Automatically Suggesting Diverse Example Sentences for L2 Japanese Learners Using Pre-Trained Language Models

Enrico Benedetti[1], Akiko Aizawa[2], and Florian Boudin[2,3]

[1]University of Bologna, Italy [2]National Institute of Informatics, Japan [3]JFLI, CNRS, Nantes University, France

[1]enrico.benedetti5@studio.unibo.it [2]aizawa@nii.ac.jp [3]florian.boudin@univ-nantes.fr

Introduction

Good example sentences can improve vocabulary acquisition when studying a second language. We focus on Japanese. Because it costs time and effort to find or create quality examples, we study how to obtain and score examples automatically, using Pre-trained Language Models.

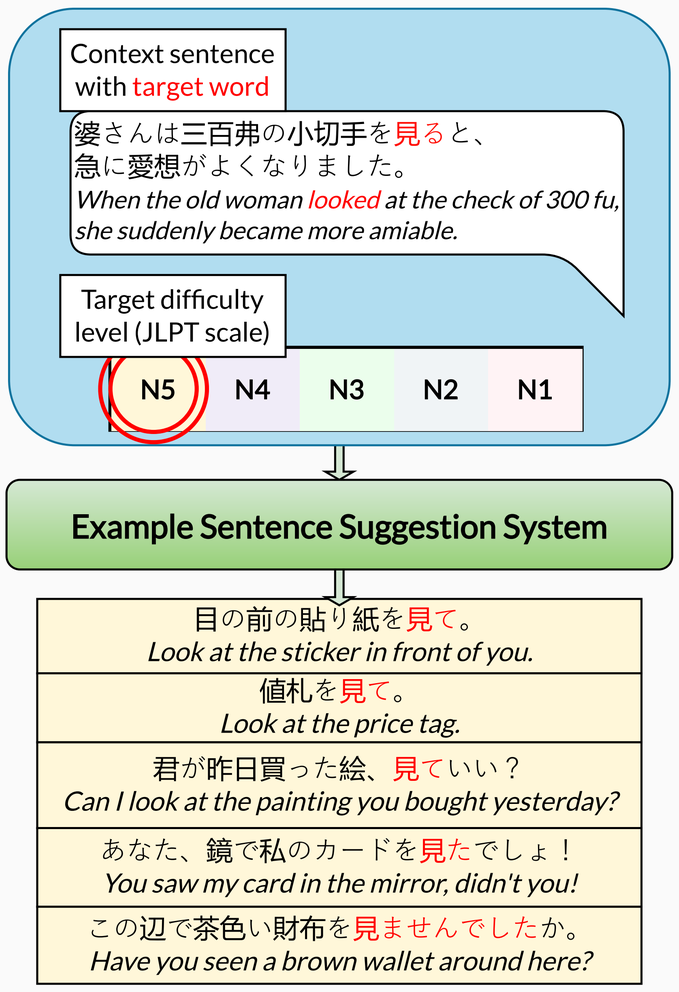

Fig. 1: The task of suggesting good examples for a target word.

What makes a good example sentence?

1. It contextualizes a target word.
2. It matches a specific proficiency level (JLPT N5 to N1).
3. Multiple examples are diverse in vocabulary and syntax.

Contributions and methods

We develop a retrieval-based approach to extract examples from a new corpus of 12M sentences, **WJTSentDil** (**W**ikipedia, **J**pWaC and **T**atoeba **Sent**ences with **Di**fficulty **L**evel).
We combine Pre-trained Language Models and NLP techniques to score examples automatically on multiple characteristics:

1. **similar sense** with word embeddings from MirrorWIC (Liu et al., 2021).
2. **difficulty** with a BERT classifier finetuned on web data.
3. **syntax diversity** with syntax tree overlap (Chen et al., 2023).

Evaluation setup

Context sentence	Target level	Target word	Block ID
また、東西お互いに相手を非難するプロパガンダ放送を流し合っていた。	N1	相手	1
System 1			
Suggested Sentences	**Difficulty rating**	**Sense rating**	**Reject**
東京と大阪はライバル同士であるため、それぞれの地域では互いに相手を非難するプロパガンダ放送を流し合っていた。	N2	Similar	☐
これまでずっと、両国間では互いに相手を非難するプロパガンダ放送を流し合ってきた。	N2	Similar	☐
彼と彼女の関係がうまくいかないときは、私たちは常に相手を責める。	N3	Similar	☑
だから、彼らは互いに相手を非難するプロパガンダ放送を流し合っていた。	N2	Similar	☐
私たちは互いに相手を尊重し合わなければならない。	N4	Similar	☐
Syntactic Diversity			
Medium			
System ranking			
3rd			

Fig. 2: The human evaluation interface, with multiple-choice boxes. There is one for each of the three systems.

