「間違い~~が~~直す」格助詞誤り訂正システム
を

長岡技術科学大学 自然言語処理研究室 小川耀一朗・山本和英

①目的

文法誤り訂正タスクとは

文章に含まれるあらゆる文法的な誤りを自動で訂正するタスク

間違いが直す → 間違いを直す (格助詞)
マイク内臓 → マイク内蔵 (同音異義語)
結果を表示される → 結果を表示する (受動態/能動態)

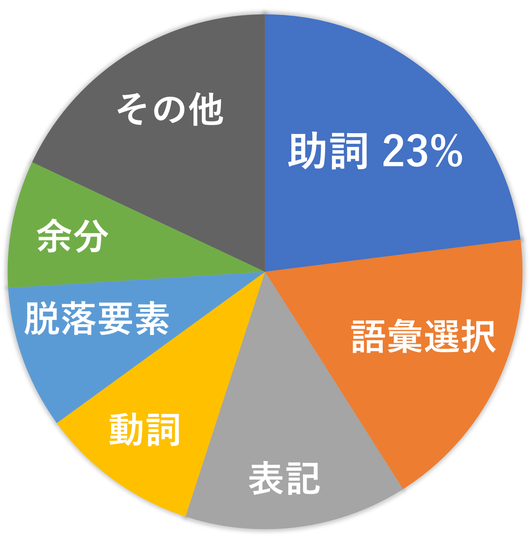
日本語の誤り訂正ツールがない

英語だと Grammarly, Gingerなどがある

自然言語処理の教育分野への応用

- ✓ 日本語教師の作文チェック作業の効率化
- ✓ eラーニングでの日本語学習支援
- ✓ EdTechが国政に

日本語学習者の誤り傾向



NAIST誤用コーパスでの誤り傾向

NAIST誤用コーパスでの
助詞誤りの内訳

✓ 日本語学習者にとって**助詞が最も間違えやすい**
✓ 特に格助詞の誤用と助詞不足が多い

→ 格助詞「が, を, に, で」及びそれらの不足の誤りを対象にした訂正システムを構築する

③課題

広い文脈情報が必要

この中には硫黄や硫化化合物をあります　予測:に, 正解:が

→ 「この中には」を4-gramでは捉えきれない

人々はこの規制を賛同の意志を表明しました 予測:が, 正解:に

→ 文中の助詞との兼ね合いを考慮できていない

日本語学習者は平仮名を多用する

さいきん,カンボジアがおおきなこうずいがありました 予測:に,正解:で

→ 訓練コーパスに出現しないため予測が難しい

②訂正方法と結果

言語モデルとは

ある単語列の次に来る単語の自然さを確率で表すモデル

自転車に { **乗る：0.9** ← 自然 / **買う：0.1** ← 不自然

$$P(\text{自転車, に, 乗る}) = \frac{(\text{自転車, に, 乗る})\text{の頻度}}{(\text{自転車, に})\text{の頻度}}$$

言語モデル確率を用いた格助詞誤りの訂正

私 は 自転車 を 乗る

が：0.1	
を：0.03	
に：0.7	文全体の確率が最も高くなる格助詞に置換
で：0.02	

文中の全ての格助詞「が, を, に, で」に対して順番にチェックを行う

言語モデル確率を用いた格助詞の補完

私 は 自転車_乗る

が：0.1	
を：0.03	
に：0.7	文全体の確率が最も高くなる格助詞を挿入
で：0.02	
不要：0.01	

名詞|代名詞|接尾辞(名詞的)の後の単語が助詞|助動詞でなければチェックを行う

実験設定と結果

言語モデル：　単語4-gram言語モデル
訓練データ：　日本経済新聞記事コーパス（約1800万文）
テストデータ：NAIST誤用コーパスから対象とする誤りを含む文とその訂正文を抽出(879対)

手法	適合率	再現率	F値
ランダム	6.56%	44.1%	11.4%
頻度	7.43%	49.2%	12.9%
提案手法	**50.9%**	**60.3%**	**55.2%**
提案手法 (置換のみ)	49.3%	69.3%	57.6%
提案手法 (補完のみ)	58.3%	39.7%	48.0%

デモを公開しています！
http://box.jnlp.org/grammatical-error-checker