マルチソース入力を用いた多言語文符号化器による感情分析

[S3-P09]

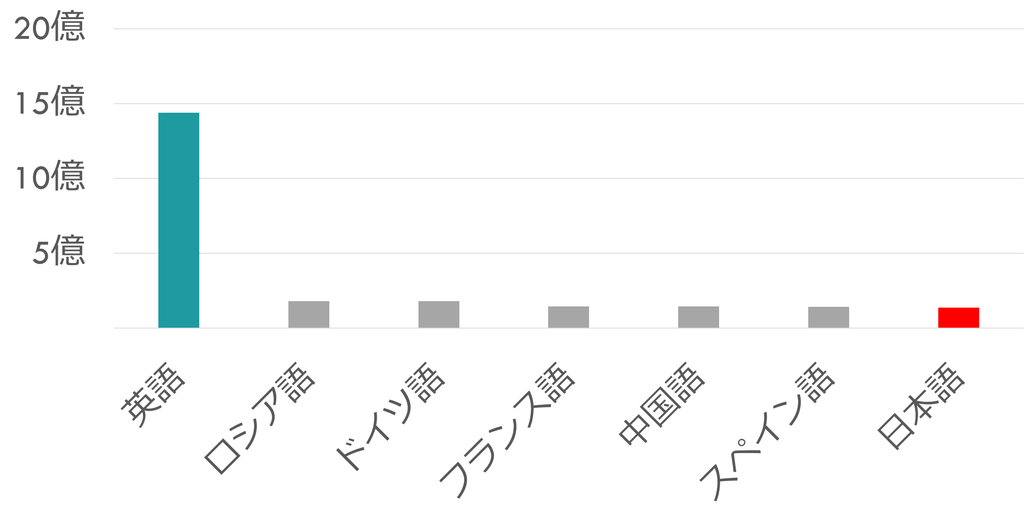
梶川 怜恩, 梶原 智之（愛媛大学）

目的：多言語モデルにとって処理しやすい文に変換し、英語以外の性能を改善

課題：英語以外の言語において、多言語モデルの事前訓練データ量が少ない

XLM-Rの事前訓練**データ**（Common Crawl）の分布

（単位：ページ）



事前訓練データの分布より

英語以外の言語の訓練データの量が少ない

また

文法構造が異なる言語間では多言語モデルの性能が低下する[1]

日本語のデータ割合は4.4%しかなく
英語と日本語は文法構造が大きく異なるため
日本語における多言語モデルの性能が課題となる

⇩

任意の原文を多言語モデルが処理しやすい文に変換したい

[1] Pires et al. “How Multilingual is Multilingual BERT?” In Proc. of ACL, pp. 4996–5001, 2019.

提案手法：原文と機械翻訳による高頻度な表現を併用するマルチソース入力

（仮説）**事前訓練のコーパス中に頻出する表現は高精度に解析可能になる**

（提案手法）**原文とより高頻度な表現を併用するマルチソース入力を多言語モデルに適用**

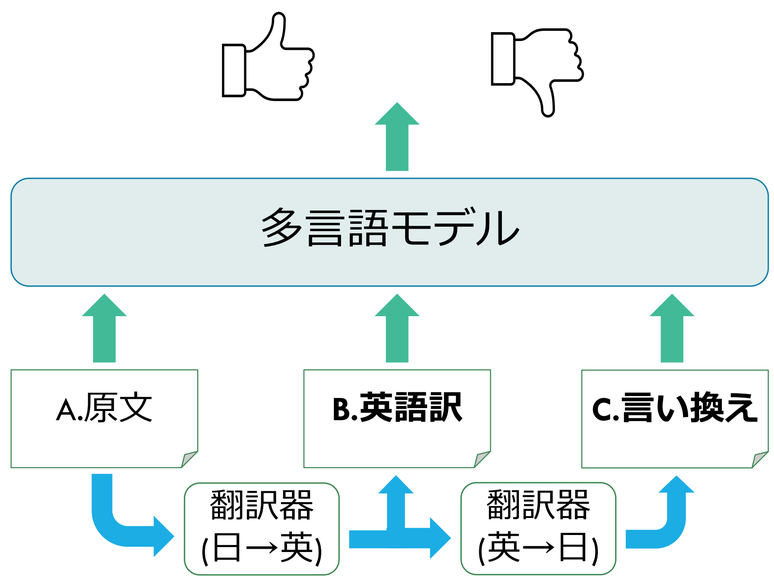
機械翻訳によってテキストの個性が消え、一般的な表現になる[2]

→ 一般的な表現にすることで、コーパス中における高頻度な表現を獲得

機械翻訳に基づく２種類の方法で原文をより高頻度な単語に変換

1. **英語への機械翻訳**
   原文を日英翻訳し、マルチソース入力「“原文 [SEP] 英語訳”」
   事前訓練データにおいて、英語の表現は多く含まれるため、処理しやすい文に

2. **対象言語への折り返し翻訳**
   原文を折り返し翻訳し、マルチソース入力「“原文 [SEP] 言い換え”」
   折り返し翻訳による**言い換え生成によって、高頻度な表現が増加**



[2] Ella Rabinovich et al. “Personalized Machine Translation: Preserving Original Author Traits” In Proc. of EACL, pp. 1074–1084, 2017.

評価実験：日本語の感情分析においてマルチソース入力の有効性を確認

WRIMEデータセット[3]を用いた日本語の感情極性分類（５段階評価）

訓練：30,000文　検証：2,500文　評価：2,500文　評価指標：QWK（正解ラベルとの一致率）

翻訳文との併用によって性能が改善

英語訳を英語モデルに入力するより提案手法の方が性能が高い

多言語モデル	A.原文	B.英語訳	C.言い換え	A+B	A+C	A+B+C	英語モデル	B
DistillmBERT	0.418	0.391	0.383	**0.424**	**0.420**	**0.445**	DistillBERT	0.437
mBERT	0.423	0.363	0.410	**0.484**	**0.459**	**0.466**	BERT	0.497
XLM-R	0.534	0.467	0.450	**0.546**	**0.536**	**0.549**	RoBERTa	0.494
XLM-R (large)	0.591	0.487	0.465	**0.597**	**0.596**	**0.598**	RoBERTa (large)	0.530

[3] https://github.com/ids-cv/wrime

分析と展望：性能改善が機械翻訳による単語頻度の増加に起因するか調査

文ごとの平均単語頻度・未知語数を分析

Wikipediaを用いてmBERT tokenizerによる単語頻度を算出

1. **「高頻度な表現は高精度に解析可能」は真か？**
   高頻度な表現や未知語を含まない文はより高い性能
   → 多言語モデルは高頻度な表現に対して高性能を発揮
2. **「機械翻訳によって高頻度な表現になる」は真か？**
   翻訳文は平均単語頻度が上昇・未知語の減少を確認
   → 機械翻訳により高頻度な表現の変換に成功

〈今後の展望〉

様々な言語やタスクに対して汎用的な手法

言語　：英語と文法構造が近い言語にも有効か？

タスク：JGLUEなどのベンチマークによる調査

原文（検証データ）に対するmBERTの性能			
平均単語頻度の上位500件	**0.358**	未知語を含まない文	**0.430**
平均単語頻度の下位500件	-0.065	未知語を含む文	0.344

	１文あたりの平均単語頻度	頻度が増えた文の割合	未知語数	未知語を含まない文の割合
原文	1.6 M	-	724	0.840
英語訳	**13.0 M**	0.961	**66**	0.994
言い換え	**2.7 M**	0.773	**70**	0.998