日本語の語彙平易化評価セットの構築

梶原 智之　　山本 和英
長岡技術科学大学　電気系
{kajiwara, yamamoto}@jnlp.org

1 はじめに

語彙平易化は、文中の難解な語をより平易な同義語に置換する技術である。語彙平易化技術によって、外国人などの言語学習者や子どもをはじめとする幅広い読者の文章読解を支援することができる。

英語では、SemEval-2012 の評価型ワークショップにおいて English Lexical Simplification Task[1] が開催されており、語彙平易化システムの評価のための言語資源が整備され、様々な手法を用いた多くのシステムが参加している。また、Wikipedia[1] の平易版である Simple English Wikipedia[2] の存在により、難解な文と平易な文のパラレルコーパスを用いて統計的に平易化規則を学習するような手法も近年提案されている [2]。このような活発な研究の中で、いくつかの英語の語彙平易化システム [3][4] や難解な文と平易な文のパラレルコーパス [3][5][4][6]、語彙平易化アルゴリズムの評価のためのデータセット [1][7][5][8] が公開されている。

一方で日本語では、語彙平易化システムの評価のための言語資源も整備されておらず、難解な文と平易な文のパラレルコーパスも一般的に利用可能なものは存在しない。そのため、読解支援を必要とする読者のためにも、研究を加速させるためにも、日本語の語彙平易化のための言語資源やシステムの公開が必要である。

我々は先行研究 [6] で、まず日本語の語彙平易化システムを構築し、公開 [9] した。続いて本研究では、日本語の語彙平易化アルゴリズムの評価のためのデータセットを構築し、公開した。本稿で構築するデータセットは、次の URL から利用できる。

http://www.jnlp.org/SNOW/E4

[1] http://en.wikipedia.org
[2] http://simple.wikipedia.org
[3] http://homepages.inf.ed.ac.uk/kwoodsen/demos/simplify.html
[4] https://rewordify.com
[5] http://www.cs.pomona.edu/~dkauchak/simplification/
[6] https://www.ukp.tu-darmstadt.de/data/sentence-simplification/
[7] http://www.cs.york.ac.uk/semeval-2012/task1/
[8] http://people.cs.kuleuven.be/~jan.debelder/lseval.zip
[9] http://www.jnlp.org/SNOW/S3

2 関連研究

英語では、語彙平易化の評価のために 2 種類のデータセットが公開されている。これらはどちらも、SemEval-2007 の English Lexical Substitution Task[7] のために構築された語彙的換言の評価のためのデータセット [10] を改良して構築されている。本節では、これらの各データセットについて概説する。

2.1 McCarthy の語彙的換言データセット

SemEval-2007 の English Lexical Substitution Task[7] は、文脈中で対象語と置換可能な語や句を見つけるタスクである。対象語は内容語であり、その内訳は表 1 のとおりである。文脈は英語のウェブ均衡コーパスである English Internet Corpus[8] から選択されている。このタスクの語彙的換言の評価のためのデータセットは、201 種類の対象語に対して各 10 種類の文脈が付与された 2,010 文で構成されている。これらの各文について、5 人の英語母語話者が文脈中で適切な換言を 3 語まで付与している。作業者は、適切な換言を単語で思いつかない場合、句で回答してもよい。

以下に McCarthy の語彙的換言データセットの例を示す。この文脈中の形容詞 bright の換言として、ある 3 人の作業者は intelligent と回答し、またある 3 人の作業者は clever と回答し、ある 1 人の作業者は smart と回答している。

- <context>During the siegem George Robertson had appointed Shuja-ul-Mulk, who was a <head>bright</head> boy only 12 years old and the youngest surviving son of Aman-ul-Mulk, as the ruler of Chitral.</context>
- Gold: intelligent 3; clever 3; smart 1;

2.2 Specia の語彙平易化データセット

SemEval-2012 の English Lexical Simplification Task[1] は、文脈中で対象語の複数の換言を「易しさ」

[10] http://www.dianamccarthy.co.uk/task10index.html

表 1: 各データセットの品詞の内訳

データセット	総文数	名詞 (%)	動詞 (%)	形容詞 (%)	副詞 (%)
McCarthy の換言/Specia の平易化 (Trial)	300	80 (26.7)	80 (26.7)	90 (30.0)	50 (16.7)
McCarthy の換言/Specia の平易化 (Test)	1,710	500 (29.2)	440 (25.8)	470 (27.5)	300 (17.5)
De Belder の平易化	430	100 (23.3)	60 (14.0)	160 (37.2)	110 (25.6)
語彙平易化評価セット (SNOW E4)	2,330	630 (27.0)	720 (30.9)	500 (21.5)	480 (20.6)

