研究速報

複数の言語資源を用いた難解語の平易化モデルの構築

Construction of Simplification Model for Complex Words Using Multiple Language Resources

呉　浩東*1
Kotoh Go

Email: kgo@dokkyo.ac.jp

レベルの違い日本語読者の言語理解能力を向上させ、より効果的情報を収集するために、テキストにおける難解な語を平易な表現へ変換することが必要である。語彙平易化技術は、子どもや言語学習者をはじめとするさまざまな読者の文章読解を支援する道具である。また、自然言語処理における言語理解や機械翻訳などのタスクの精度を向上にも役立つことが期待される。本研究では、複数の言語資源を用いて、難解語の平易化を目指す。提案手法では日本語読者にとってより正確に情報を伝達するため、難解語の変換過程における平易さ、意味の保持と文脈の整合性を重視する。段階的な成果として、難解語の変換精度は 76. 9%、変換率は 78. 5%、意味維持度の高い結果を得た。

In order to improve the different levels of language comprehension ability of readers for understanding Japanese texts, and enhance readers to access various information more effectively, it is necessary to convert complex words in text into simple expressions. The technology for simplify complex words provide a tool to improve reading comprehension ability of children, language learners and various other readers. It can also be expected to help improvement of the accuracy of tasks such as language understanding and machine translation in natural language processing. In this research, we aim to simplify complex words using multiple language resources and develop an effective language model　based on this method. In order to convey information more accurately to Japanese readers, we take consider of simplicity, grammaticality, meaning preservation in the conversion process of complex words in the context. As a result that is in improvement, the conversion accuracy of complex words was 76.9%, the converse rate was 78.5%, and in same time, the original meaning in context is well retained.

*1 獨協大学国際教養学部言語文化学科

1. はじめに

国際化と情報化が急速に進み、誰にも大量かつ多様なテキストデータにアクセス環境を利用できる。より効果的情報を理解するために、レベルの違い読者の言語能力の格差を埋める技術が必要である。その技術の中核に語彙平易化がある。言語学習者の読解を妨げる要因にはさまざまなものが考えられるが、そのなかの一つは難解語の存在である。難解語かどうかを判定する技術は難解語同定（Complex Word Identification, CWI）という。

語彙平易化は、テキスト中の難解な語（難解語という）をより平易な語あるいは表現に置換する作業である。語彙平易化技術は、子どもや言語学習者をはじめとするさまざまな読者の文章読解を支援する道具である。また、自然言語処理における言語理解や機械翻訳などのタスクの精度を向上するためにも役立つことが期待される。

本研究では、日本語における難解語の対象は、名詞、動詞、形容詞、副詞などの内容語である。読者のレベルは小学 6 年生までの者に想定する。一方、機能語を除外する。平易語の対象は基本語彙とされ、高頻度語でありながら使用者数の多さは特徴である。

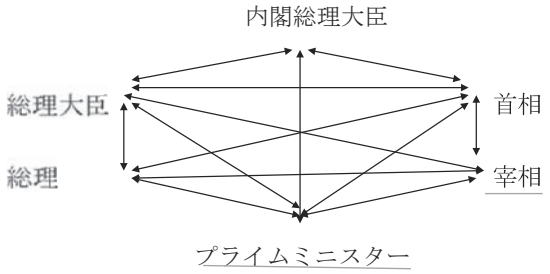

図 1　難解語の例

図 1 に下線を付いた単語は難解語である。

2. 難解語の言い換え

難解語に対する読解支援ツールの一つは言い換え処理が挙げられる。すなわち、読者が理解できない語句を別のより平易な語句に言い換え処理である。言い換え処理の入力は難解語を含む対象文である。一方、出力文は言い換え処理を施した平易語または平易表現から構成される目的文である。入出文の具体例を以下の通り示す。

例 1

入力：　人工知能という概念は 1956 年に**作り出した**。

出力：　人工知能という概念は 1956 年に**作った**。

言い換え処理に三要素がある：（１）**平易さ** (simplicity) 、(2)**文法性** (grammaticality)、（３）**意味の保持** (meaning preservation)　の三つの視点から評価することが多い。「平易さ」では、言い換え前に比べて言い換えにどれくらい平易になっているかを評価する。「文法性」では、変換後の文が文法的に的確かどうかを表す。「意味の保持」（すなわち「整合性」）は言い換え後の文が言い換え語前と意味を保持しているかどうかを人間により判定する。さらに、難解語の言い換えにおける性能評価に、言い換え率（R）と言い換え精度（P）は使う。

$$R=\frac{\text{システムと人間で共通する言い換えの数}}{\text{人間が言い換えた語の数}}$$

$$P=\frac{\text{システムと人間で共通する言い換えの数}}{\text{システムが言い換えた語の数}}$$

奥村，永田（2017）指摘の通り、言い換え処理についてでも学習効果を考えることが大切である。言い換え処理で特に問題となるのは、難解語のうちどの表現に言い換えるかということである。

