P5-18

ニューラル文法誤り訂正のための多様な規則を用いる人工誤り生成

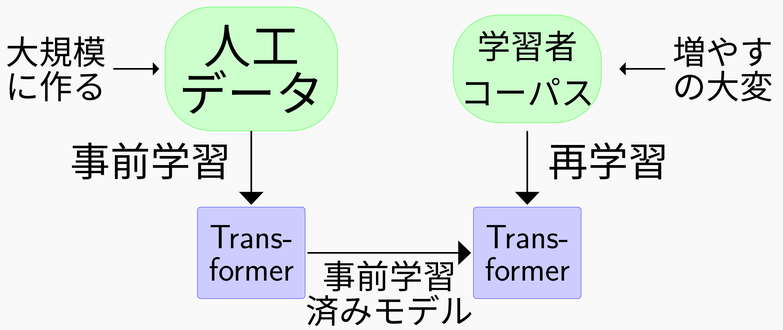
古山翔太　高村大也　岡崎直観　（東工大/産総研）
(東工大 岡崎研 修士１年)
東京工業大学 Tokyo Institute of Technology　産総研

概要

- 文法誤り訂正の事前学習のデータをルールベースで人工的に生成
- 多様なルールの組合せが効果的

背景

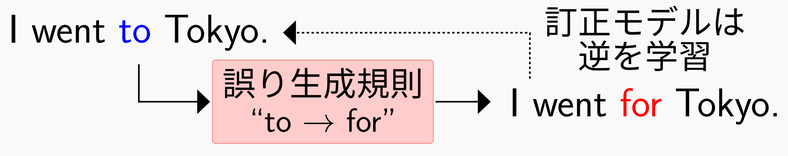
ニューラル文法誤り訂正のパラダイム「事前学習＋再学習」



大規模な **人工データ**による**事前学習**
＋ **学習者コーパス**による**再学習**が主流

- “人工データ” をどのように作るかが重要

人工誤り生成とは？　その課題は？

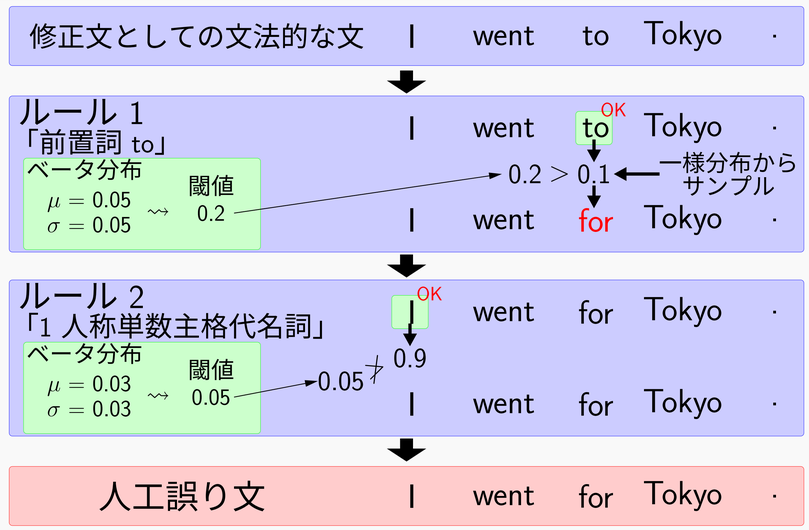


文法誤り訂正の学習データを作るために，人工的に誤りデータを生成すること

- 本研究ではルールベースで誤り生成
  - 少ないルールでは誤り文が偏る...
- 多くのルールを用いて良い誤り文を偏りなく，多様に生成したい！
- 人工誤りデータの質が向上し，文法誤り訂正の性能向上が期待

手法

どのように多様な誤り文を生成するか？



ルールを順次，独立に適用
文ごとに誤り確率を変え多様な誤り文を生成
誤り生成規則は全 **5** カテゴリ・**188** 種類 [a]
（機能語・活用・語順・単語選択・表記体系）

