 task 1: English lexical simplification. In *SemEval*, pp. 347–355, 2012.

[3] Tomoyuki Kajiwara and Kazuhide Yamamoto. Evaluation dataset and system for Japanese lexical simplification. In *ACL-IJCNLP 2015 Student Research Workshop*, pp. 35–40, 2015.

[4] Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *LREC*, pp. 1483–1486, 2010.

[5] Diana McCarthy and Roberto Navigli. Semeval-2007 task 10: English lexical substitution task. In *SemEval*, pp. 48–53, 2007.

[6] Serge Sharoff. Open-source corpora: Using the net to fish for linguistic data. In *International Journal of Corpus Linguistics, 11(4)*, pp. 435–462, 2006.

[7] Jan De Belder and Marie-Francine Moens. A dataset for the evaluation of lexical simplification. In *CICLing*, pp. 426–437, 2012.

[8] Yuriko Sunakawa, Jae-ho Lee, and Mari Takahara. The construction of a database to support the compilation of Japanese learners dictionaries. In *Acta Linguistica Asiatica 2(2)*, pp. 97–115, 2012.

[9] Toshiko Matsui, Yukino Baba, Toshihiro Kamishima, and Hisashi Kashima. Crowdordering. In *PAKDD*, pp. 336–347, 2014.