クラウドソーシングによる 大規模なやさしい日本語換言辞書の構築

角張竜晴　　山本和英
長岡技術科学大学
{kakubari, yamamoto}@jnlp.org

1 はじめに

文章の自動平易化は、複雑な文章の意味を保持しつつ、簡単な単語や文法を用いて書き換えるタスクである。平易化された文章は、理解しやすい情報を提供することができ、言語学習者や外国人に対して重要な役割を果たす。そのため、近年は、英語を始めとする様々な言語で統計的機械翻訳 (SMT) を用いた自動平易化の研究がされている (Zhu et al., 2010)。自動平易化の研究が比較的進んでいる英語では、Simgle English Wikipedia [1] という大規模な平易化されたデータがある。一方で日本語は、語彙平易化の研究がされており (?; Hading et al., 2016)、文脈を考慮した平易化があまり進んでいない。何故ならば、日本語の平易化データで公開されているものがなく、機械翻訳に用いるために必要な言語資源が不足しているからである。よって、機械翻訳を利用して意味や文脈を考慮した自動平易化の研究が思うように進められない状況にある。そこで、我々は先行研究でやさしい日本語対訳コーパスを構築し (山本, 2017)、その過程で基礎語彙を 2,000 語選定した。しかしながら、平易化データは未だに少なく、より多くの言語資源を構築するには非常に時間が必要であり、我々だけの作業量では限界がある。

そこで、大規模な言語資源を構築する方法に、クラウドソーシングを用いた研究がある (Finin et al., 2010)。クラウドソーシングでは、不特定多数の人に作業依頼の募集をかけることができ、作業内容や作業期間、報酬などの条件を提示し、合意が得られた作業者に対して、指示し作業を行なってもらうことができる。よって、大規模な言語資源を構築する上で有効な方法である。

以上のことから、本研究では、クラウドソーシングを利用して、大規模なやさしい日本語換言辞書を構築した。対象となる換言対象は、BCCWJ と田中コーパス [2] に含まれる基礎語彙ではない高頻度語（以下、難解語という）である。本論文では、実際にクラウドソーシングを行なった作業過程で得られた知見を報告するとともに、そのクラウドソーシングの利用方法について考察する。

2 作業

2.1 対象語彙

換言対象である難解語は、以下のような特徴がある。

- BCCWJ とやさしい日本語対訳コーパスの原文に出現する難解語のうち高頻度 20,000 語を換言対象とする。
- 上記の難解語のうち、オノマトペや固有名詞、アルファベットなどは換言対象から除外する。また、書き換えが不要だと考える語（漢数字「五十」や「七十」のように換言することが有意義ではない場合）も対象外としている。

本研究で利用したクラウドソーシングサービスは、クラウドワークス [3]である。作業データは上記の難解語を 5,000 語ずつ計 4 つのファイルにまとめ、作業者にはいずれか 1 つのファイルに対して換言作業を行なったもらった。作業者には Google スプレッドシートでデータを共有し、作業を行なってもらうことで、我々が作業の進捗状況を把握で

[1] https://simple.wikipedia.org/

[2] http://www.edrdg.org/wiki/index.php/Tanaka_Corpus

[3] https://crowdworks.jp/

きるようにした。共有したデータの項目とその説明を以下に示す。

表 1: 作業者に提示したデータの項目とその説明

セルの項目	説明
例文	コーパス中で難解語を含む文
難解語	やさしい日本語辞書の基礎語彙ではない語
品詞	換言対象である難解語の品詞
書き換えた表現	作業者が換言結果を記述する
書き換え対象の文字列	書き換えた表現と対応する例文の文字列

2.2 作業方針

本研究で構築する換言辞書は、先行研究ではできなかった意味や文法を考慮した自動平易化を行なうために構築している。

そのため、難解語とその例文を提示することで意味や文法を意識して換言するように指示した。従って、作業者によって書き換えられた表現は、例文中の難解語と同じ意味になる。だが、文法を考慮する場合、難解語だけでなく周辺の語と書き換えた方が文法的に正しくなる場合がある。例えば、難解語が複合名詞の一部である場合やサ変動詞として用いられている場合がある。その際、表 1 で示した「書き換え対象の文字列」の欄に、換言表現に対応する例文中の文字列を記述してもらった。つまり、例文中の書き換え対象の文字列と書き換えた表現は、そのまま置換することが可能である。さらに意味的にも文法的にも正しい表現の対であり、機械翻訳の訓練データとして容易に用いることができる。