の基準で並び替えるタスクである。「易しい」とは、子どもや英語非母語話者を含む幅広い人々にとって理解しやすいことを指す。このデータセットでは特に、英語が流暢な非母語話者（専攻の異なる大学 1 年生）によって「易しさ」による並び替え作業が行われている。Trial データについては 4 人、Test データについては 5 人の作業者が、2.1 節で紹介した McCarthy の語彙的換言データセットの対象語とその換言を「易しさ」の基準で並び替えている。

最終的に、各作業者による並び替えの結果を統合して一つのデータセットを作成する。Specia の語彙平易化データセットでは、各作業者の難易度の順位の平均値を用いて、各単語の難易度の順位を決定している。また、各作業者の順位の平均値が等しい単語同士は、同じ順位としている。

以下に Specia の語彙平易化データセットの例を示す。ある文脈中で、4 人の作業者から次のような難易度の順位が得られたとき、clear の順位はそれぞれ、1、2、1、4 であり、平均値は 2 である。同様に、light は 3.25、bright は 2.5、luminous は 4、well-lit は 3.25 が平均値となる。最終的な統合順位は、これらの単語を平均値の小さい順に並び替えることで得られる。

1: {clear} {light} {bright} {luminous} {well-lit}
2: {well-lit} {clear} {light} {bright} {luminous}
3: {clear} {bright} {light} {luminous} {well-lit}
4: {bright} {well-lit} {luminous} {clear} {light}
Gold: {clear} {bright} {light, well-lit} {luminous}

2.3 De Belder の語彙平易化データセット

De Belder ら [5] は、2.1 節で紹介した McCarthy の語彙的換言データセットを語彙平易化のためのデータセットに変換するために、まず十分に平易な対象語を除外している。Simple English Wikipedia の Basic English combined word list[11] などに含まれる十分平易な語を除外した結果、201 種類の対象語のうち 43 種類が残っている。これらの対象語とその換言を、それぞれ 5 人の作業者が「易しさ」の基準で並び替えている。この並び替えでは、難易度が等しいと思った単語同士に同じ順位を付与することが許されている。作業者は、Amazon Mechanical Turk[12] を用いて、アメリカに住んでおり過去の作業承認率が 95%以上であるという条件で集められている。

De Belder の語彙平易化データセットでは、雑音のある通信路モデルを用いて、各作業者の並び替え結果と各作業者の信頼度を考慮して、各作業者による並び替えの結果を統合している。

3 日本語版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構築する。我々は英語での先行研究を参考に、まず語彙的換言の評価のためのデータセットを構築し、その対象語と複数の換言を「易しさ」の基準で並び替えることにより、語彙平易化の評価のためのデータセットを構築する。

我々はクラウドソーシングを用いて延べ 500 人の作業者によって、先行研究 [1][7] と同等の規模の日本語のデータセットを構築した。

3.1 語彙的換言データセットの構築

3.1.1 換言対象語の選定

我々は換言の対象語を、IPADIC(2.7.0)[13] と JUMAN(7.0)[14] の形態素辞書の共通部分から選択する。本研究で対象とするのは内容語（名詞、動詞、形容詞、副詞）であり、サ変名詞は名詞、サ変動詞は動詞、形容動詞は形容詞にそれぞれ含まれる。なお、複合名詞や複合動詞は、その一部の語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対象語から除外する。

我々の目的は語彙平易化のためのデータセットを構築することなので、すでに十分に平易な語は扱わない。本研究では、小学生のための理解語彙である学習基本語彙 [9] に含まれる語を対象語から除外する。

[11] http://simple.wikipedia.org/wiki/Wikipedia:Basic_English_combined_wordlist

[12] https://www.mturk.com

[13] http://sourceforge.jp/projects/ipadic/releases/24435/

[14] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

全ての語に必ずしも語彙的換言が存在するわけではないので、換言が存在しない可能性の高い語は扱わない。本研究では、JUMAN 辞書の内容語に人手で語彙的換言を付与している内容語換言辞書 (SNOW D2)[10][11] に含まれない語を対象語から除外する。

先行研究 [7] に倣い、各対象語を 10 種類の文脈中で扱う。本研究では文脈として、1990 年から 2004 年までの 15 年分の日本経済新聞 [15] の本文から、対象語が含まれる文を無作為に抽出する。なお、文脈が 10 種類に満たない低頻度語は、対象語から除外する。

