Q6-9 Majority or Minority: 固有表現認識におけるデータの不均衡性を考慮した損失関数の提案

根本颯汰[1]，北田俊輔[2]，彌冨仁[1]　[1]法政大学 理工学部　[2]法政大学 大学院 理工学研究科

法政大学 HOSEI University

Summary

固有表現認識におけるデータの不均衡性を考慮した新しい損失関数の提案

- 多数派クラスのみを考慮し，少数派クラスから多数派クラスへの誤分類を抑制が可能
- 言語非依存で既存の不均衡性に対処する損失関数より性能向上

Background

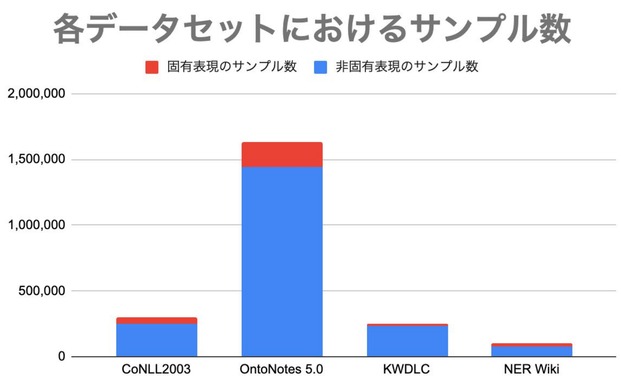
固有表現認識 (Named Entity Recognition; NER)

- テキストから人名や組織名を取り出すビジネス面で多くの応用がなされている情報抽出技術の一つ

法政大学 [組織名] は 東小金井 [地名] にある。

NERのデータセットの8割近くは抽出対象外の非固有表現の単語

➜ 😰 不均衡性により単語数の少ない固有表現クラスの性能低下



従来は損失関数を適切に設計して不均衡性に対処する手法が主流

➜ 😰 二値分類問題へ帰着させることで不均衡に対処

Majority or Minority (MoM) loss

従来のモデルの損失 $\mathcal{L}$ に多数派クラスのみを対象とした損失 $\mathcal{L}_{\mathrm{MoM}}$ を追加

➜ 😊 少数派クラスの多数派クラスへの誤分類を抑制

INPUT

	テキスト	法政	大学	は	小金井	にある
	T (GT)	B-Org	I-Org	O	B-Loc	O
$\mathcal{L}(T,P)$	T (GT)	B-Org	I-Org	O	B-Loc	O
	P(Predict)	B-Org	I-Org	I-Org	O	O
$\mathcal{L}_{\mathrm{MoM}}(T,P)$	T (GT)	B-Org	I-Org	O	B-Loc	O
	P(Predict)	B-Org	I-Org	I-Org	O	O

多数派である非固有表現の O クラス

loss の計算対象

誤分類した O クラスのトークンに対して損失大

（全体）モデルの損失

（既存）全クラスの多クラス分類

（提案）多数派クラスと少数派クラスの二値分類

$$\mathcal{L}_{\mathrm{total}}(T,P) = \lambda \cdot \mathcal{L}(T,P) + (1-\lambda) \cdot \mathcal{L}_{\mathrm{MoM}}(T,P)$$

Experiments

データセット (2言語計4データセット)

- 🇺🇸 CoNLL 2003 [Sang+ NAACL'03] ニュース記事データセット
- 🇺🇸 OntoNotes5.0 [Pradhan+ EMNLP'12] 16個の固有表現のコーパス
- 🇯🇵 京大ウェブ文書リードコーパス (KWDLC)
- 🇯🇵 Stockmark-NER-wiki (NER-Wiki)

ベースモデル

BERT [Devlin+ NAACL'19]

BERT-MRC [Li+ ACL'20]
➔ 機械読解問題として解くモデル

損失関数

Focal loss (FL) [Lin+ ICCV'17]

Dice loss (DL) [Li+ ACL'20]

MoM loss (提案手法)

Results

4つのデータセットを用いたBERTでの性能比較

	🇺🇸 CoNLL03			🇺🇸 OntoNotes			🇯🇵 KWDLC			🇯🇵 NER-Wiki		
BERT	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
w/ CE	90.16	91.86	91.00	87.41	89.07	88.23	70.92	73.96	72.41	77.32	81.04	79.13
w/ FL	90.33	92.03	91.17	**87.62**	89.15	88.39	71.88	**74.27**	73.05	77.79	81.53	79.61
w/ MoM	**90.41**	**92.27**	**91.33**	87.39	**89.84**	**88.60**	**72.54**	74.13	**73.32**	**78.13**	**81.61**	**79.83**

CoNLL03を用いたBERT-MRCでの性能比較

	🇺🇸 CoNLL03		
BERT-MRC	Prec.	Rec.	F1
w/ BCE	92.47	92.19	92.33
w/ FL	92.81	92.17	92.49
w/ DL	92.59	92.47	92.53
w/ MoM	**93.09**	**92.57**	**92.83**

😊 すべてのデータセットにおいて BERT と BERT-MRC の両方で既存の不均衡性に対処する損失関数よりも提案手法である MoM loss が更に予測性能向上に貢献した

CoNLL03での固有表現ごとの性能結果

	BERT w/ MoM			BERT			
	Prec.	Rec.	F1	Prec.	Rec.	F1	test tokens
LOC	93.43	93.24	**93.38**	93.02	93.65	93.34	1,922
MISC	79.37	84.64	**81.92**	78.87	83.77	81.25	918
ORG	89.82	93.63	**91.68**	90.30	92.87	91.57	2,496
PER	97.07	98.16	97.61	97.62	97.87	**97.75**	2,769
O	99.75	99.25	**99.50**	99.68	99.30	99.49	38,312

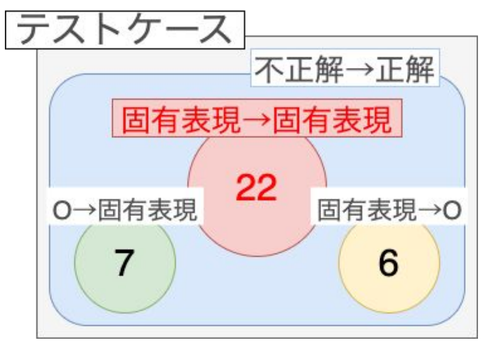
😊 LOC (位置) や MISC (その他)，ORG (組織)，Oクラスは性能向上
😰 PER (人) クラスは性能低下

Discussion

PERSON クラスが性能低下した理由について

➜ CoNLL03のテストケース 3,453 件のうち35件が正解に変化

➜ 22件全て，よりサンプル数の多い固有表現クラスへ変更
ex) ORGANIZATION→PERSON



MoM loss 導入による正解例

テキスト	法政	大学	は	小金井	に	ある
CE Pred	B-Loc	I-Loc	O	B-Loc	O	O
MoM Pred	B-Org	I-Org	O	B-Loc	O	O

MoM loss 導入による影響

1. 少数クラスからOクラスへの過剰な偏りが減少
   - 😊 少数クラス間のサンプル数の分布により忠実
2. サンプル数の多いクラスに変化
   - 😰 少数クラス内で最も多い人名 (PERSON) に偏る
     - 上記のPERSONの性能低下の原因だと考えられる

Conclusion & Future Work

1. 多数派クラスと少数派クラスに分けて多数派の損失のみを計算し不均衡性に対処する損失関数の提案
   - 既存の不均衡に対処する損失関数よりも予測性能が向上
   - 分類するクラス数や言語を問わない予測性能の向上
2. 他分野のデータの不均衡性に対する応用