3. 関連研究

文中の難解語を他の平易表現に変換に関する研究は数多く行っている。近年、平易な表現への研究は統計的機械翻訳の手法が特に盛んである。特に英語においては、Wikipedia をコンパラブルコーパスとし、これから抽出された単言語パラレスコーパス（Specia 2010;Zhu et al.2010; Coster and Kauchak 2011）を平易化ツールとして統計的平易語生成の研究を盛んに行われた。また、Web 検索を用いる複数パターンの平易化変換の生成（熊本, 田中 2008）、利用者の言語の能力に配慮した平易化（西村,田中,北野,大林 2009 ; 中野,遠籐,菅，乾、

藤田 2006）の手法も報告されている。国外は語の変換に着目した評価型ワークショップが開催し（McCarthy et al. 2007, Jauhar and Specia 2012, Sinha 2012）といた研究が報告されている。

4. 提案手法

4.1 変換方式

われわれは二種類の変換方式を提案する：（１）１語対１語変換、（２）1 語対 N 語変換。具体的に、難解語（T）に平易な同義語が存在する場合、使用頻度最も高いものによって置き換えする。一方、同義語か平易な同義語（S）か類語（C）が存在しない場合。国語辞書の語釈文に切り替える。もし、語釈文に S があれば、S は T の平易な類語（平易語）に置き換える。

4.2 難解語の同定

単なるコーパスから低頻度語を難解語として抽出するアプローチは問題が存在する。頻度以外に、語の親密度が重要である。難解語の判定基準は日本語学習者であるため、われわれはまずテキストの中に難解語をありうる低頻度語と小学校 6 年生までの教科書に収録していない専門用語を難解語に指定し、候補リストに入れる。

難解語と置き換える平易な表現はシソーラスと国語辞書から得る。1 語対 1 語において国語辞書とシソーラスの組み合わせが有効である。一方、１語対 N 語の変換は国語辞書を用いて難解語の意味を特定するのは一番適切であると考えられる。

4.3 難解語の解消法

われわれは前述の考え方に基づいて、難解語の平易化過程を以下のモデルにする。

難解語の平易化モデル：

1. 入力されたテキストに対して、コンピュータによる「形態素解析」を行う。内容語（名詞、動詞、形容詞、副詞）を入力テキストから抽出する。
2. 入力文から基本語彙以外と低使用頻度の単語を難解語の候補語として検出する。
3. テキスト内の難解語を読者に判断してもらい、非難解語を候補リストから削除する。
4. リストに残る候補語はシソーラスを調べ、同義語をリストに入れる。同義語は存在しない場合、シソーラスに類似度の高い類語をリストに入れる。
5. 均衡コーパスでリストの候補の使用頻度の低いものを取り除く。
6. 出現頻度最大の候補語は難解語の出現頻度より著しい高い場合、1 語対 1 語の置換を行う、選定作業を終了する。さもなければ、STEP 7 にシフトする。
7. すべての候補の出現頻度は難解語より少ないか同程度（±20%）の場合、国語辞書の語釈文を使い 1 語対 N 語の変換を行い、難解語（S）を平易な同義語なし類語を置き換え、その過程で不要語（非 S）も削除する。

ここで提案した手法は他手法と違うものである。

例 2：難解語の 1 対 N 変換のケース

難解語**断言**に対して国語辞書とシソーラスから共通の同義語は：**言い切る**、**確言**、**明言**である。しかし、現代日本語書き言葉均衡コーパス（BCCWJ）を用いて頻度情報を調べると、それぞれの頻度は：断言（784）、言い切る（163）、確言（30）、明言（375）である。しかし、BCCWJ のような大規模コーパスから出現頻度からみると、閾値（1,200）より以下であり、断言とその三つの候補語とも平易語ではないことが判明した。そのため、国語辞書の語釈文に切り替える。**断言**は“（スル） 確信をもってきっぱりと言い切ること”と定義される。そのなかの「言い切る」を**断言**の類語の「主張する」に置き換える。その根拠は**主張**が BCCWJ コーパスに出現頻度は 8,267 であり、高頻度語である。