このようにして、名詞および動詞各 75 語、形容詞および副詞各 50 語を対象語として無作為に選定した。

3.1.2 語彙的換言の列挙

3.1.1 節で選定した各対象語について、文脈中で置換可能な語を 5 人ずつの作業者によって列挙する。本研究ではクラウドソーシングを用いて作業者を募った。それぞれの作業者は、50 語（対象語 5 ×文脈 10）ずつの作業を行う。作業者は、文脈中で対象語と置換可能な単語または句を、思いつく限り列挙する。その際、辞書を参照することは許可しているが、他人に意見を求めることは禁止している。作業者が適切な換言が存在しないと判断した場合には、無記入を許している。

本研究ではクラウドソーシングサービスとしてランサーズ [16] を利用し、37 人の作業者を得た。これらの作業者から、それぞれの文脈中の対象語について、平均 5.38 語の語彙的換言が得られた。なお、作業者間の一致率を先行研究 [7] と同様に計算したところ、17.8%の一致率であった。

3.1.3 換言データセットの統合

3.1.2 節で 5 人ずつの作業者から得た換言を統合して、ひとつの日本語の語彙的換言の評価のためのデータセットを構築する。ある単語または句が文脈中で対象語の換言であるかどうかは作業者間で基準が一定ではないため、新たに 5 人ずつの作業者を募り、適切な換言であるかどうかの確認を行う。なお、3.1.2 節と同様に、作業者はクラウドソーシングで募り、各作業者は 50 語の作業を行う。本研究では、5 人中 3 人以上の作業者が「適切な換言である」と認めた単語または句を採用する。なお、「適切な換言ではない」基準として、次の 2 つを作業者に提示している。

- 文脈中の対象語と置き換えたときに、不自然な文になる場合は、適切な換言ではない。
- 文脈中の対象語と置き換えたときに、文の意味が保持できない場合は、適切な換言ではない。

本研究ではクラウドソーシングサービスとしてランサーズを利用し、83 人の作業者を得た。これらの作業者によって、それぞれの文脈中の対象語について、平均 4.50 語の語彙的換言が適切な換言であると認められた。ただし、全ての換言が不適切であると評価された 17 種類の対象語を含む 170 文は除外した。なお、作業者間の一致率は、66.4%であった。

以下に、構築した日本語の語彙的換言データセットの例を示す。本データセットは、対象語と文脈から成る文脈データと、対象語とその換言から成る換言データの 2 種類で構成されている。下の例の先頭の数値は、2 つのデータを紐付ける ID である。下の例では、この文脈中の名詞「悪気」の換言として、ある 1 人の作業者は「意地悪」と回答し、またある 1 人の作業者は「悪い考え」と回答し、ある 4 人の作業者は「悪意」と回答している。それぞれの換言に続く数値は、本節で「適切な換言である」と認めた作業者の人数ではなく、3.1.2 節でその語または句を回答した作業者の人数であることに注意されたい。

- 820, 悪気, 親は悪気で言ったわけではなく、子供をあやすということを本当に知らなかった様子。
- 820, 悪気, 意地悪 1; 悪い考え 1; 悪意 4;

3.2 語彙平易化データセットの構築

3.2.1 複数の換言の並び替え

3.1 節で構築した語彙的換言のためのデータセットに含まれる対象語とその換言を、文脈中で「易しさ」の基準で並び替えることによって、語彙平易化のためのデータセットを構築する。本研究では、それぞれ 5 人の作業者によって並び替えの作業を行う。なお、3.1.2 節と同様に、作業者はクラウドソーシングで募り、各作業者は 50 語の作業を行う。また、先行研究 [1] と同様、各作業者による並び替えの際には、全ての語または句に重複なく難易度の順位を割り当てる。ただし、それらの結果を統合する際には、複数の語または句に同じ難易度の順位が割り当てられる場合がある。

この並び替え作業は、3.1.3 節で募った 83 人の作業者が行った。なお、作業者間の一致率を先行研究 [5] に倣ってスピアマンの順位相関係数で計算したところ、33.2%の一致率であった。

3.2.2 平易化データセットの統合

3.2.1 節で得られた 5 人ずつの作業者による並び替えの結果を統合して、ひとつの日本語の語彙平易化の

[15] http://www.nikkeibookvideo.com/kijidb/
[16] http://www.lancers.jp

評価のためのデータセットを構築する。本研究では先行研究 [1] に倣い、各作業者の難易度の順位の平均値を用いて、各語または句の難易度の順位を決定する。ここで、各作業者の順位の平均値が等しい語または句同士は、同じ順位を割り当てる。本研究では、同じ作業者が換言の評価と並び替えの両方の作業を行っているため、ある作業者がある単語または句を「適切な換言ではない」と評価した場合、その単語または句は難易度の順位を付与されない。これらの順位を付与されなかった単語または句は、最低順位（最も大きな値）として平均値を計算する。

