ProQE: Proficiency-wise Quality Estimation dataset for Grammatical Error Correction

Yujin Takahashi♤, Masahiro Kaneko♡, Masato Mita♧,♤, Mamoru Komachi♤

♤ Tokyo Metropolitan University; ♡ Tokyo Institute of Technology; ♧ RIKEN

https://github.com/tmu-nlp/ProQE/

Introduction

- Quality estimation (QE) models can evaluate grammatical error correction (GEC) systems without relying on reference sentences
- Yoshimura+ (2020) showed BERT-based QE model for GEC achieves high correlation with human, but the proficiency of the dataset was limited (advanced)

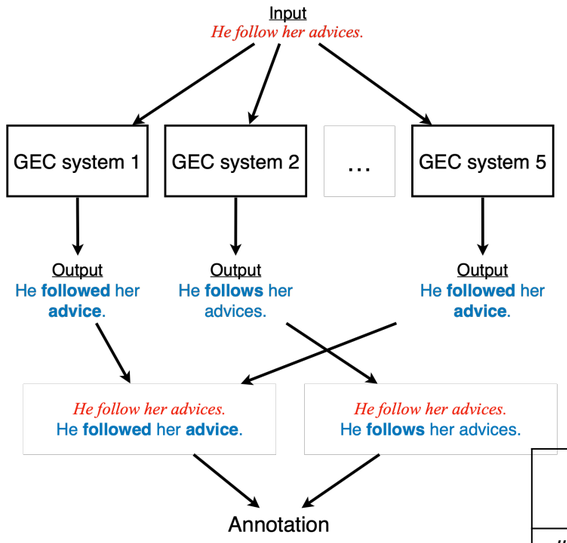
Contribution

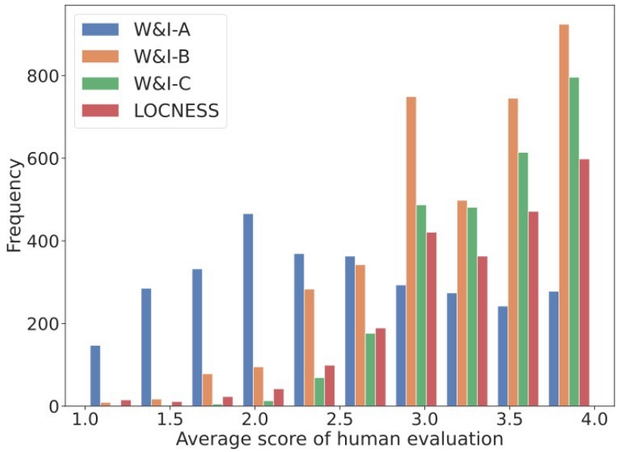
- Created a QE dataset with proficiency-level for GEC to alleviate the problem of imbalanced dataset of QE for GEC (ProQE)
- Investigated the effect of proficiency for QE for GEC and proposed a robust QE model based on the findings

Annotation

1. Get learner-sentences from W&I+LOCNESS (Bryant+, 2019) corpus, which is annotated with proficiency based on CEFR-levels
2. Use a variety of GEC systems to generate system outputs
3. Retain unique pairs of learner-sentences and system-outputs
4. Assign QE scores (0-4) to the pairs by crowdsourcing (Amazon Mechanical Turk)

Flow of the annotation





Statistics of our dataset (ProQE)

	A beginner	B intermediate	C advanced	N native
# sents	3,049	3,740	2,644	**2,233**

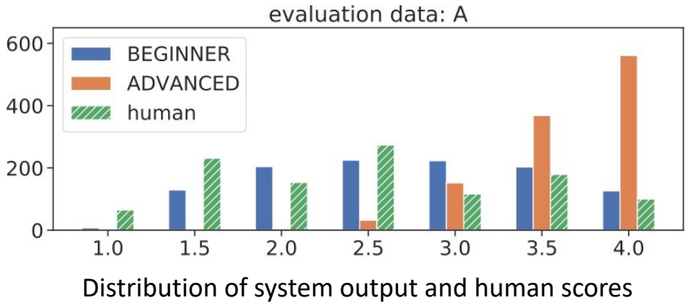
Experiments

Experiment 1: Effect of proficiency for QE

- Proficiency-wise evaluation for BERT-based QE models trained on each data
- 5-fold cross validation with Pearson's coefficient

Model	A	B	C	N
Beginner	**0.70**	0.54	0.51	0.60
Intermediate	0.63	**0.58**	0.52	0.57
Advanced	0.63	0.52	**0.52**	0.58
Native	0.59	0.54	0.49	**0.61**

→ No notable difference except for A



Distribution of system output and human scores

→ Advanced model assigns high scores to ungrammatical input

Experiment 2: Robust model

- Mixed: all the sub-corpora combined
- Mixed+Tag: Mixed + the gold proficiency level as a special token

Model	A	B	C	N
Mixed	0.69	0.60	0.57	0.61
Mixed+Tag	**0.71**	**0.63**	**0.60**	**0.63**