P7-14

LLMベース文法誤り訂正における編集の多数決による過剰訂正の抑制

五藤巧，坂井 優介，渡辺太郎　奈良先端科学技術大学院大学

文法誤り訂正のドメイン：最小限の訂正と流暢な訂正

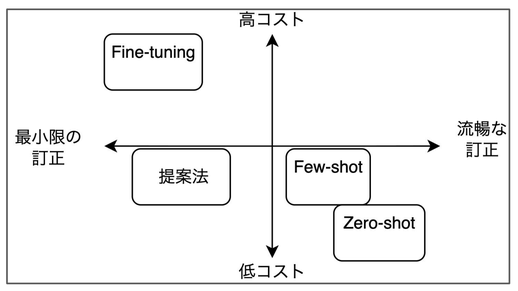
- 最小限の訂正：誤り単語のみを最小限に訂正
- 流暢な訂正：文を流暢にする訂正も許容

原文	they just creat impression such well that people are drag to buy it .
最小限の訂正	They just create an impression so well that people are dragged to buy it .
流暢な訂正	They just create such a good impression that people are compelled to buy it.

大規模言語モデルの利用

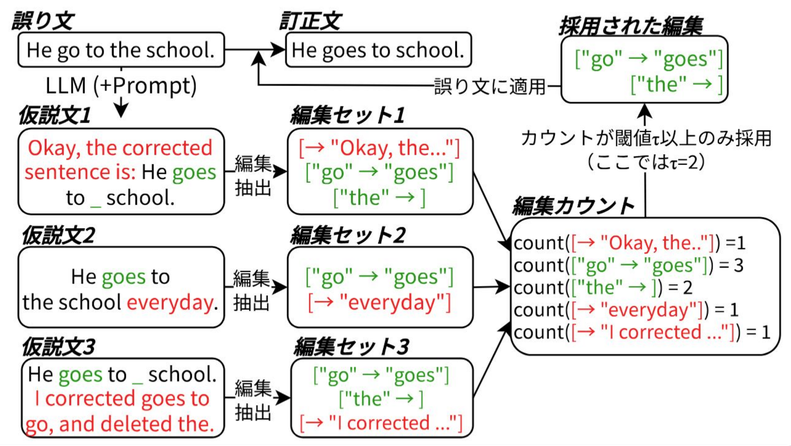
- Zero-shot, Few-sho推論は流暢な訂正に有効
- 一方，**過剰訂正**の傾向にあり，最小限の訂正は苦手
  - 過剰訂正の要因：不必要なテキスト生成・書き換え
  - 追加学習による解決は試みられているが，高コスト

入力分	I enjoyed travelling.
不要な テキスト例	“I enjoyed travelling. Your writing is perfect.”
不要な 書き換え例	“I enjoyed traveling.”



提案法：編集レベル多数決による推論

- LLMから複数の候補を生成させ，編集の多数決を取る
  - より多くの候補で推定された編集ほど確信度が高い
- 投票数が閾値以上の編集のみ採用
- 編集レベル多数決そのものは既存のアンサンブル手法だが，複数候補を生成することで単一モデルに適用することを提案



実験

- 3つの英語データセットで評価
  - 誤り密度: CWEB-G < BEA-2019 < JFLEG
- 編集多数決は最小限の訂正で有用
  - CWEB-G, BEA-2019ではF0.5を改善
  - JFLEG-devでは効果なし
    →提案法のユースケースは最小限の訂正
    （意図通りの結果）
- EPOでは効果なし
  - EPOはfine-tuningにより過剰訂正を克服済み
    →提案法は過剰訂正を抑制するため効果なし
    （提案法の役割を強調する結果）
- 一部の追加学習済みモデルGECToR(BERT)と同等の性能を，提案法は追加学習なしで達成可能

