Cross-Corpora Evaluation and Analysis of Grammatical Error Correction Models

Is Single-Corpus Evaluation Enough?

Masato Mita¹,² (masato.mita@riken.jp), Tomoya Mizumto¹, Masahiro Kaneko³,⁴, Ryo Nagata³,¹, Kentaro Inui²,¹
¹RIKEN Center for Advanced Intelligence Project (AIP), ²Tohoku University, ³Tokyo Metropolitan University, ⁴Konan University

Ans: No. it’s not enough. GEC systems should be tested over multiple diverse datasets!!

### Systems:
- LSTM
- CNN [Chollampatt et al., 2017]
- Transformer
- SMT [Junczys-Dowmunt et al., 2017]

### Data set:
- Train: Lang-8 + NUCLE (1.3 M)
- Dev: NUCLE (5.4K)

### Analysis1: System Performance by Error Type (CoNLL-2014)
- Each system has different strengths and weaknesses

### Analysis2: Top System by Error Type (Cross-Corpora)
- Each corpus has different tendency errors

### Background
- Input: ungrammatical sentence
- Output: grammatical sentence
- GEC input is expected to have many variations due to the influence of writer’s proficiency

### Issue:
- GEC system performance is still primarily benchmarked against the narrow domain data set ---CoNLL-2014 test set

### Experiments on Cross-Corpora

[https://github.com/tomo-wb/GEC_CCE](https://github.com/tomo-wb/GEC_CCE)