例 3. 難解語の 1 対 1 変換のケース

図 1 に示されるように、難解語であるプライムミニスターと宰相に対して、内閣総理大臣、総理大臣、首相と総理が平易語候補リストに残る。さ

らに、BCCWJ コーパスを調べて、総理の出現頻度（4,963）が一番高いため、総理は平易語として選出される。

5. 実験

われわれは、日本語 Wikipedia からランダムに難解語一つだけを含む 378 文を抽出し、その中に異なる 260 の難解語を含む 260 文を選択する（これは表 1 の原文に相当）。「人手による言い換え」は上記の対応する文に対する評価である。評価者は日本語母語話者 5 人による 5 段階評価（5 が最もよい）の平均を評価値としている。「言い換え率」は言い換えられた文の割合である。

表 5.1 難解語の言い換えにおける性能評価

種類	平易さ	文法性	意味の保持	言い換え率
原文	3.15	4.93	-	-
人手による変換	4.23	4.85	4.69	84.2%
本手法	3.82	4.50	4.54	78.5%
Bira らの手法	3.23	4.03	4.23	72.3%
Gastavo らの手法	3.54	3.96	4.31	68.3%

実験結果は表 5.1 に示す。実験の対象者は小学性 3 年生 5 人、4 年生 5 名、5 年生 5 名、日本語中級レベルの留学生 5 人計 20 名である、「人手による評価」は 5 人の日本語母語話者による評価の平均値である。ちなみに、本手法の正解率は 76.9%に達成した。平易さ、文法性、意味の保持、言い換え率とも他手法より高い精度を得た。その結果は他の手法より複数の言語資源の併用によるものと考えられる。一方、現時点では実験用データはわりあいに少ない。今後、実験規模の拡大とデータの多様化も必要である。また、対訳コーパスの導入により、自然言語処理への適用に努めたい。

6. 終わりに

語彙平易化技術は、子どもや言語学習者をはじめとするさまざまな読者の文章読解を支援する道具である。また、自然言語処理における言語理解や機械翻訳などのタスクの精度を向上にも役立つことが期待される。本研究では、シソーラスから同義語を抽出し、1 語対 1 語の平易語生成の作業を行う。同義語も難解語の場合、国語辞書から語釈文を用いて文脈解析を加えて、難解語の平易化作業を入念に実施する。提案手法では日本語読者にとってより正確に情報を伝達するため、難解語の変換過程における平易さ、意味の保持と文脈の整合性を重視する。段階的成果として、難解語の変換精度は 76.9%、変換率は 78.5%、意味維持度の高い結果を得た。今後、語彙的曖昧性を解消し、正解率をさらに向上することに努める。さらに、実験範囲と適用分野を拡大し、実用性の高い語彙平易化システムを構築する。また、読者のレベルの個人差を配慮し、それぞれの難解語を特定する方向に努めたい。

参考文献

（1） 奥村　学（監），永田　亮（著），“言語学習支援のための言語処理”, pp.103-116, コロナ社（2017）

（2） 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(2015.3)

（3） Juri Gantikevitch, Benjami Van Durme, Chris Callison-Burch. PPDB: The Paraphrase Database, In Proc. Of NAACL, pp.758-764(2013)

（4） Gastavo Henrique Paetzold, Lucia Specia, “Lexical Simplification with Neural Ranking” In Proc. Of EACL, Vol.2, pp.34-40(2017)

（5） Connine, C.M., Mullennix, J. Yelen J. “Word Familiarity and Frequency in Visual and Auditory Word recognition” Journal of

Experimental Psychology, Vol.16, No.6, pp.1084-1096(1990)

（６） 佐藤理史,“均衡コーパスを規範とするテキスト難易度判定”, 情報処理学会論文誌,Vol.52,No.4, pp.1777-1789 (2011)

（７） Zhu, Z., Bernhard, D., Gurevych, I. “A Monolingual Tree-based Translation Model for Sentence Simplification.” In Proc. of the 23rd International Conf. on Computational Linguistics, pp.1353-1361(2010)

（８） Biran, O., Brody, S. Elhadad, N. “Putting it Simply: a Context-Aware Approach to Levical Simplification, In Proc. Of the 49th Annual Meeting Of ACL: Human Language Technologies, pp.496-501 (2011)