MATCHA：専門家が平易化した記事を用いたやさしい日本語パラレルコーパス

宮田 莉奈 [1]　惟高 日向 [1]　山内 洋輝 [1]　柳本 大輝 [1]　梶原 智之 [1]　二宮 崇 [1]　西脇 靖紘 [2]
[1] 愛媛大学大学院理工学研究科　[2] 株式会社 MATCHA
{miyata, koretaka, yamauchi, yanamoto}@ai.cs.ehime-u.ac.jp
{kajiwara@cs., ninomiya.takashi.mk@}ehime-u.ac.jp　nishiwaki@matcha-jp.com

掲載号の情報

31 巻 2 号 pp. 590-609.
doi: https://doi.org/10.5715/jnlp.31.590

概要

本研究では、テキスト平易化のための日本語パラレルコーパスを構築し、公開[1)]した。テキスト平易化とは、文の意味を保持しつつ、難解な文を平易に言い換えるタスクのことである。既存の日本語コーパスとしては、文を非専門家が平易化した SNOW [1, 2] や、専門家が平易化した JADES [3] がある。SNOW の平易化は、作者らによって定義された基礎語彙 2,000 語[2)]に基づくため、この語彙で網羅できない表現については、却ってわかりにくい言い回しが見られる。また、JADES は小規模である。既存の日本語テキスト平易化パラレルコーパスには、高品質かつ大規模なコーパスは存在しない。

そこで本研究では、高品質かつ大規模な日本語のテキスト平易化パラレルコーパスを構築するために、訪日観光客向けメディア MATCHA[3)]の日本語記事とその専門家による平易化版の記事[4)]の対から人手で文アライメントをとった。ただし、記事単位で平易化されており、全ての難解文に対して意味的に完全に対応する平易文が存在するわけではないため、完全一致と部分一致の 2 種類の文アライメントをとった。文対の抽出後、著者らが人手で誤字脱字を訂正し、括弧の用法などのスタイルを統一した。我々の MATCHA コーパスは、完全一致 11,000 文対と部分一致 5,000 文対の計 16,000 文対からなる。

表 1　日本語テキスト平易化コーパスの特徴

	専門家による言い換え	言い換えの多様性	言い換えの品質	規模
SNOW	-	★	★★	85,000
JADES	✓	★★★	★	3,907
MATCHA	✓	★★★	★★★	16,000

コーパスを人手評価した結果を表 1 に示す。専門家が平易化したパラレルコーパスは非専門家によるものと比べて多様な平易化操作を含んでいることが明らかになった。また、SNOW の平易化操作が語句の置換に集中している一方で、JADES や MATCHA には語句の挿入や削除、並び替えや文分割など多様な変換が含まれることもわかった。さらに、本コーパスは他のコーパスよりも流暢かつ意味を保持した平易化が行われていることを確認した。

SNOW および MATCHA を用いて事前訓練済み系列変換モデルをファインチューニングし、テキスト平易化モデルを構築した。実験の結果、MATCHA で訓練したモデルは SNOW で訓練したモデルに比べて、流暢かつ意味を保持した平易化ができることを確認した。MATCHA 上での訓練においては、完全一致の文対のみで訓練したモデルの方が流暢性および同義性に優れる一方、平易性については部分一致の文対も訓練に含めた方が高い性能を発揮した。

1) https://github.com/EhimeNLP/matcha
2) https://www.jnlp.org/GengoHouse/list/語彙
3) https://matcha-jp.com/
4) 日本語能力試験新 4 級相当（旧 3 級相当）の文法や単語を用いて、通常の日本語記事を書き換えて作成されている。

参考文献

[1] Takumi Maruyama and Kazuhide Yamamoto. Simplified Corpus with Core Vocabulary. In **Proc. of LREC**, pp. 1153–1160, 2018.
[2] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In **Proc. of LREC**, pp. 461–466, 2018.
[3] Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, and Taro Watanabe. JADES: New Text Simplification Dataset in Japanese Targeted at Non-native Speakers. In **Proc. of TSAR**, pp. 179–187, 2022.