ユーザ生成コンテンツの高品質な自動翻訳に向けた言語現象の体系的分析

藤井 諒[1]，三田 雅人[2,1]，阿部 香央莉[1]，塙 一晃[2,1]，森下 睦[3]，鈴木 潤[1,2]，乾 健太郎[1,2]
1. 東北大学　2. 理研AIP　3. NTTコミュニケーション科学基礎研究所

概要

- 言語現象に注目したNMT評価
- 現象ラベルの定義と対照データセットの作成
- 現象「かな表記」は学習データの増加で解決されない

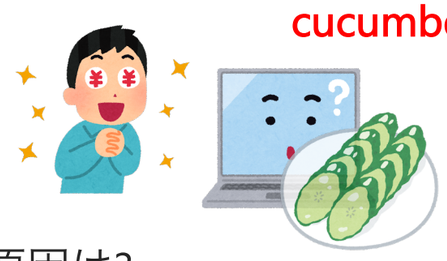
・研究背景/目的

- ユーザ生成コンテンツ (UGC) の増加
  - ➢ 例 : Twitter, ブログ, レビュー

非標準的テキスト処理の重要性 ↑

- UGCにおけるNMTの精度低下

きゅうりょう



Q . 精度低下の原因は?

➡ 言語現象の側面から分析

・手法

1. 言語現象ラベルを定義

MTNTデータセット [Michel+, '18] train/dev から無作為抽出したサンプルの観察

➡ **13種類**の言語現象ラベルを策定

現象ラベル	分類例 (一部)
固有名詞	安倍首相、アナと雪の女王
名詞の省略	マスコミ、JK
スラング	すこ、w
かな表記	とうぜん、キモチワルイ
非正規形	かなちい、だろー

2. 既存データセットを分類

test/blind test の計1895文*に対してクラウドソーシングを用いて定義したラベルを付与

* 不適切語を含む文を除去するフィルタリング適用後

3. 対照データセットで評価

2. の該当表現を正規化した「対照データセット」を作成
言語現象毎の感受性を元文との**BLEUの差分**で測定

- 固有名詞：「某」+ 上位語に言い換え (2言語タスク)

渋谷 Shibuya ➡ **某所 a certain place**

- 固有名詞以外：日本語側のみ辞書的表記に正規化

アプデ ➡ **アップデート**

・実験設定

対照データセット文数 : 290/97/77/72
(固有名詞/名詞の省略/かな表記/非正規形)

モデル : Transformer base [Vaswani+, '17]

- constrained (3.9M) : WMT2019 robustness task [Li+, 2019] のデータ
- unconstrained (12.2M) : +JParaCrawl [Morishita+, '19]

評価 : BLEU [Papineni+, '02] (元文との**差分**)

・結果 / 考察

- 固有名詞 : 学習データ規模 大 ➡ スコア差 小

評価データより新しい学習データの存在による過小評価の可能性

- かな表記 : 差分が大きく学習データ規模による改善 ✗

学習データ規模に依拠せず**特別な対処の必要性**を示唆

(元文/正規化後 のBLEU)

	constrained	unconstrained
固有名詞	+3.9 (14.6/18.5)	+1.2 (17.1/18.3)
名詞の省略	+0.5 (12.2/12.7)	-0.7 (14.0/13.3)
かな表記	+2.4 (12.5/14.9)	+2.8 (12.0/14.8)
非正規形	+0.7 (12.2/12.9)	+1.1 (11.1/12.2)