日本語文法平易化コーパスの構築

稲岡　夢人　　山本　和英
長岡技術科学大学
{inaoka, yamamoto}@jnlp.org

1　はじめに

日本の在留外国人数は約 264 万人に達し、2012 年から連続して増加の傾向にある[1]。近年では掲示や配布物において日本語と英語を併記する表現が行われるようになっている。しかし日本国内で外国人へ情報伝達を行う場合、英語より日本語の方が伝達効率が高いことが [1] で示されている。そうした理由により「やさしい日本語」は近年重要性を増している考え方となっている [2]。それに関連して、自然言語処理の分野においては自動平易化について研究が行われている。自動平易化とは、語彙や文法が難解なテキストの意味を保持しつつ自動で平易なテキストに言い換えるタスクである。

我々は日本語文法平易化コーパスの構築を行った。構築においては、比較的短期間で大規模なデータを得ることを目的にクラウドソーシングを用いた。クラウドソーシングとは、インターネットを用いて不特定多数の人に作業を依頼し、合意が得られた人について作業を指示する手法である。また構築したコーパスを観察し、言い換えにおける現象についての分析を行った。

日本語の自動平易化の研究が進展することを期待し、以下の研究成果物を一般に公開する。

- 日本語文法平易化コーパス
- やさしい日本語文法の定義ルール
- やさしい日本語文法チェッカー

2　関連研究

Simple PPDB[3] は、英語の大規模な言い換え辞書 (PPDB)[4] から平易な言い換えを収集することで平易化のための辞書を構築している。Simple PPDB: Japanese [5] では日本語において同様の辞書を構築して平易化を行っている。また文単位で平易化を行う研究も行われており、英語では English Wikipedia[2] と Simple English Wikipedia[3] を用いて抽出した文対から平易化のための対訳コーパスを構築 [6] し、SMT や NMT といった機械翻訳手法によって平易化が行われている [7][8]。日本語においては、平易化のための対訳コーパスを人手で構築する研究 [9] や、それを用いて機械翻訳手法で平易化を行う研究 [10] が行われている。

Simple English Wikipedia で使用される英語には文法的に平易である制約がある一方、日本語の平易化コーパスである [9] は平易な文法を意識して作られてはいない。そのため、日本語の文法平易化研究のための対訳コーパス構築を行った。コーパスの作成に際して、「平易な文法」のルールについて明確に定義する必要があった。今回は日本語における「ミニマムの文法」として [2] で規定されている文法項目をやさしい日本語文法として定義した。そして、ある文がやさしい日本語文法に従っているか否かを自動で判別するやさしい日本語文法チェッカーを作成した。解析には MeCab[4], JUMAN 辞書, CaboCha[5] を用いた。ただし CaboCha は文節の分割のためのみに利用し、係り受けは考慮せず文節単位でのみチェックを行っている。そのため全ての文法項目を完全には網羅・再現できていない。またサ変動詞については文法項目にないものの、今回は便宜上[6]やさしい日本語文法に追加した。文法ルールに従わない文節と従う文節の具体例を表 1 に示す。文法ルールの定義は、従うべき文節の条件を列挙するホワイトリスト形式によって行った。やさしい日本語文法に従う文節の条件は形態素列によって定義した。今回は 130 の形態素列を定義し、その一部を表 2 に示す。

[1] https://www.e-stat.go.jp/

[2] https://en.wikipedia.org/
[3] https://simple.wikipedia.org
[4] https://taku910.github.io/mecab/
[5] https://taku910.github.io/cabocha/
[6] サ変動詞から一般動詞への書き換えを作業に含めると、作業量が大幅に増加すると考えたため。

3 作業

書き換えを行う原テキストは、[9] と同様に田中コーパス[7]の日本語文を用いた。田中コーパスは日本の大学生が授業の一環で教科書等の文を翻訳したものである。分野を限定せず、比較的語彙数が少なく短文が多いのが特徴である。