結果

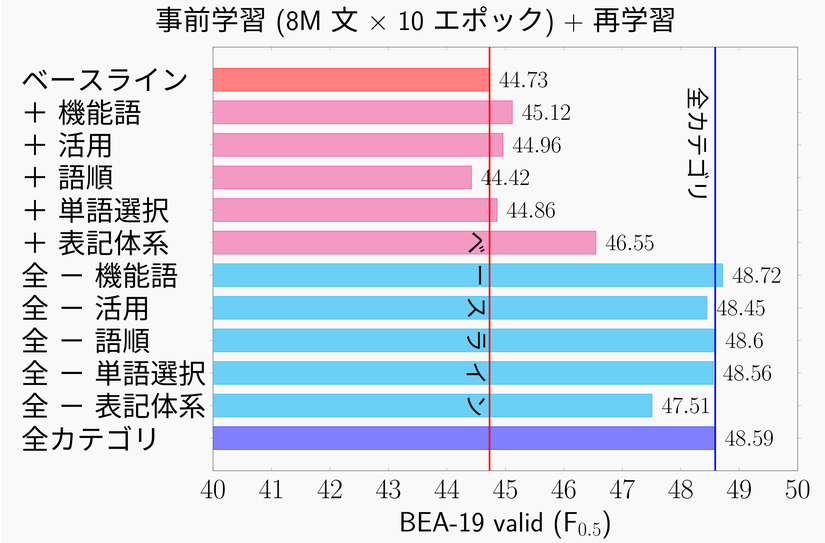
モデル: Transformer big
単言語コーパス: 1 億 2400 万文 (WMT ニュースタスク)
学習者コーパス: 62 万文

	BEA-19 valid ($F_{0.5}$)	BEA-19 test ($F_{0.5}$)	CoNLL 14 ($F_{0.5}$)	JFLEG (GLEU)
ベースライン	44.73	-	53.88	57.37
事前学習	41.79	-	54.92	58.35
＋ 再学習	52.83	-	63.25	62.23
＋ アンサンブル ＋ リスコア	54.68	72.51	65.43 [b]	63.69
＋ 再学習 ＋ ドメイン適応 ＋ アンサンブル ＋ リスコア	56.49	72.76	-	-
Grundkiewicz+ 19 (ルールベース誤り生成)	53.00	69.47	64.16	61.22
Omelianchuk+ 20 (タグ付けモデル)	-	**73.7**	66.5	-
Lichtarge+ 20 (折返し翻訳の誤り生成)	-	73.0	**66.8**	**64.9**

既存のルールベース誤り生成よりは良い
タグ付モデル・MT ベース誤り生成に及ばず

議論

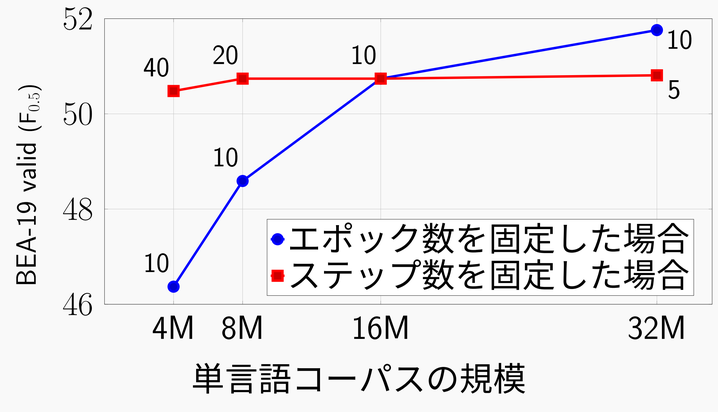
誤り生成規則の多様さは重要か？



1 カテゴリのみの誤り生成は効果が薄い
1 カテゴリを除いても，性能低下しにくい

- 多様な誤り生成の統合は効果的

単言語コーパスの規模は重要か？



単言語コーパスを 4M,8M,16M,32M と変え
エポック数固定 (10) と40→20→10→5で比較

- 単言語コーパスの規模が大きければいいというわけではない
- 低資源の言語にも良いルールを用いれば高性能なモデルを構築可能？

[a] 誤り生成規則は実験に用いたソースコードとともに公開しています．(github.com/nymwa/arteraro)
[b] CoNLL-14 に関する論文中の値に誤りがありましたので，当ポスター，修正原稿 (リンク) では訂正した値を掲載しました．