難解な表現の言い換えによるテキストの平易化

呉　浩東

Simplification of Text by Paraphrasing Esoteric Expressions

GO Kotoh

Abstract

Text simplification is a paraphrasing technique that rewrites esoteric expressions in sentences or documents while preserving the meaning of the text to be processed. Text simplification technology is a tool that supports reading comprehension by various readers, including children and language learners. It is also expected to be useful for improving the accuracy of tasks such as language understanding and machine translation in natural language processing. In this research, we aim to simplify difficult expressions using multiple language resources. We proposed a method emphasizes simplicity, preservation of meaning, and consistency of context in the process of converting esoteric words and phrases. As a step-by-step result, the conversion accuracy of esoteric words was 78.4%,

1．はじめに

テキスト平易化は、処理対象であるテキストの意味を保持しつつ平易な文や文書などに書き換えるという言い換え技術である。言語には同じ情報を伝える表現がいくつも用意されている。意味が近似的に等価な言語表現の異形を言い換えという。テキストの平易化は文章読解支援や文章作成支援、機械翻訳の事前と事後処理、情報検索、対話システムなどの自然言語処理タスクに適用できるという。

近年、日本語学習の需要や学習者の人口は増える傾向にある。より効果的に情報を理解するために、レベルの違う読者の言語能力の格差を埋める技術が必要である。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や難解句の存在である。難解語や難解な表現が含まれるテキストの平易化により学習者の言語能力の向上に役立つものである。

機械翻訳を用いた日本語文章の平易化の方法として、ニューラル機械翻訳（NMT）のツールであるDeepLによる折り返し翻訳を採用する。折り返し翻訳とは原言語を対象言語に翻訳し、その翻訳結果を再び原言語に翻訳することを指す。本研究では、原言語を日本語、対象言語を英語に設定している。

テキスト平易化は難解語や難解な表現を平易な語や句に置き換えるタスクである。日本語はテキスト平易化のためのアライメントされたコーパスに乏しい言語では、機械翻訳ベースの手法の有効性は限定的である。我々はこの問題を対処するため、言い換えルール、単言語コーパス、国語辞典の語釈文、WEB活用を合わせて利用する統合的な平易化モデルを提案する。

２．関連研究

文中の難解語の他の平易表現への変換に関する研究は数多く行われている。近年、平易な表現への研究はニューラル機械翻訳の手法が特に盛んである［Maruyama 19, Nisioi 17, Zhang 17, Surya 18, Zhao 18］。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が盛んに行われた。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; 乾、藤田 2006）の手法も報告されている。国外では語の変換に着目した評価型ワークショップが開催された（McCarthy et al. 2007, Jauhar and Specia 2012, Sinha 2012）という研究が報告されている。

機械翻訳を用いた日本語文章の平易化の方法として、折り返し翻訳が考えられる。鈴木ら（2020）は、被験者が正しく簡潔に情報伝達を達成する文章が書けるようになるべく、機械翻訳を用いた折り返し翻訳を利用した文章推敲手法が効果的であるかについて研究を行った。結果として、各被験者は機械翻訳を利用することで、冗長表現等の問題を含む文章を修正できたことから、機械翻訳を利用した折り返し翻訳の文章推敲への有用性が示唆された。

本研究は複数の言語資源を使って、ニューラル機械翻訳、言い換えルール、シソーラスを統合するアプローチを提案する。

3．言い換えによるテキストの平易化

3．1　言い換えの分類

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＝システムと人間で共通する言い換えの数/人間が言い換えた語の数

P＝システムと人間で共通する言い換えの数/システムが言い換えた語の数

奥村、永田（2017）の指摘の通り、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。

乾らは、語彙・構文的言い換えを、次の6つに分類した（乾、藤田 2004）。

（1）節間の言い換え
（2）節内の言い換え
（3）内容語の複合表現の言い換え
（4）機能語/モダリテイの言い換え
（5）内容語の言い換え
（6）慣用表現の言い換え

3．2　語彙の言い換え

語彙的換言処理では、表記の違い（妻 ⇔ 家内 ⇔ ワイフ）や動詞の変化（入手する ⇔ 手に入れる）などを扱う。

語彙平易化は難解語を平易な語や句に置き換えるテキスト平易化のサブタスクである。低頻度語を言い換え対象とする。

例1：

s. 失敗する**公算**が大きい
t. 失敗する**可能性**が大きい

例２：
s. タイン川流域のローマ時代の**城塞**
t. タイン川流域のローマ時代の**城**

表３．１　難解語の例

二字熟語	意味	例文	BCCWJ出現頻度
矜持	自負、プライド	「男の矜持を大事にしろ」	59
蹂躙	暴力・強権などをもって他を侵害すること	「弱小国の領土を蹂躙する」「人権蹂躙」	30
僥倖	思いがけない幸い。偶然に得る幸運	「僥倖を頼むしかない」	51
狼狽	不意の出来事などにあわててうろたえること	「株価の急落に狼狽する」	108
逼迫	行き詰まって余裕のなくなること。事態が差し迫ること	「財政が逼迫する」	71
刹那	瞬間	「衝突した刹那に気を失う」	0
伴侶	なかま。また、配偶者	「人生の伴侶を得る」	165
反駁	反論	「例をあげて反駁する」	57

