単語ベースモデル及び文字ベースモデルの固有表現抽出の性能比較

加藤拓真[1,*] 宮脇峻平[1], 阿部香央莉[1], 大内啓樹[2,1] ,鈴木潤[1,2], 乾健太郎[1,2]

*takuma.kato@ecei.tohoku.ac.jp [1]東北大学 [2] 理化学研究所

概要

- 単語ベースおよび文字ベースモデルで日本語の固有表現抽出 (NER) を行い性能を比較
- 単語ベース・文字ベースで，それぞれ解けた・解けなかった固有表現を分析

背景

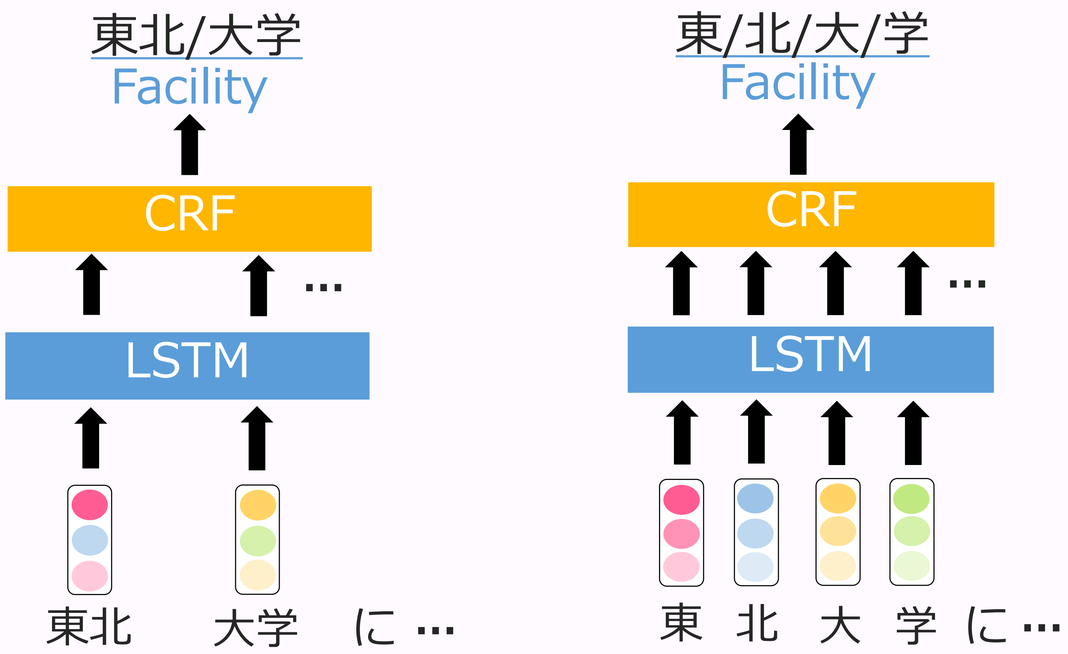
- 現在の日本語NERは単語ベースが主流
- 中国語では，文字ベースで性能が向上したタスクがある[1]

日本語でも文字ベースで性能が上がるのか?

[1]:Yuxian Meng, 2019

提案手法：日本語NERモデル

ベースライン：ニューラルNERモデル[2]



- 比較実験のため、固有表現と単語区切りが一致した（単語ベースに合わせた）データを使用
- word2vecで学習したベクトルを使用

[2]:Jie Yang, 2018

文字ベースNERのメリット

1.形態素解析の必要がない

単語ベース

東北大学に行く ➡ 東北/大学/に/行く ➡ 東北/大学/に/行く
Facility O O

文字ベース

東北大学に行く ☺ 東/北/大/学/に/行/く
Facility O O O

2.語彙サイズの削減

単語ベース：約200万
文字ベース：約1万

約1/200 ☺

3.単語ベースでは解けない問題を解ける

単語ベース

トランプが来日した ➡ トランプ 来日
Person O

文字ベース

トランプが来日した ➡ トランプ 来 日 ☺
Person O Country

実験

データ

拡張固有表現コーパス[3]を使用

	文数	固有表現数
訓練	34784	72318
開発	7009	11954
評価	6783	11669

固有表現ラベル数：28

実験設定

次元数	512
Optimizer	Adam[4]
Dropout率	0.25
語彙数（単語）*	336666
語彙数（文字）	15905

* 頻度10で足切り

[3]:橋本泰一, 2008 [4]:Diederik P. Kingma, 2015

結果・分析

① 単語ベースと文字ベースのF値

	開発	評価
単語ベース	66.69	68.13
文字ベース	62.29	62.73

文字ベースが単語ベースに届かなかった要因は？

- 文が単語ベースよりも長く，予測する文字数が多くなったため？
- 単語ベースに合わせたデータを使用したため？

② ラベル正解数の比較 (○：正解，×：不正解)

		単語ベース ○	単語ベース ×
文字ベース	○	6124	882
	×	1400	3548

例1. 単語ベース：○，文字ベース：×

正解 ：災害対策基本法
文字ベース：災害対策基本法上

「上」にも余計なラベルが付いた

例2. 単語ベース：×，文字ベース：○

正解 ：北東アジア
Location
単語ベース：北東 アジア
O Location

「アジア」にのみタグが付いた

今後の方針

- ELMoやBERTなどより良い言語モデルを文字ベクトルとして用いて改善していきたい
- 文字ベースと単語ベースを組み合わせた場合でも実験を行いたい