Results

Research questions

1. How much do humans (N=5) agree with each other? → They agree on difficulty level (ICC of 0.68) more than the other aspects (0.23 on average).
2. Can PLMs complete this task 0-shot? → There are problems with diversity and adjusting difficulty; Smaller models struggle more.
3. Is retrieved text preferred to generated text? → Retrieved examples were ranked first in the majority of cases.
4. Can GPT-4 evaluate the quality of examples? → Probably, but it will need more alignment with human learner ratings.

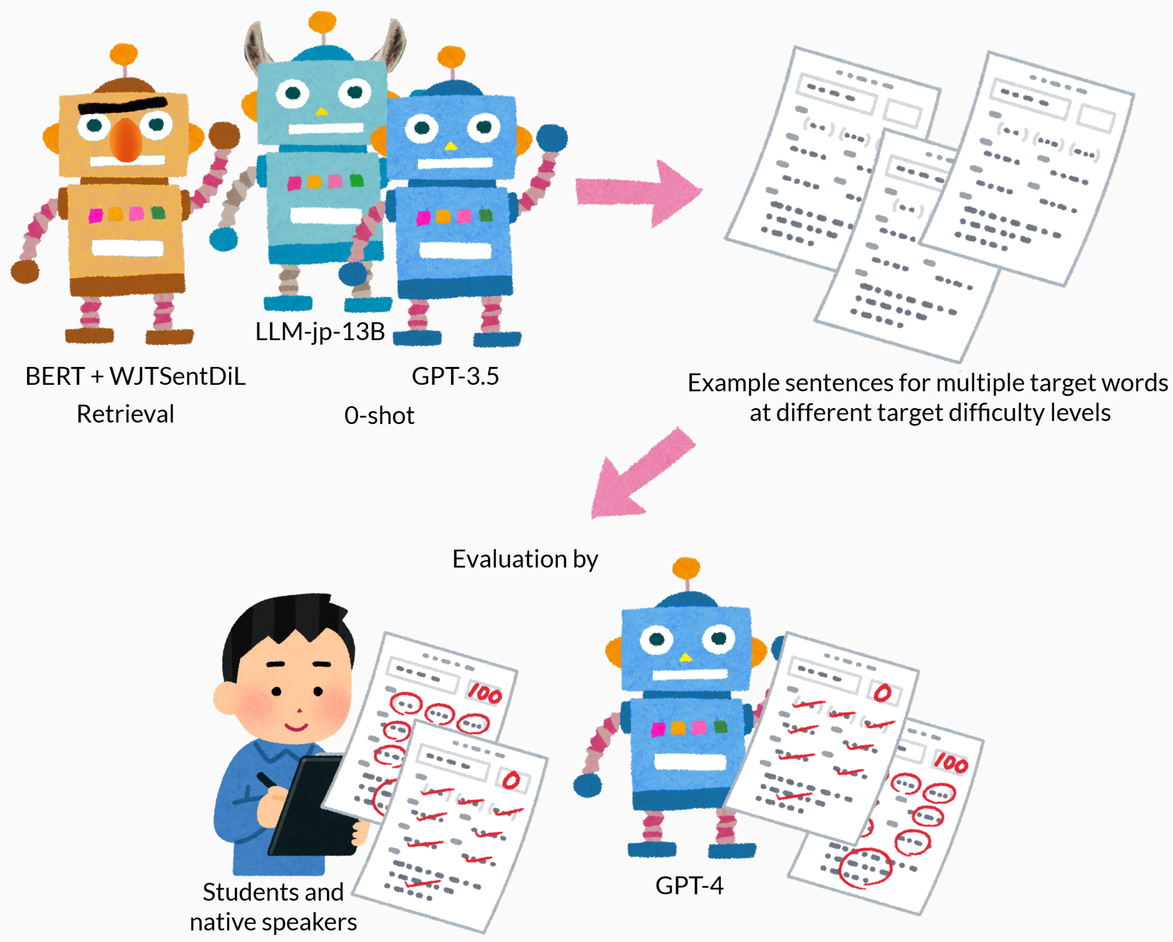

Fig. 3: Our main experiment: we test each system on 10 different target words and 3 main difficulty levels. We also measure agreement between raters (Intraclass Correlation Coefficient).