我々は低頻度語または基本語彙外の言い換えの戦略として、以下のアルゴリズム（折り返し機械翻訳あり）を提案する（手法１）。

STEP 1. 形態素解析ツール JUMAN++ を使って単語と句を抽出する。
STEP 2. 日本語均衡コーパスBCCWJを使って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（E1）として抽出する。
STEP 3. オンライン辞書からE1の語訳文を取得する。
STEP 4. 機械翻訳ツールDeepLを使用して英語の訳文を取得する。その訳文を英語から日本語に折り返し翻訳を行う、結果をE2とする。
STEP 5. シソーラスから、言い換える語の候補を取得する。
STEP 6. BCCWJより、各候補の頻度を取得し上位の３語をリスト（L）に格納する。頻度が閾値（100）より低いものはLから削除する。
STEP 7. E1とE2は一致しない場合、STEP 8に移動する。E1とE2は一致する場合、E2をLから除外しSTEP 9に移動する。
STEP 8. E2は語釈文の表現の同義の場合、E2はE1の言い換えにし、処理を終

了する。
STEP 9. Lの類語の順番に重み5、4、3に付ける。重み×頻度で算出するスコアの１位を言い換え語にし、処理を終了する。

以下の例はアルゴリズムの仕組みを示す。
例３：蹂躙
折り返し機械翻訳の結果：侵害
シソーラス：侵害、侵略、侵犯
BCCWJ：侵害　頻度：1167　　侵略　頻度：1177　　侵犯　頻度：57
例４：狼狽
折り返し機械翻訳の結果：混乱
シソーラス：混乱、錯乱、動乱
BCCWJ：混乱　頻度：1879　　錯乱　頻度：98　　動乱　頻度：207
例５：僥倖
折り返し機械翻訳の結果：僥倖
シソーラス：好運、幸運、幸い
BCCWJ：好運　頻度：０　　幸運　頻度：1201　　幸い　頻度：650
語訳文：思いがけない幸い。偶然に得る幸運
例６：逼迫
折り返し機械翻訳の結果：逼迫
シソーラス：急迫、緊迫、切迫
BCCWJ：急迫　頻度：67　　緊迫　頻度：131　　切迫　頻度：130
語訳文：行き詰まって余裕のなくなること。

３.３　冗長表現の言い換え

冗長表現とは、無駄な語句や言い回しが含まれており、意味が伝わりづらい文章表現のことである。
５つの冗長表現：
（１）文末表現の冗長
（２）文中表現の冗長
（３）同じ単語の連続
（４）類語の重複
（５）二重否定

冗長表現をなくすと文章がシンプルになるので、読みやすくわかりやすい印象に変わる。

文末表現の冗長：

文末でよく使われる冗長表現は、「～することができます」もしくは「～であるものである」という表現である。これらは文法的には誤っていないが、多用しすぎると読みにくくなってしまう。冗長表現をなくすと文章がシンプルになるので、読みやすくわかりやすい印象に変わる。