作業者はクラウドソーシングサービス「クラウドワークス」[8]を利用して募集した。各作業者が 5,000 文ずつの書き換えを担当し、全体で 50,000 文の書き換えを依頼した。本稿ではそのうち作業が完全に完了した 5 人のデータに対して分析を行った。作業中の文についても依頼を継続し、公開時点において完了したデータを公開する。

作業は前述のやさしい日本語文法チェッカーを用いて行った。具体的には、やさしい日本語文法チェッカーで「やさしくない」と判定された文に対して、全ての文節が「やさしい」と判定されるように書き換えを繰り返すことで文全体の書き換えを行う。ただしやさしい日本語文法チェッカーは自動で品詞や文節等を解析して判断を行うため、解析誤り等が生じ、人間の感覚とは反した出力を行うことがある。その際は作業者からその旨の連絡を受け、解析誤りの箇所については無視して作業を行ってもらった。

4 結果と分析

4.1 結果

作業者による書き換えの例を表 3 に示す。1 や 2 のような完全に意味を保持しているような書き換え、3 や 4 のような多少ニュアンスが変化するが大意を保持した書き換え、5 や 6 のような一部の情報を落とした書き換え、7 のような使用する内容語を変更するような書き換えがみられた。一方で 8 のような、やさしい日本語文法チェッカーの解析誤りによって「やさしくない」と判定されたものに対して読点「、」を追加し、解析結果を変化させるような書き換えを行っているような例も見られた。これは我々が望む書き換えではないため、作業者への提示方法の改善、フィルタリング等を行う必要があると考える。

表 1: 文法ルールに従わない文節と従う文節の例

文法ルールに従わない文節	文法ルールに従う文節
元気である 。	元気です 。
滅ぼさ れた 。	滅び ました 。
食べ なかった 。	食べ ませ ん でした 。
彼 ほど	彼 より
考えて みる	考える
知らずに	知らないで

4.2 書き換え前後での文節数の変化

作業者による書き換えの例 (表 3) での 1, 3, 7 は、一文節の表現から一文節の表現へ書き換えが行われている。一方で 2 はより文節の多い表現への書き換え、4 はより文節節の少ない表現へ書き換えが行われている。また 5 や 6 のような、1 文中の複数箇所で書き換えが行われているような例もある。このように、書き換え前後での文節数の変化には様々なものがある。書き換え前後での文節数の変化を頻度で調査した結果を表 4 に示す。やさしい日本語文法チェッカーにおいて「やさしくない」とされる文節を書き換える際、一文節の表現から一文節の表現へ書き換えられている割合が約 66%、より多くの文節による表現へ書き換えられている割合が約 20%、より少ない文節による表現へ書き換えられている割合が約 6.1%となった。残りの約 8.1%は、同じ文節数の表現へ書き換えが行われたものである。これはやさしい日本語文法への書き換えが文節内の書き換えだけで事足りる場合が多く、そうでない場合はより多くの文節を使った表現へ書き換えられると考えられる。

表 2: 文法ルールに従う文節の定義例（$は文末）

	形態素列	
（動詞-連用形）	ます	$
（形容詞-連用形）	なり	ました $
（名詞）	じゃ	ないで す $
（名詞）	を	
（動詞-未然形）	たい	と
（動詞-未然形）	たい	んです が

[7]http://www.edrdg.org/wiki/index.php/Tanaka_Corpus
[8]https://crowdworks.jp

表 3: 作業者による書き換えの例

	書き換え前の文	書き換え後の文
1	私の ペンを 使うな。	私の ペンを 使わないでください。
2	妹が そこに いるのが 見えた。	妹が そこに いる 姿が 見えました。
3	彼女は その 話が 不思議に 思えてならなかった。	彼女は その 話が 不思議に 思えました。
4	私は 笑わないわけには いかなかった。	私は 笑いました。
5	演劇が みたいのですが 情報を ください。	演劇の 情報を 教えてください。
6	これこそ 僕が 欲しいと 思っていた ものだ。	これが 僕が 欲しい ものです。
7	暗くならない うちに 寝てしまいました。	明るい うちに 寝てしまいました。
8	テーブルの 上に 猫が いま すか。	テーブルの 上に 猫が、 いますか。

