PT4-17　Masked Language Model による系列確率に基づく文法誤り検出

○土肥 康輔，須藤 克仁，中村 哲 (NAIST)

本研究の概要

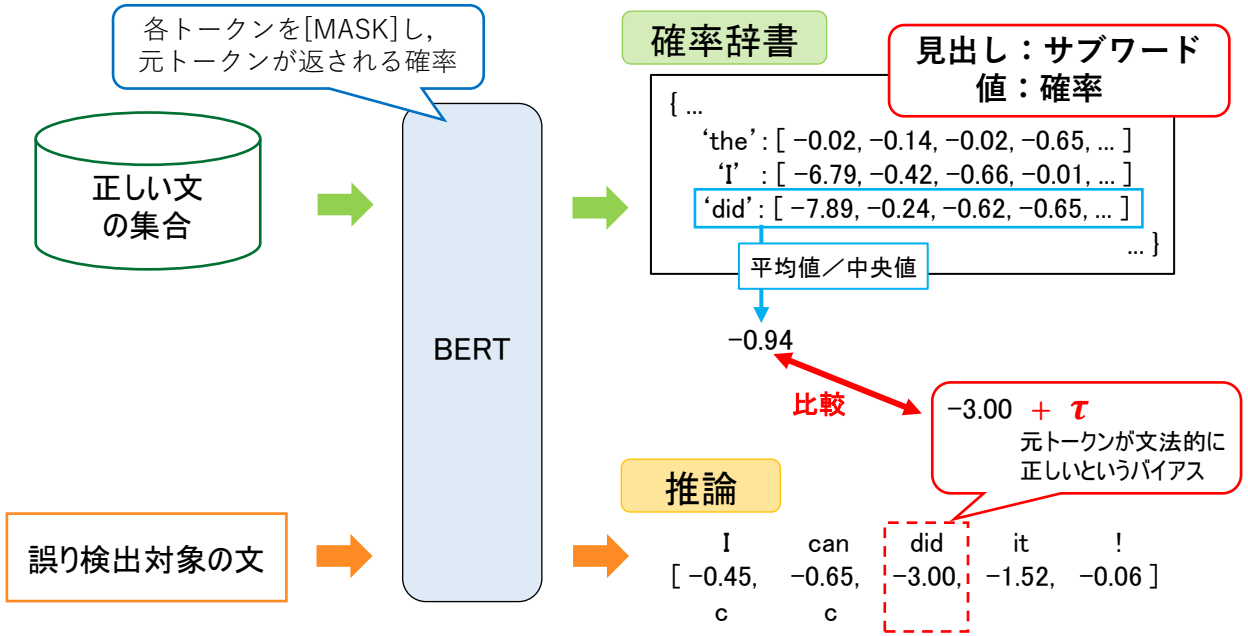
BERTの確率を利用した**文法誤り検出**

→ **ラベル付きデータなしで，ナイーブなLSTMをやや上回った**

提案手法

- 確率の低い語は，誤っているだろう

ある文での“did”の確率を，“did”全体の確率と比較　→　**[確率辞書]**



- **サブワードと元トークンの対応付け**

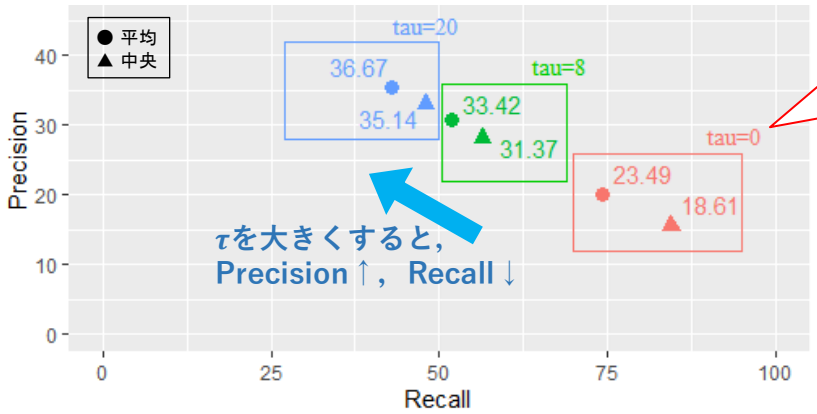
すべて“c”のときのみ“c”

le	##mma	##tized		lemmatized
[c	c	c]	→	[c]
[c	i	c]	→	[i]

実験

確率辞書作成方法，τの違いによって，誤り検出性能に差があるか？

τ，確率辞書の基準値の比較



τが同じ場合，確率辞書の基準値を「平均」とするほうが$F_{0.5}$値が高い

<BERT>
bert-base-cased (入力の最大長 = 250)

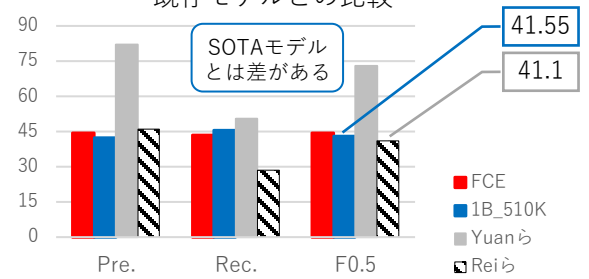
<データ>
母語話者: the One Billion Word Benchmark (1B)
学習者: W&I+L, FCE, NUCLE, CoNLL-2013, 2014

→ **Recallが高くなる「中央」からスタートし，適切な制約を加えるほうがよい**

確率辞書間の比較

→ **母語話者コーパスで◎**

既存モデルとの比較



- Precisionに課題
- **ラベル付きデータなしで，ナイーブなLSTMをやや上回った**

今後の課題

- 文脈情報を利用した制約
- 文脈，品詞等を考慮した確率辞書作成

【参考文献】
[Yuan+ 2021] Multi-class grammatical error detection for correction: A tale of two systems. In *Proc. of EMNLP*.
[Rei+ 2016] Compositional sequence labeling models for error detection in learner writing. In *Proc. of ACL*.