Japanese-English Sentence Translation Exercises Dataset for Automatic Grading

Naoki Miura[1,2], Hiroaki Funayama[1,2], Seiya Kikuchi[1,2],
Yuichiroh Matsubayashi[1,2] ,Yuya Iwase[1,2] ,Kentaro Inui[3,1,2]

[1]Tohoku University [2]RIKEN [3] MBZUAI

Summary

- We aim to **automate a grading of sentence translation exercises (STEs)** for an educational use
- **We formalize the STE tasks, create datasets, and establish baselines**
- We show the performance of finetuned BERT and GPT models and discuss further directions

Sentence Translation Exercises

Background & Motivation

- Utilized as educational tools in the early stages of L2 language learning [Cook, 2010; Butzkamm and Caldwell, 2009]
- In STE, the rubric allows learners to focus on the learning objectives set by the teacher, facilitating efficient learning

Question :

Translate this Japanese (L1) sentence into English.

私は / 一昨年に / オーストラリアで /見るまで / コアラを / 見た / ことがなかった
(I / the year before last / in Australia / before I saw one / a koala / seen / had never)

L2 learner's response

I hadn't seen a koala, before I **saw** **in Australia** two years ago.

Incorrect (O4) Correct (G4) Correct (E3)

Rubric

Chunk	Analytic criteria	2 (Correct)	0 (Incorrect)
"オーストラリアで" (in Australia)	E3	"in Australia"	Otherwise
	…	…	…
"見るまで" (before I saw one)	O4	The word order is "conjunction + SVO"	Incorrect
	G4	Using "saw"	Otherwise

E : Expression, **O** : Word Order, **G** : Grammar

STE Dataset

Contents : questions, graded responses, rubrics

- 3,498 responses for 21 questions, including 196 analytic criteria.

Collecting student responses

- From **high school students** and **cloud workers**

Annotation

- A score for a criterion and an identified specific phrase within a response that serves as a grading clue (as **justification cues**).

Annotation quality (IAA)

- Substantial agreement [Landis and Koch, 1977] for scoring: 0.72 in Cohen's kappa coefficient
- High level of agreement [Sato et al., 2022] for justification cues

Method

- We employ a **BERT** [Devlin et al., 2019]-based classification model and the **GPT models** [OpenAI, 2023] with in-context learning as a baseline
- The models **predict a score for each analytic criterion**
- Given that STE deal with language knowledge, we hypothesize utilizing LLM can show superior performance in the grading STE

Models	Finetuned BERT	GPT with in-context-learning
Input	Response	Response, L1 sentence Rubric, Scoring example
Output	Score and cue	

Result and Discussion

Evaluation measure : F1 (5-fold cross-validation)

Category (#criteria)	BERT			GPT-3.5 (5 shots)		
	Correct	Partially correct	Incorrect	Correct	Partially correct	Incorrect
E : (96)	0.92 ± 0.15	0.64 ± 0.36	0.82 ± 0.24	0.84 ± 0.12	0.79 ± 0.23	0.65 ± 0.18
O : (42)	0.95 ± 0.05	*nan*	0.79 ± 0.25	0.80 ± 0.12	*nan*	0.53 ± 0.21
G : (45)	0.94 ± 0.11	0.81 ± 0.21	0.88 ± 0.13	0.82 ± 0.13	0.48 ± 0.11	0.64 ± 0.28
All	0.93	0.68	0.83	0.83	0.73	0.61

Lower performance for incorrect responses

- GPT-3.5 performs significantly worse than BERT
- **GPT-3.5 struggled to interpret** STEs scoring task
- Several analytic criteria were challenging for both models, as evidenced by the standard deviation
- Increasing the number of scoring examples does not improve the performance (see the paper)

Future Work

- We aim to develop a scoring model that drastically reduces the amount of learning data by leveraging LLM
- We plan to conduct experiments using open-source LLMs
- We consider developing a rubric in a format that is easy for the LLM to interpret.
- **Subdivide scoring tasks** in STEs (" Grammatical Error Correction," "Checking the coherence with L1," "Verifying the use of expressions corresponding to the rubric")