2Q-01

日本語の語彙平易化システムの構築

梶原智之　　山本和英
長岡技術科学大学電気系

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 の平易版である Simple English Wikipedia 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]が Web で公開されている。

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。また、日本語の語彙平易化手法はこれまでにもいくつか提案されている[5][6]が、一般的に利用可能なシステムは存在しない。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。

本研究の貢献は、日本語の語彙平易化システムを構築し、Web で初めて公開することである。本稿で構築するシステムは、次の URL から利用できる。
http://www.jnlp.org/resources/3

2　提案手法

本稿で構築する日本語の語彙平易化システムは、典型的な語彙平易化手法の 4 つの機構[7]を備えている。すなわち、まず入力文から難解語を検出する。続いて、その難解語の語彙的換言を列挙する。そして、難解語の換言の中から入力文の文脈で使用可能な語を選択する。最後に、それらの語を難易度で並び替え、最も平易な語を難解語と置き換えることで出力文を生成する。この語彙平易化システムの概要を、例とともに図 1 に示す。

Building a Japanese Lexical Simplification System
Tomoyuki KAJIWARA (kajiwara@jnlp.org),
Kazuhide YAMAMOTO (yamamoto@jnlp.org)
Department of Electrical Engineering, Nagaoka University of Technology, 1603-1 Kamitomioka-machi, Nagaoka City, Niigata Prefecture, 940-2188 Japan

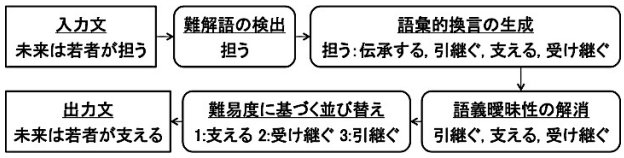

図 1. 語彙平易化システムの概要

2.1 難解語の検出

本システムでは、名詞、動詞、形容詞、副詞などの内容語と呼ばれる単語を平易化の対象とし、機能語、複合語、慣用句などは扱わない。まず、形態素解析によって入力文から内容語を抽出する。ただし、複合語や慣用句などは、そこに含まれる単語の単位で変換を行うと意味を保持できなくなる場合が多いので、複合表現リストを用意し、これらを平易化の対象から除外する。また、すでに十分平易な単語も平易化する必要がないので、十分平易な単語リストも用意し、これらも平易化の対象から除外する。これらの処理を経て残った単語が平易化を試みる難解語である。

なお、形態素解析には MeCab (0.993)[1]および IPADIC (2.7.0)[2]を使用した。また、複合表現リストには、複合名詞として日本語 Wikipedia の見出し語[3]、複合動詞として複合動詞レキシコンの見出し語[4]、慣用句として佐藤らの慣用句リスト[8]を使用した。十分平易な単語リストとしては、小学生のための理解語彙である学習基本語彙[9]を使用した。

2.2 語彙的換言の生成

続いて、前節で検出した各難解語について、語彙的換言を列挙する。これらの語彙的換言が、平易語の候補となる。

本システムでは、先行研究[10]の調査に基づき、基本的意味関係の事例ベース[5]、内容語換言辞書[6]、動詞含意関係データベース[7]、日本語 WordNet 同義語データベース[8]から語彙的換言を収集した。

[1] https://code.google.com/p/mecab/
[2] http://sourceforge.jp/projects/ipadic/releases/24435/
[3] http://dumps.wikimedia.org/jawiki/
[4] http://vvlexicon.ninjal.ac.jp/
[5] https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html#A-9
[6] http://www.jnlp.org/resources/2
[7] https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html#A-2
[8] http://nlpwww.nict.go.jp/wn-ja/jpn/downloads.html

2.3 語義曖昧性の解消

前節で難解語の語彙的換言を得たが、先行研究[10]でも示したように、換言可能と思われる単語対でも、周辺の文脈によって換言可能な場合と換言不可能な場合が存在する。そこで、入力文の文脈を考慮して不適格な換言を除去する手法を提案する。

文の適格性を評価するために、本システムでは文の基本的な意味構造である述語項構造を用いる。述語とその項の関係を見ることで、文の適格性を評価することができる。例えば、図1の入力文を述語項構造解析することで、述語が「担う」であり、そのガ格の項が「若者」であることが分かる。また、格フレーム辞書には述語の取り得る項の情報が記載されているので、述語「担う」がガ格に「若者」を取ることを格フレーム辞書で調べることにより、この述語項構造が適格であることが分かる。このように、入力文から述語項構造解析によって(項，格，述語)の三つ組を抽出し、項または述語を換言した三つ組が格フレーム辞書に存在しなければ、それを不適格な換言として除去する。