例７：

s. タバコと酒を販売**することができる**。

t. タバコと酒を販売**できる**。

文中表現の冗長：

文中によく見られる冗長表現は「～という」や「こと」である。

例８：

s. この挑戦は指導部にとって重要な意味が**あるものである**。

t. この挑戦は指導部にとって重要な意味がある。

上記の文中にある「あるものである」の部分が冗長表現である。

単語の重複：

一文中に同じ単語を繰り返し使うと、読みづらい文章になる。

例９：

s. パソコンのほうが使いやすいため、パソコンを使い続ける人もいます。

t. パソコンのほうが使いやすいため、そのまま使い続ける人もいます。

類語の重複：

同じ意味の言葉を重複して使うと、くどい印象を与えてしまう。

例10：

s. **まず最初に**、ご飯を食べよう。

t1. **まず**、ご飯を食べよう。

t2. **最初に**ご飯を食べよう。

二重否定：

二重否定とは、否定文を連続使用することである。

例11：

s. 彼は山登りが**できないわけではない**。

t. 彼は山登りができる。

例12：
s. 明日のプレゼン資料は完成**しないとは限らない**。
t. 明日のプレゼン資料は完成する**可能性もある**。

４．提案手法

我々は難解表現の検出のために
（１）言い換え候補の生成
（２）言い換え候補からスコアを算出し、ランキングを生成する
（３）言い換えの有効性を分析し後処理を行う
また、言い換え生成に使う言語資源は下記のものである。
（１）国語辞書
（２）シソーラス
（３）コーパス
（４）WEB
（５）変換ルール
ここで、難解な表現を対処するため、以下の言い換え方法を導入する。
構文的言い換え
　構文的換言処理は言語知識によって換言規則を作り、適用可能なものを処理する。また、変形文法を利用し語順を調節する。
例13：
変換規則　ＮしかＶない　⇒　ＮだけＶ
s. 橋本君しか英検一級に合格しなかった。
t. 橋本君だけ英検一級に合格した。
辞書の定義文による言い換え
例14：
. よく話して納得させる　⇔　説得する
. 物事を理解して承認する　⇔　納得する
. 相手の言い分を聞き入れる　⇔　承認する
パラレルコーパスによる言い換え
例15：
s. いい　です　か？
t. よろしい　です　か？
t. いい　のです　か？

t. いい　でしょう　か？
折り返し機械翻訳による言い換え
例16：
体がぽかぽか
The body is warm　（和文英訳）
体が暖かい　（折り返し翻訳）
WEB検索とシソーラス併用
例17：
「山道がツルツル」はGoogleで検索すると、6280件である。
「道がツルツル」はGoogle検索で、1,060,000件である。
「雪道がスベスベ」、件数が０回である。
コーパス、Web検索と辞書の併用
BCCWJを使って、「羊頭狗肉」の現す頻度は０である。Googleの件数が22,600件である。
Goo辞書で釈文は「羊頭（ようとう）を掲（かか）げて狗肉（くにく）を売（う）る」である。意味は「表面と内容とがくいちがうこと」である。
難解語の解消法（その２）
われわれは前述の考え方に基づいて、難解語の平易化（折り返し機械翻訳使用なし）過程を以下のモデルにする。（旧来の手法、手法２）
難解語の平易化モデル２：

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．リストに残る候補語はシソーラスを調べる。
４．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く高い場合、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。

本研究では、機械翻訳ツールDeepLを採用し、下記の種類の翻訳を実施する。
１．難解語の翻訳（逐語訳）
２．難解句の翻訳（意訳）
３．難解文の翻訳（等価変換）
４．長文翻訳（文分割による要約）
難解句の翻訳例
例18：
体がぽかぽか
The body is warm （和文英訳）
体が暖かい （折り返し翻訳）
長文翻訳の例
例19：
歌唄いが来て歌唄えと言うが、歌唄いくらい歌うまければ歌唄うが、歌唄いくらい歌うまくないので歌唄わぬ
（引用元：高齢者のための役立ち情報ブログ～３歩進んで２歩下がる～）
DeepLによる英訳：
A singing comes and says to sing, but if you sing as well as singing, you will sing, but if you sing well, you will sing, but you will not sing because you are not good at singing.
DeepLによる英語から日本語への折り返し翻訳
歌が来て歌えと言うが、歌うだけでなく歌うなら歌うが、上手に歌えば歌うけど歌うのが苦手なので歌わない。

５．実験結果

われわれは、日本語Wikipediaから難解な表現を抽出し、その中に異なる1056の難解な表現を含む984文を選択する。「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。ここでは
（１）出力された表現と入力表現に意味的に等価性を判定する
（２）出力された表現は難易度のレベルに対応しているかをチェックする

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	正解率
旧来の手法（折り返し翻訳なし）	4.05	4.17	3.58	74.3
本手法（折り返し翻訳あり）	4.42	4.38	3.84	78.4

実験結果は表５．１に示す。「旧来の手法（折り返し翻訳なし）の言い換えの正解率は74.3%である。ちなみに、本手法の正解率は78.4%に達成した。平易さ、文法性、意味の保持、言い換え率とも高い性能を得た。その結果は旧来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こと、今後、実験規模の拡大とデータの多様化も必要である。ニューラル機械翻訳による折返し翻訳の有効性も確認した。

６．おわりに

本論文では、複数の言語資源を使い、変換精度の高いテキストの平易化モデルを提案する。われわれは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、テキストの中に難解な表現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。また、多様な文脈中での難解な表現の平易化を実現した。段階的成果として、難解語の変換精度は78.4%の意味維持度の高い結果を得た。今後、実験範囲と適用分野を拡大し、実用性の高いテキスト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、より多くの難解な表現における変換精度の向上を目指したい。