2.3 作業手順

クラウドソーシングに依頼する前の処理から検収までの具体的な流れを以下に示す。本研究ではデータを data_1 から data_4 の 4 つに分割しており、各データの換言作業に作業者を 3 人ずつ割り当てた。換言作業は 2017 年 12 月から 2018 年 1 月中旬までの 1ヶ月半で実施した。

1. BCCWJ と田中コーパスを解析することで、コーパス中の難解語を抽出し高頻度順にソートする。次に、人手で大まかに高頻度な難解語のうち対象外の難解語を除外する。その際に、作業者が難解語の意味を推定できない場合は、出来るだけ容易に意味を理解できる例文と入れ替える。実際に、作業者に提示した例文の平均的な文字数は、約 33 文字である。

2. 上記の調整後、高頻度順に 5,000 語毎に 1 つのファイルにまとめ、データセット data_1 から f_4 を作る。そして、クラウドソーシングで作業者を募る。

3. 作業者に data_1 から data_4 のいずれかを共有し、換言作業を行なってもらう。換言作業は指定された難解語をやさしい日本語のみで可能な限り意味や文法を保持して行なう。換言作業によって意味が一部欠落してしまうことは避けられないが、2.2 節で述べたように、文法は書き換えた表現と書き換え対象の文字列が例文において置換できる関係が成り立つように行なう。

4. 適宜、作業者からの質問に回答し、指定されている難解語が固有名詞などの場合に報告してもらい、例文と難解語を新たに提示する作業を行なう。定期的に作業者の進捗状況を確認し、連絡をする。

5. 作業が終了した場合は、こちらが意図通りに書き換えが行われているか、記入漏れがないかを確認する。

書き換えた表現がやさしい日本語で構成されているかを判断するために、作業者にはやさしい日本語チェッカー [4]を使用するように案内した。

3 所感

今回のクラウドソーシングは、複数の作業者に同時で、作業者と対話しながら作業を進めた。そのため、次のような問題があった。

- 作業者ごとに質問や報告の頻度が異なるので、作業者によっては作業の二度手間になること

[4]http://box.jnlp.org/easy-japanese/checker

がある。作業者の中には契約して間も無く、連絡が途絶える人もいる。

- 作業者間で進捗状況の差が大きい。
- 作業者からの質問への回答や報告の迅速な対応が必要である一方で、依頼した作業者が多いことでその対応に追われてしまう。

まず、クラウドソーシングに依頼する前に、データセットの作成の段階で少人数で書き換え作業を行なう必要があると考える。その理由は、本研究では人手で大まかなチェックを行ない、換言対象外の語を除いたが、実際に作業者から報告される修正依頼は多かった。そこで、大規模な言語資源を構築する場合には、下記の 2 段階で行なうことを推奨する。

1. まず少人数の作業者に依頼をし、作業対象外の語をデータセットから除外・修正を行なう。
2. 次に、大規模に募集をかけて、換言作業を依頼する。

その結果、作業者からの修正依頼の報告が減り、作業の二度手間を防ぐことができる。また、一度作業を行なってもらった作業者に、別のデータの作業を依頼することで、質問の対応は最低限で済み、作業結果の品質も確保することができる。

次に、作業者との対話によって得られたやさしい日本語の特徴を以下に示す。

- 「レモンとライムはすっぱい果物である。」のように、同一カテゴリの語が併記されている場合、制限された語彙だけでその違いを表現することは困難である。
- 法律用語や医学用語、人体の組織名などの専門性が高い難解語（ex. 「胆汁」「気功」）は、文の意味を保持することが難しい。

これらの特徴があるため、同一カテゴリの語が併記されている場合には、可能な限り難解語の特徴を明記して表現することで差別化している。併記された語を差別化できない場合は、それらの語をまとめて、カテゴリ名に書き換えるように指示した。また、専門性が高く換言が困難な語の品詞は名詞であり、動詞では報告されなかった。その結果から、現在の基礎語彙は一般的な文章で用いられる動詞の多くを表現可能であると言える。一方で、名詞の難解語を書き換えるには、意味を保持しようとすると冗長になりやすく、また定義文のようになるため好ましくない。従って、今後は名詞の難解語に対して、人手によって適度に意味を保持した書き換えが必要である。

4 収集したデータ

本研究で収集したデータの例を表 2 に示す。現段階で完成したデータは、data_1 が 2 人、data_2 が 2 人、data_3 が 1 人が完成した。data_4 は 1 人が作業中である。