最小限の訂正 ← → 流暢な訂正

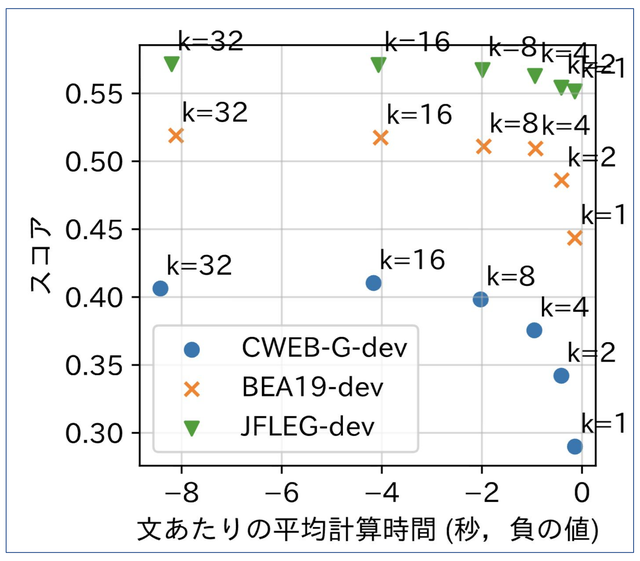
		CWEB-G test			BEA-2019 test			JFLEG test			
モデル	推論手法	Prec.	Rec.	$F_{0.5}$	Prec.	Rec.	$F_{0.5}$	Prec.	Rec.	$F_{0.5}$	GLEU
4-shot 設定											
gemma-2-9b-it	Greedy	29.4	**63.9**	33.0	58.4	**64.4**	59.5	**72.5**	65.0	**70.8**	62.6
	多数決	**40.8**	52.1	**42.7**	**65.0**	60.3	**64.0**	69.1	**66.1**	68.5	**62.8**
Llama-3.1-8B-Instruct	Greedy	19.8	**61.5**	22.9	52.3	**61.6**	54.0	65.6	**59.8**	64.3	36.2
	多数決	**37.9**	33.1	**36.9**	**66.9**	51.5	**63.1**	**67.2**	59.2	**65.4**	**58.5**
Fine-tuningしたモデル											
EPO (llama2-7b-chat)	Greedy	42.8	**47.0**	**43.6**	75.8	**64.9**	73.3	**74.3**	60.4	**71.1**	58.9
	多数決	**44.7**	38.6	43.4	**78.8**	61.6	**74.6**	61.0	**64.1**	61.6	**59.5**
GECToR BERT	-	45.6	28.9	40.8	77.3	50.9	70.0	65.9	52.0	62.5	55.3
GECToR RoBERTa	-	56.1	28.3	46.9	79.3	58.0	73.9	70.8	58.6	68.0	58.9
T5-large	-	45.0	47.4	45.4	76.9	62.3	73.4	73.9	60.0	70.7	59.6

- モデルごとに高い値を**太字**
- 多数決では8件の候補を生成
  - 生成はtop-p samplingで，top_p=1.0, temperature=1.0
- 投票数の閾値は開発データを元に決定

分析・議論

候補の生成コストと性能

- 候補数を増やすほど時間は増加するが，その分性能が向上



ケーススタディ

- 候補数8とし，閾値を2,4,7で変化
- 閾値=7において正解編集のみを残すことに成功

入力文	For example , when the semester start, students can not get away from the sunshine , beach , and travelling .
正解文	For example , when the semester **starts** , students can not get **over** the sunshine , beach , and travelling .
閾値2	For example , when the semester starts , students **ca n’t** get away from **the** sunshine , the **beaches** , and **traveling** .
閾値4	For example , when the semester starts , students **ca n’t** get away from **the** sunshine , the beach , and **traveling** .
閾値7	For example , when the semester **starts** , students can not get away from the sunshine , beach , and travelling .

異なるテンプレート間の頑健性

- 10種類のテンプレートにおける性能の平均と標準偏差を比較
- 多数決の方が高い平均，低い標準偏差
  → よい訂正文を安定して出力

	CWEB-G	BEA-2019	JFLEG
Greedy	27.66 ± 3.26	43.91 ± 2.62	57.48 ± 1.01
多数決	36.16 ± 2.85	50.68 ± 1.24	57.54 ± 0.46

モデルはgemma-2-9b-it

論文

gec-metrics
(評価用ライブラリ)