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試作

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

1. 概要

https://github.com/EhimeNLP/matcha

- 文単位の日本語のテキスト平易化コーパスを試作（1,792文対を公開）
- 専門家が平易化した記事から，人手の文アライメントで，パラレルコーパスを構築
- 我々のMATCHAコーパスは，既存コーパスよりも平易な文・多様な平易化操作を含む

2. 背景

- テキスト平易化：意味を保持しながら平易に言い換えること
- 近年の研究は，テキスト平易化を同一言語内の機械翻訳の問題として扱い，
  難解な文と平易な文のパラレルコーパス用いて系列変換モデルを学習
- 既存のコーパス(SNOW)は，非専門家により平易化されているため，低品質な文が含まれる
- 専門家により平易化された高品質なデータセットが必要

そこに署名してください
↓
そこに名前を書いてください

3. MATCHAコーパスの構築

<u>データ</u>

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）

＊ https://matcha-jp.com/

<u>構築手順</u>

1. 記事の前処理：
   記号の削除(自動)→文分割(自動と人手)→スタイルの調整(人手)
2. 文アライメント：
   難解な記事と平易な記事から，意味的に対応する文対を収集(人手)

MATCHA
鉢植えの植物に、人の手が加わり生み出される芸術、盆栽。
原型が作られたのは、2000年以上前の中国。

難解

MATCHA
盆栽(ぼんさい)は人(ひと)が鉢植(はちう)え(※)の植物(しょくぶつ)で作(つく)る芸術(げいじゅつ)(※)です。一番最初(いちばんさいしょ)の盆栽(ぼんさい)は、2000年(ねんいじょうまえ)以上前(いじょうまえ)に中国(ちゅうごく)で生(う)まれました。

平易

鉢植えの植物に、人の手が加わり生み出される芸術、盆栽。
原型が作られたのは、2000年以上前の中国。

盆栽は人が鉢植えの植物で作る芸術です。
一番最初の盆栽は、2000年以上前に中国で生まれました。

スタイルの調整（一部抜粋）	具体例（<u>下線</u>は追加する修正，~~取り消し線~~は削除する修正）
誤字の修正	また、炒め<u>物</u>~~者~~に肉が使われる時は豚肉のことが多いです。
脱字の修正	送れる国はこちらのサイトをチェックすれば詳<u>し</u>くわかります。
読み仮名は削除	万~~(man)~~は多いという意味で使われ、葉~~(you)~~とは言葉のこと。

4. 統計情報

- MATCHAの方が文が長い（ドメインの違い）
- MATCHAの方が平易な単語を多く使って
  平易化している（小さければより平易）

		文対数	文数	語彙サイズ	平均単語数	単語難易度
MATCHA	難解	1,792	1,801	4,552	20.37	33.63
	平易		2,097	3,560	18.66	31.64
SNOW	難解	85,000	85,265	21,135	10.55	31.74
	平易		85,834	6,418	11.68	32.03

5. 人手評価

- MATCHAとSNOWの各100文を平易化操作と品質で評価
- 平易化操作　SNOW　：語句の置換に集中
  MATCHA：多様な変換を含む
- 品質評価（4段階）　同義性・文法性：どちらも高い
  平易性　：MATCHAの方が平易

平易化操作	MATCHA	SNOW
語句の挿入	31	10
語句の削除	21	4
語句の置換	137	115
並び替え	13	5
文分割	14	0

品質	MATCHA	SNOW
同義性	3.9	3.9
文法性	3.9	3.9
平易性	3.4	3.1

6. まとめ

- 文単位の日本語の高品質なテキスト平易化コーパス構築
- MATCHAはSNOWよりも平易な文・多様な平易化操作を含む
- 今年度中に，完成版のコーパス(約16,000文対)を公開予定

M1の惟高・山内・柳本は夏休みのインターン先を探しています！！！