実際に得られたデータでは作業者が異なっていたとしても、例文 1 のように書き換えた表現の一部が異なる場合があった。似た表現になる原因は、使用可能な語彙数を制限し、換言作業の基準となる例文を提示したためである。換言作業を依頼する場合には、例文を提示することで、作業者間の一致を取ることができる。例文 2 から、作業者によって、難解語周辺の語を積極的に巻き込んで、スムーズな文に書き換える場合と、難解語のみを換言した最小コストの場合の違いがあった。どちらもやさしい日本語で表現されているが、サ変動詞より動詞で表現されている方がより平易な印象がある。例文 3 では、難解語を書き換えた表現が作業者によって異なるが、それぞれは意味的にも文法的にも正しい。一方、例文 4 では、書き換えられた表現が文中で表している現象が異なる。「目に入る」と換言された場合、「林の中にある竹を見る」という動作を表すが、「たくさんある」と換言された場合は、「林の中に竹が存在している」という存在を表している。このように、表している現象自体は異なるとしても、どちらの換言表現でも文意を読み取ることができる。したがって、人によって理解しやすい表現が異なるため、換言作業を複数の作業者に依頼することで様々な表現を収集することができる。また、換言に使用できる語彙数を制限していることもあり、書き換えができないが難解語を削除することで文を理解しやすくなる場合がある。例文 5 では、1 人の作業者が難解語の「本部」を削除することで文が意味的にも文法的にも分かりやすくなると判断している。確かに、やさしい日本語を必要としている人が「大学の本部」と「大学」のような違いを理解する必要性は高く

ない。そのため、平易化する際には、難解語を削除することも有効な手段である。

表 2: 難解語をやさしい日本語へ換言した結果

例文 1		
パソコン を 持っ て ない 人 は 、 この ＭＰ３ プレーヤー 機能 を フル に 使い こなせ ませ ん 。		
換言作業者	書き換えた表現	書き換え対象の文字列
作業者 1	個人型のコンピューター	パソコン
作業者 2	自分のコンピューター	パソコン
例文 2		
政府 が 住民 に 問う 一般 投票 を 実施 し ま し た 。		
換言作業者	書き換えた表現	書き換え対象の文字列
作業者 1	実行	実施
作業者 2	行いました	実施しました
例文 3		
母親 は 病気 の 子供 の 背中 を さすっ た 。		
換言作業者	書き換えた表現	書き換え対象の文字列
作業者 1	背の中央	背中
作業者 2	体の後ろ	背中
例文 4		
林 に 竹 が 目立つ 。		
換言作業者	書き換えた表現	書き換え対象の文字列
作業者 1	目に入る	目立つ
作業者 2	たくさんある	目立つ
例文 5		
大学 本部 は ニューヨーク に 分校 を 設立 する こと を 決定 し た 。		
換言作業者	書き換えた表現	書き換え対象の文字列
作業者 1	活動の中心になる場所	本部
作業者 2	※削除	本部

5 おわりに

本研究では、クラウドソーシングで機械翻訳の訓練データに用いるための換言データを構築した。クラウドソーシングで大規模なデータを構築するには、作業者とのやりとりが必要であり、大規模な依頼前に少数の作業者に依頼し、指示やデータが理解しやすいか精査し調整するべきである。また、実際に収集されたデータから、人手で平易化する際の特徴を調査した。

謝辞

本研究は、平成 27〜31 年科学研究費補助基盤 (B) 課題番号 15H03216、課題名「日本語教育用テキスト解析ツールの開発と学習者向け誤用チェッカーへの展開」、及び平成 29〜31 年科学研究費助成事業挑戦的萌芽課題番号 17K18481、課題名「やさしい日本語化実証実験による言語資源構築と自動平易化システムの試作」の助成を受けています。

参考文献

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 80–88.

Hading, M., Matsumoto, Y., and Sakamoto, M. (2016). Japanese lexical simplification for non-native speakers. Proceedings ofthe 3rd Workshop on Natural Language Processing Techniques for Educational Applications, pages 92–96.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In Proceedings of the 23rd International Conference on Computational Linguistics, pages 1353–1361.

梶原, 智. and 山本, 和. (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, 56(3):983–992.

山本, 他. (2017). やさしい日本語対訳コーパスの構築. 言語処理学会第 23 回年次大会, pages 763–766.