ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in

Natural Language Processing

EXAMPLE SENTENCE SUGGESTION FOR LEARNERS OF JAPANESE AS A SECOND LANGUAGE USING PRETRAINED LANGUAGE MODELS

CANDIDATE
Enrico Benedetti

SUPERVISOR
Prof. Paolo Torroni

CO-SUPERVISOR
Prof. Akiko Aizawa

Academic year 2022-2023

Session 3rd

To Oscar, my dog. And to the important people in my life.

Abstract

In this thesis, we tackle the challenge of proposing diverse example sentences to learners of Japanese that are tailored to their proficiency level. Trying to address the lack of work using Pretrained Language Models (PLMs) on this specific task and expanding in new directions, we develop and compare different paradigms. First, we propose employing PLMs as quality scoring components of a retrieval system, retrieving from a newly curated corpus of Japanese sentences from varied sources. Second, we directly leverage PLMs as sentence generators through zero-shot learning. Then, we evaluate the quality of suggested sentences by considering multiple aspects such as difficulty, diversity, and naturalness, with a panel of raters consisting of learners of Japanese, native speakers – and GPT-4. The experimental results suggest that there is inherent disagreement among participants on the ratings of sentence qualities, except for difficulty ratings. Despite the variability, the retrieval approach was preferred by all the evaluators especially when focusing on beginner and advanced target difficulty, suggesting there is potential for using PLMs to enhance the adaptability of sentence suggestion systems to better suit learners during their journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The term second language acquisition (or L2 acquisition) refers to the acquisition of a second language by someone who already knows a first one. While children have a natural predisposition for acquiring languages, the degree of success among L2 learners varies greatly, as it is usually harder in adult life, requiring a combination of conscious effort, motivation, support from teachers and adequate materials (Fromkin et al., 2013).

Online dictionaries are usually the first resource towards which learners turn to in order to understand an unknown word or expression via definition and example sentences. However, producing high-quality learning material requires effort and expert knowledge. Because of that, researchers have explored automatic methods for example selection and generation to help professionals such as lexicographers or teachers and non-experts such as language learners (Kilgarriff et al., 2008; Ward, 2017; Pilán et al., 2013).

Pre-trained Language Models (PLMs) were shown to be effective for many NLP tasks (Wang et al., 2023). The main motivation for this work is to investigate whether PLMs can be leveraged to propose sentences that are understandable and diverse to help L2 learners be exposed to a broad range of uses for the target words they are interested in (e.g. an unknown word encountered

while reading). Example sentences contribute to improving vocabulary knowledge (Baicheng, 2009). In this study we focus on Japanese and on producing multiple example sentences at the right difficulty level.

An increasing number of people is learning Japanese (Nakamachi et al., 2022), and there is much work on obtaining high-quality text from corpora or from generative models (see Section 2.4). However, to the best of our knowledge there are few studies that address the task of example sentences suggestion in Japanese considering the latest developments in Natural Language Processing (NLP) and the impact of PLMs. The existing work mostly focuses on functional expressions (Liu et al., 2018,; Liu and Matsumoto, 2016; Shortt, 2021) or exercises (Andersson and Picazo-Sanchez, 2023).

Our contributions can be summarized as follows.

1. We develop a retrieval-based method for selecting good example sentences from a corpus, by combining different PLM modules and NLP techniques for scoring sentence quality according to four criteria: difficulty, sense similarity, syntactic and lexical diversity.
2. We build a corpus of sentences from different web sources, annotated with Japanese Language Proficiency Test[1] (JLPT) labels.
3. We evaluate the quality of sentences for specific target words by comparing the retrieval approach to two generative PLM baselines, employing volunteer native speakers and learners, alongside the GPT-4-turbo text generation model (OpenAI, 2023). We present the insights obtained from the investigation.

1.2 Thesis structure

This thesis is organized into 7 chapters and several appendices.

- In **Chapter 2 – Background and Related Work** we introduce general information on the Japanese language and its relevant aspects in NLP, as

[1] More details on the JLPT official website and in Section 3.2.1.

well as presenting the sub-field of AI for language education. Then, we discuss the related work on example sentence suggestion and difficulty estimation.