参考文献

［１］Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio（2014）: Neural Machine Translation by Jointly Learning to Align and Translate, arXiv: 1409.0473

［２］［Gehring 17］Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin（2017）: Convolutional Sequence to Sequence Learning, arXiv: 1705.03122

［３］［Luong 15］Minh-Thang Luong, Hieu Pham, Christopher D. Manning（2015）: Effective Approaches to Attention-based Neural Machine Translation, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.1412-1421

［４］Takumi Maruyama, Kazuhide Yamamoto（2017）: Sentence Simplification with Core Vocabulary, Proceedings of the International Conference on Asian Language Processing, pp.363-366

［５］Manami Moku, Kazuhide Yamamoto, Ai Makabi（2012）: Automatic Easy Japanese Translation for information accessibility of foreigners, Proceedings of the Workshop

on Speech and Language Processing Tools in Education, pp.85-90
［6］Sergiu Nisioi, Sanja Stajner, Simone Paolo ˇPonzetto, Liviu P. Dinu（2017）: Exploring Neural Text Simplification Models, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics（Volume 2: Short Papers）, pp.85-91
［7］Sai Surya, Abhijit Mishra, Anirban Laha, ParagJain, Karthik Sankaranarayanan（2018）: Unsupervised Neural Text Simplification, arXiv: 1810.07931
［8］Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser（2017）Illia Polosukhin: Attention Is All You Need 31st Conference on Neural Information Processing Systems（NIPS 2017）, pp.5998-6008
［9］Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, Hong Yu（2018）: Sentence Simplification with Memory-Augmented Neural Networks, Proceedings of NAACL-HLT, 79-85.
［10］Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch（2016）: Optimizing Statistical Machine Translation for Text Simplification, Transactions of the Association for Computational Linguistics, Vol.4, pp.401-415
［11］Xingxing Zhang（2017）, Mirella Lapata: Sentence simplification with deep reinforcement learning, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp.584-594
［12］Juri Gantikevitch, Benjami Van Durme, Chris Callison-Burch（2013）
PPDB: The Paraphrase Database, In Proc. Of NAACL, pp.758-764
［13］Gastavo Henrique Paetzold, Lucia Specia（2017）“Lexical Simplification with Neural Ranking”In Proc. Of EACL, Vol.2, pp.34-40
［14］Biran, O., Brody, S. Elhadad, N（2011）“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
［15］加藤汰一・宮田玲・佐藤理史（2021）「説明文を対象とした日本語文末述語の平易化」情報処理学会論文誌 Vol.62 No.9 1605-1619
［16］奥村学（監）・永田亮（著）（2017）「言語学習支援のための言語処理」pp.103-116, コロナ社
［17］梶原智之・山本和英（2015）「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
［18］佐藤理史（2011）“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
［19］鈴木諒輔・角康之（2020）「折返し翻訳を用いた文章推敲手法の提案」情報処理学会 インタラクション』1081-1086.
［20］藤田篤・柴田知秀・松吉俊・渡邉陽太郎・梶原智之（2015）「言い換え認識技術の評価に適した言い換えコーパスの構築指針」言語処理学会第21回年次大会ワークショップ「自然言語処理におけるエラー分析」発表論文集, pp.1-11
［21］今村賢治・越前谷博・江原暉将・後藤功雄・須藤克仁・園尾聡・綱川隆司・中澤敏明・二宮崇・王向莉（2022）「特許機械翻訳の課題解決に向けた機械翻訳技術解説」自然言語処理 第29巻 第３号, 925-985.
［22］中川慎太郎（2021）「機械翻訳を用いた日英・英日翻訳の特長と課題」『2021年度ゼミ

論』
［23］呉浩東（2019）「機械翻訳の原理と研究動向」『マテシス・ウニウェルサリス』第20巻 第２号, 27-41.
［24］宮部真衣・吉野孝（2012）「折返し翻訳文と対象言語翻訳文の精度不一致要因」『電子情報通信学会論文誌』 第J95-D巻 第１号, 11-18.
［25］中澤敏明（2017）「機械翻訳の新しいパラダイム―ニューラル機械翻訳の原理」情報管理 第60巻 第５号, 299-306.
［26］丸山拓海・山本和英（2019）「ニューラル機械翻訳による公的文書平易化」第33回人工知能学会全国大会論文集
［27］加藤汰一・宮田玲・佐藤理史（2021）「説明文を対象とした日本語文末述語の平易化」情報処理学会論文誌 Vol.62 No.9 1605-1619
［28］網川隆司（2020）「ニューラル機械翻訳における長文翻訳のための分割による言い換え方法の検討」JAPIO pp.288-291