[S1-P10]

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の構築

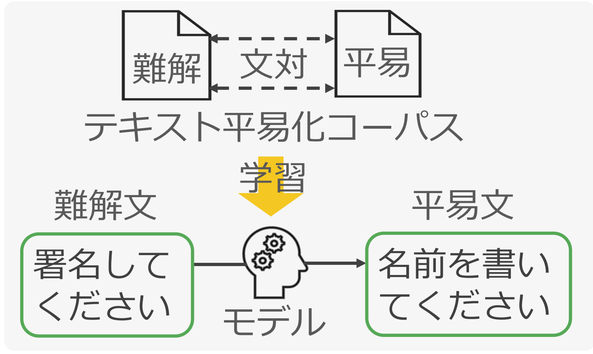
宮田莉奈, 惟高日向, 山内洋輝, 柳本大輝, 梶原智之, 二宮崇(愛媛大学), 西脇靖紘(株式会社MATCHA)

目的：高品質かつ1万件規模の文単位の日本語テキスト平易化コーパスを構築する

1. 背景：在留外国人や子供などへの情報伝達手段としてやさしい日本語を用いたい

- NHK NEWS WEB EASY [1] やMATCHA [2] などで，やさしい日本語コンテンツが毎日発信されている
- 難解文と平易文のパラレルコーパスがあれば，テキスト平易化でやさしい日本語コンテンツを自動生成できる
- しかし，**高品質な大規模コーパスは存在しない**

コーパス	専門家	規模
SNOW [3]	×	大
JADES [4]	○	小
MATCHA	○	中



[1] https://www3.nhk.or.jp/news/easy/ [2] https://matcha-jp.com [3] https://www.jnlp.org/GengoHouse/snow/t15 [4] https://github.com/naist-nlp/jades

2. データ構築：専門家が平易化した記事ペア ＋ 著者による人手の文アライメント

- データ　：訪日観光者向けメディアMATCHAから**専門家**が平易化した記事と元の記事を収集（2015年4月から2023年3月までの8年分の記事）
- 専門家　：日本語教育能力検定試験に合格し，技能実習生が対象の日本語教師を経験
- 構築手順：難解記事と平易記事から意味的に完全または部分的に対応する文対を人手で収集

MATCHAコーパス
https://github.com/EhimeNLP/matcha

統計情報	文対数		語彙サイズ	平均単語数	単語難易度
SNOW	85,000	難解	21,134	10.55	31.75
		平易	6,418	11.68	32.03
JADES	3,907	難解	12,571	30.80	32.15
		平易	5,498	25.21	31.31
MATCHA	16,000	難解	15,643	20.90	32.97
		平易	13,345	19.10	31.10

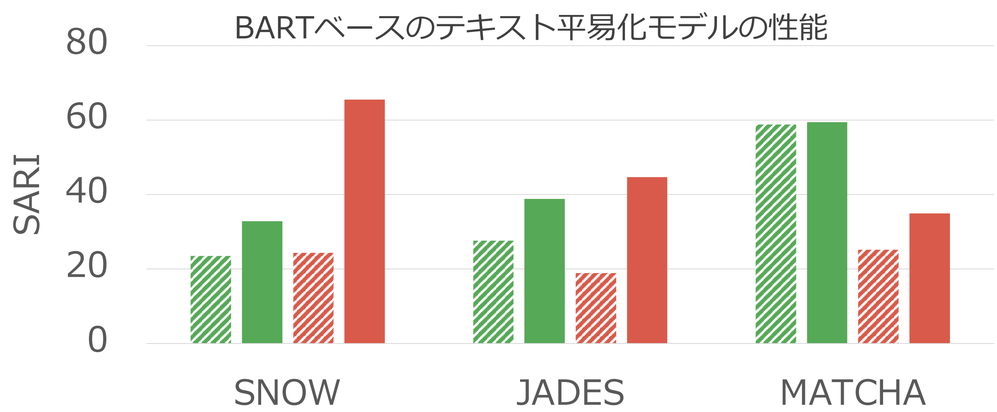
平易化する際に，

- 語彙サイズ： SNOW と JADES は減少
- 平均単語数： JADES と MATCHA は減少
- 単語難易度： JADES と MATCHA は減少

MATCHAは平易な単語を用いた短文で構成

3. 自動評価：同数のSNOWよりもMATCHAで訓練したテキスト平易化モデルが高品質

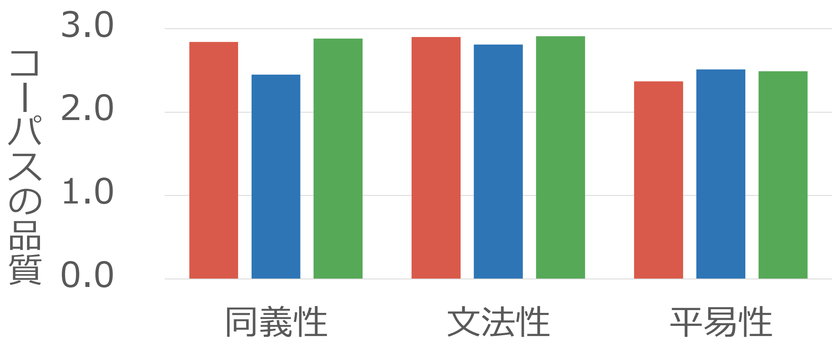
データセット		訓練	検証	評価
SNOW（MATCHAと同数）		15,050	2,000	700
SNOW		82,300	2,000	700
JADES		-	-	3,907
MATCHA（完全一致）		10,000	450	500
MATCHA（全体）		15,050	450	500



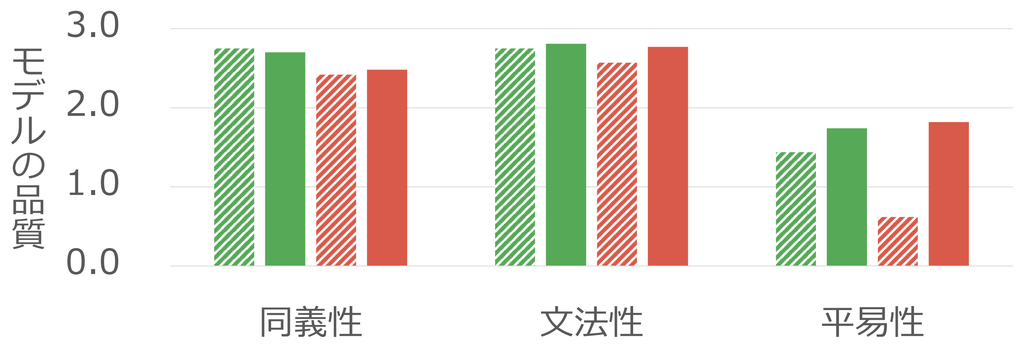
- 15kモデルの中ではSNOWよりもMATCHAの方が高性能 → MATCHAは高品質
- MATCHAの中では15kのモデルの方が高性能 → 部分一致の文対も有用
- SNOWの82kモデルが最高性能 → 質より量が重要！？

4. 人手評価：MATCHAコーパスは高品質だが，もっと量が必要

各コーパス200文とBARTの各モデル出力200文を同義性・文法性・平易性について4段階で評価



- SNOW　：平易性が低い
- JADES　：同義性が低い
- MATCHA：安定して高い → 高品質



- SNOW　：15kでは平易性が著しく低い
  82kでは平易性が最も高い
- MATCHA：同義性と文法性が高い
  部分一致を加えると平易性が向上