表 4: 書き換え前後での文節数の変化

		書き換え後									
		1		2		3		4		5-9	
書き換え前	1	66.3%	(18,029)	9.96%	(2,710)	3.03%	(823)	0.221%	(60)	0.0294%	(8)
	2	3.49%	(949)	6.30%	(1,713)	4.15%	(1,129)	0.823%	(224)	0.221%	(60)
	3	0.592%	(161)	1.19%	(325)	1.48%	(402)	0.698%	(190)	0.287%	(78)
	4	0.0625%	(17)	0.169%	(46)	0.294%	(80)	0.235%	(64)	0.176%	(48)
	5-9	0.0184%	(5)	0.0331%	(9)	0.0956%	(26)	0.110%	(30)	0.0698%	(19)

表 5: 書き換えの規則

頻度順位	頻度	書き換え前	書き換え後
1	4,939	動詞-タ形 。	動詞-基本連用形 ました 。
2	1,973	動詞-基本形 。	動詞-基本連用形 ます 。
3	1,372	動詞-タ系連用テ形 いる 。	動詞-タ系連用テ形 い ます 。
4	1,101	名詞 動詞-タ形 。	名詞 動詞-基本連用形 ました 。
5	783	名詞 判定詞-基本形 。	名詞 判定詞-デス列基本形 。
⋮	⋮	⋮	⋮
25	115	名詞	名詞 に
25	115	動詞-タ系連用テ形 いる	動詞-基本形
27	112	形容詞-ダ列特殊連体形	名詞 の
⋮	⋮	⋮	⋮

4.3 書き換えの規則

コーパスに含まれる内容語を品詞+活用に一般化し、作業者による書き換えに存在する規則を抽出したものを表5に示す。頻度が高い上位の書き換えは文末の文体についてのものである。文体の統一は平易化において必要なものであるが、比較的自動での変換が容易なものである。そのため依頼前に事前に変換しておくことで作業者の負担を減らし、より変換が困難な部分についての書き換え作業に専念させられると考える。今回はコーパスの分析のために簡易的に書き換えの規則を抽出したが、抽出した規則を使用して文を変換することでも平易化は実現できると考える。

5 おわりに

日本語文法平易化コーパスの構築と分析を行った。収集した文から書き換え箇所を抽出し、一文節の表現から一文節の表現、より多くの文節による表現、より少ない文節による表現の順に行われることが多いことを示した。また書き換えの内容については、多くが文末における文体の統一が多くを占めることを示した。実際に各書き換え文を観察すると、完全に意味を保持したものからニュアンスが変化したもの、一部の情報を落としたような書き換えが存在し、また文法チェッカーの解析誤りに起因する書き換えも確認された。

成果物である対訳コーパス、やさしい日本語文法の定義ルール、やさしい日本語文法チェッカーはいずれも準備ができ次第一般公開を行う。また機会があれば、今後もコーパスやルールの拡充を目指す。

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、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 の助成を受けています。

参考文献

[1] 岩田一成. 言語サービスにおける英語志向：「生活のための日本語:全国調査」結果と広島の事例から. 第13巻, pp. 81–94, 2010.

[2] 庵功雄. 「やさしい日本語」 研究の現状と今後の課題. 一橋日本語教育研究, No. 2, pp. 1–12, 2014.

[3] Ellie Pavlick and Chris Callison-Burch. Simple ppdb: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 143–148, 2016.

[4] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 758–764, 2013.

[5] 梶原智之, 小町守. Simple ppdb: Japanese. 言語処理学会第23回年次大会発表論文集, pp. 529–532, 2017.

[6] William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 211–217, 2015.

[7] Sanja Štajner, Hannah Bechara, and Horacio Saggion. A deeper exploration of the standard pb-smt approach to text simplification and its evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 823–828, 2015.

[8] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 85–91, 2017.

[9] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[10] Takumi Maruyama and Kazuhide Yamamoto. Sentence simplification with core vocabulary. In *International Conference on Asian Language Processing (IALP)*, pp. 363–366. IEEE, 2017.