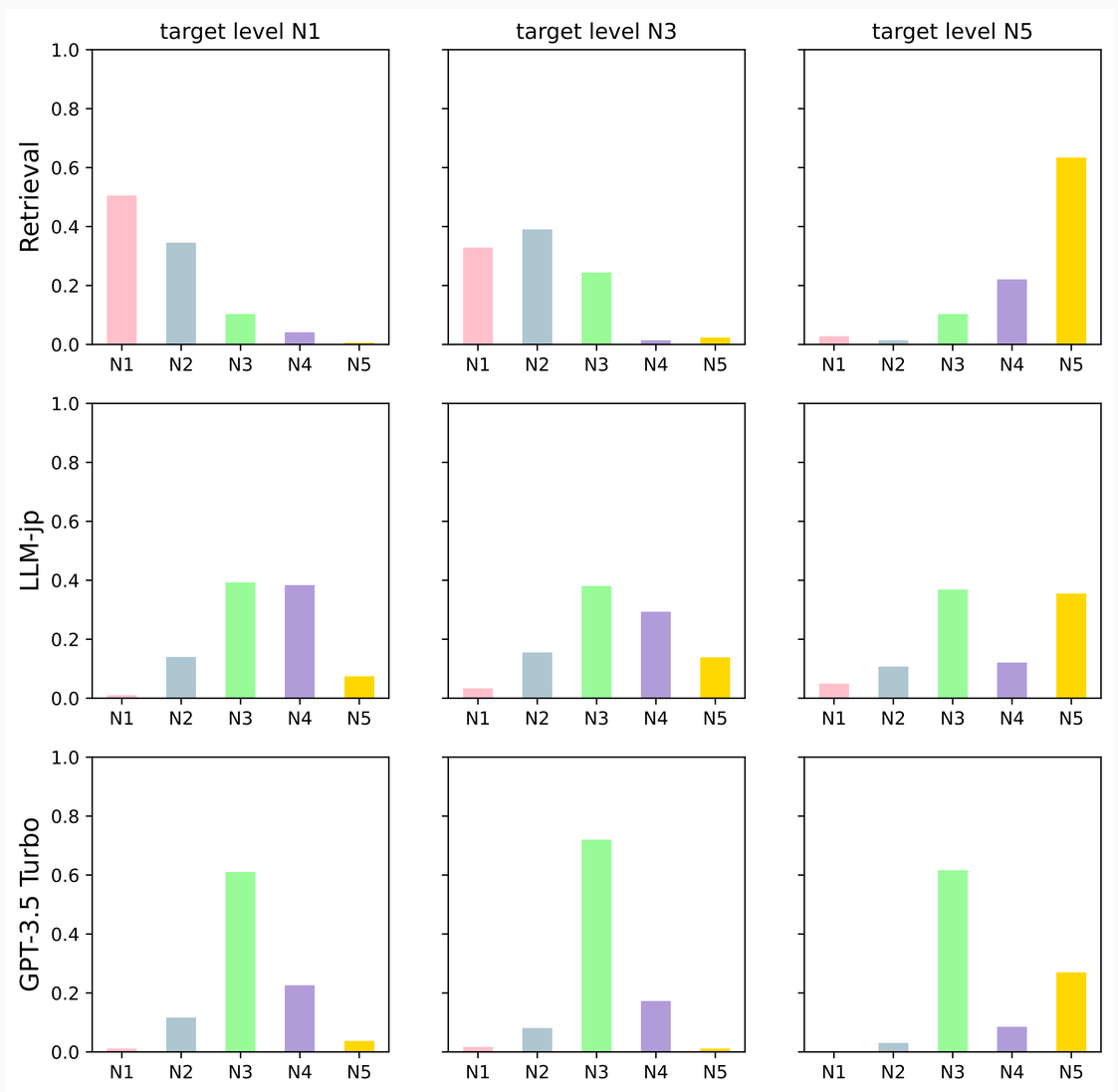

Fig. 4: Human evaluators' cumulative ratings on difficulty.

Conclusions

- We try to enhance language learning tools for learners of Japanese. This approach is a step towards obtaining more diverse and appropriate examples.
- Pre-trained Language Models can score linguistic qualities of sentences from a language learning perspective. This is useful in extracting personalized educational material.
- This study can be improved and extended to other languages.