図1の例では、難解語として述語である「担う」を検出している。また「担う」の換言として「伝承する」や「支える」などを得ている。述語項構造解析により、「担う」のガ格の項が「若者」であることが分かっているが、「担う」の換言である「伝承する」の格フレームにはガ格に「若者」が登録されていないため、「伝承する」を除外する。一方、「担う」の別の換言である「支える」の格フレームにはガ格に「若者」が登録されているため、「支える」は平易化候補として残す。

本システムでは、述語項構造解析に SynCha (0.3.0)[9]を、格フレーム辞書に京都大学格フレーム(1.0)[10]を使用した。

2.4 難易度に基づく並び替え

前節で、難解語の語彙的換言の中から述語項構造が適格な語が得られた。最後に、これらの単語の中から最も平易な語を入力文中の難解語と置換して出力文を生成する。

本システムでは、単語親密度データベース[11]を用いて各単語に難易度を付与した。親密度の最も高い語が、最も平易な語である。

3 おわりに

本稿では日本語の語彙平易化システムについて述べた。本システムは、難解語の検出、語彙的換言の生成、語義曖昧性の解消、難易度に基づく並び替えという語彙平易化の典型的な4つの機構を持つ。

特に、語彙的換言の生成においては、先行研究[10]の調査に基づき、利用可能な日本語の語彙的換言知識を広く収集した。本システムは、精度の低い句単位の換言知識の多くを除いているため、今後は高精度に句単位の換言を収集することが課題である。

関連して、格フレーム辞書や単語親密度データベースなどが単語単位の言語資源であるため、2.3 節では句の検索時に末尾の単語をその句の主辞と見て主辞で検索を行っている。また 2.4 節では句の難易度として各単語の難易度の平均値を採用している。これらの基本的な言語資源についても、句単位で整備することにより、語彙平易化をはじめとする各種自然言語処理タスクの性能改善が期待される。

また、2.3 節における述語項構造解析と格フレーム辞書を用いた換言の適格性の評価は、本稿における我々の提案手法である。この換言の適格性については、今後評価を行って確認したい。

最後に、本システムの公開が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り、同時に日本語の語彙平易化の研究が本システムをベースラインとして活発に研究されることを期待する。

[9] http://www.cl.cs.titech.ac.jp/~ryu-i/syncha/
[10] http://www.gsk.or.jp/catalog/gsk2008-b/

参考文献

[1] Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. SemEval-2012 Task 1: English Lexical Simplification. In Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval-2012), pp. 347–355, 2012.

[2] Colby Horn, Cathryn Manduca and David Kauchak. Learning 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] Automatic sentence simplification using Wikipedia. http://homepages.inf.ed.ac.uk/kwoodsen/demos/simplify.html/

[4] Rewordify.com. https://rewordify.com/

[5] 鍛治伸裕, 河原大輔, 黒橋禎夫, 佐藤理史. 格フレームの対応付けに基づく用言の言い換え. 自然言語処理, Vol. 10, No. 4, pp. 65–81, 2003.

[6] 美野秀弥, 田中英輝. 国語辞典を使った放送ニュースの名詞の平易化. 言語処理学会第 16 回年次大会, pp. 760–763, 2010.

[7] Matthew Shardlow. A Survey of Automated Text Simplification. International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing, pp. 58-70, 2014.

[8] 佐藤理史. 基本慣用句五種対照表の作成. 情報処理学会研究報告, 2007-NL-178, pp. 1–6, 2007.

[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] 梶原智之, 山本和英. 日本語の語彙的換言知識の質的評価. 信学技報, vol. 114, no. 366, NLC2014-37, pp. 43-48, 2014.

[11] 天野成昭, 近藤公久. NTT データベースシリーズ日本語の語彙特性 (第 1 期 CD-ROM 版). 三省堂, 1999.

付録：システムの入出力の例

- 海外からの｛訪問者→お客さん｝に｛配る→渡す｝ほか、神戸市の海外事務所に送付する。
- 研究開発学校として指定されれば、学習指導要領の｛枠→フレーム｝にとらわれない教育も可能。
- 北陸銀の高木｛頭取→社長｝も「リストラ策と収益拡大策の両面」で統合効果を早期に極大化する考えを示した。