統合して得られたデータセットには、それぞれの文脈中の対象語とその換言について、平均 4.94 段階の難易度の順位が与えられた。

以下に、構築した日本語の語彙平易化データセットの例を示す。ある文脈中で、5 人の作業者から次のような難易度の順位が得られたとき、「意地悪」の順位はそれぞれ、1、2、4、2、2 であり、平均値は 2.2 である。同様に、「悪意」は 2.2、「悪気」は 2.6、「悪い考え」は 3 が平均値となる。最終的な統合順位は、これらの単語を平均値の小さい順に並び替えて得られる。

1: { 意地悪 } { 悪意 } { 悪気 } { 悪い考え }
2: { 悪気 } { 意地悪 } { 悪意 } { 悪い考え }
3: { 悪意 } { 悪い考え } { 悪気 }　×：意地悪
4: { 悪い考え } { 意地悪 } { 悪気 } { 悪意 }
5: { 悪意 } { 意地悪 } { 悪気 }　×：悪い考え
Gold: { 意地悪, 悪意 } { 悪気 } { 悪い考え }

3.2.3　文脈依存性

本節では、構築した日本語の語彙平易化データセットの文脈依存性を調査する。それぞれの対象語について、10 種類ずつの文脈で換言を付与し、対象語とその換言を並び替えて語彙平易化のデータセットを構築した。ここには前提として、「文脈によって語彙的換言が異なる」「文脈によって平易な語または句が異なる」という 2 つの文脈依存性が仮定されている。この仮説を検証するため、表 2 に同じ換言リストから異なる難易度順位が得られた組み合わせの割合などを示す。この表から、語彙的換言では 84.8%、語彙平易化では 59.5%が文脈に依存して変化することがわかる。

表 2: 語彙平易化データセットの文脈依存性

	英語 [1]	日本語
対象語が同じ文脈の組	9,045	10,485
換言リストが等しい組	302 (3.3%)	1,593 (15.2%)
難易度順位が異なる組	163 (54.0%)	948 (59.5%)
最も平易な語が違う組	57 (35.0%)	463 (48.8%)

4　おわりに

本研究では、日本語の語彙的換言および語彙平易化の評価のためのデータセットを構築し、公開した。これらのデータセットを用いることで、語彙的換言および語彙平易化アルゴリズムの精度や再現率を計算し、F 値を求めることができるようになる。このような自動評価の枠組みを設けることで、日本語の語彙的換言および語彙平易化の研究のサイクルが素早く回り、より深い議論を重ねていくことが可能となる。

最後に、本データセットを用いて日本語の語彙的換言および語彙平易化が活発に研究され、子どもや言語学習者をはじめとする幅広い読者の文章読解の助けとなるシステムが開発されることを期待する。

参考文献

[1] Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. Semeval-2012 task 1: English lexical simplification. *In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval-2012)*, pp. 347–355, 2012.

[2] Colby Horn, Cathryn Manduca, and David Kauchak. Learning a lexical simplifier using wikipedia. *In Proceedings of the 52th Annual Meeting of the Associatioin for Computational Linguistics (ACL-2014, Short Papers)*, pp. 458–463, 2014.

[3] David Kauchak. Improving text simplification language modeling using unsimplified text data. *In Proceedings of the 51th Annual Meeting of the Associatioin for Computational Linguistics (ACL-2013)*, pp. 1537–1546, 2013.

[4] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. *In Proceedings of the 23rd International Conference on Computational Linguistics (Coling-2010)*, pp. 1353–1361, 2010.

[5] Jan De Belder and Marie-Francine Moens. A dataset for the evaluation of lexical simplification. *In Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2012)*, pp. 426–437, 2012.

[6] 梶原智之, 山本和英. 日本語の語彙平易化システムの構築. 情報処理学会第 77 回全国大会, 2Q-01, 2015.

[7] Diana McCarthy and Roberto Navigli. Semeval-2007 task10: English lexical substitution task. *In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 48–53, 2007.

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

[9] 甲斐睦朗, 松川利広. 語彙指導の方法：語彙表編. 光村図書出版株式会社, 2002.

[10] 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.

[11] 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.