- In **Chapter 3 – Methodology**, we outline the task and the methods proposed for suggesting example sentences to language learners of Japanese. In particular, the PLM-augmented retrieval and the generative approach.
- In **Chapter 4 – Experimental setup** we present a curated dataset of Japanese sentences with difficulty labels and we describe in detail how we implemented the baselines object of the study.
- In **Chapter 5 – Evaluation** we present the motivations and research questions that guide the human evaluation experiment for comparing the systems. Additionally, we describe the evaluation setting and protocol employed to obtain system and sentence ratings.
- In **Chapter 6 – Results and discussion** we inspect both the agreement and consistency of evaluators' judgments as well as the actual ratings by using descriptive statistics. Then we discuss the insights obtained from the evaluation and from comment and error analysis.
- In **Chapter 7 – Conclusions** we discuss the limitations, future directions and the key points of this work.

Chapter 2

Background and Related Work

2.1 The Japanese language

One of the most apparent distinguishing characteristics of Japanese with respect to English is the writing system. Japanese modern writing uses three systems: *kanji* (lit. *Chinese character*), *hiragana* and *katakana*. Kanji are ideograms and were introduced around the sixth-seventh century. Later, the two simplified syllabic alphabets hiragana and katakana evolved from kanji use around the ninth century (Kubota, 1989).

All of them can appear in the same text. Usually, nouns and the roots of verbs and adjectives are written using kanji; hiragana are used for grammatical morphemes such as suffixes, auxiliaries and particles; katakana are reserved for loanwords, some proper nouns, onomatopoeia and for stylistic emphasis.

It is possible to write everything only in hiragana, but it is not considered the proper way. Take as an example the following sentences which can both be translated as "*Mother likes flowers*":

1. ははははながすきです。
2. 母は花が好きです。

The first one is written only in hiragana, and the second uses kanji for the nouns "母" and "花" referring to *mother* and *flower*. Using kanji improves

sentence readability by making the separation between sentence constituents clearer.

Even so, recognizing and understanding kanji is one of the hardest challenges for learners who come from a language background without Chinese ideograms. There are tens of thousands of Japanese kanji, of which a smaller group of 2,136 (termed jōyō, *regular-use*) was selected as a baseline for literacy by Japan's Ministry of Culture and Education.

As for grammar, a typical Japanese sentence is made up of a subject or topic and a predicate. The predicate can be nominal (NP) if it is in the form of *noun* + *copula*, adjectival (AP) or verbal (VP).

To add information to the sentence, nouns can be used in combination with postfixed grammatical particles (such as -*he* to mark direction or -*wo* to mark a noun as the direct object). Inversely, modifiers always come before the phrase they modify, as in "大きい犬", *big dog* or "私が昨日買った本", *the book I bought yesterday*.

Additionally, nouns do not explicitly carry number or gender information and any element of the sentence may be omitted if its absence does not impede comprehension, such as the subject if it is clear from context.

The main predicate must come at the end of the sentence, in what is called a Subject Object Verb (SOV) word order – though usually that the subject or topic comes at the start, it is possible to change the order to emphasize different parts of the sentence.

2.2 Japanese text tokenization

In order to apply NLP techniques to Japanese text the first key pre-processing step is tokenization. Japanese does not have explicit word boundaries like English, which has spaces between words. Therefore, to split an input sequence of characters what is ordinarily done is to combine tokenization with morphological analysis (Fujii et al., 2023). Morphological analysis divides an

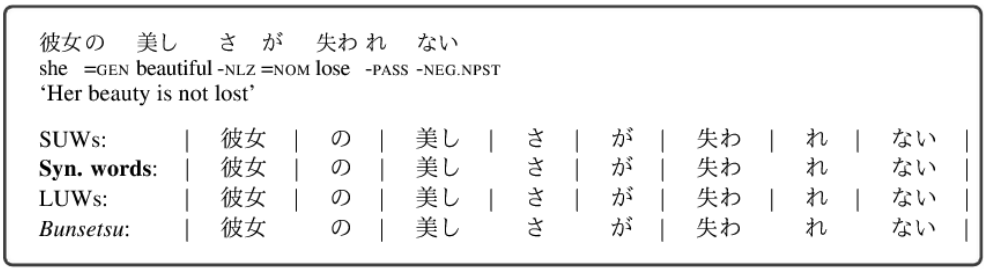

Figure 2.1: Example of sentence segmentation using different frameworks. Vertical bars indicate a boundary. Taken from Murawaki (2019).

input sequence into smaller syntactic units. The granularity of these units can vary as a parameter of the morphological analyzer, for example leaving compound nouns together or separating them into smaller elements. Traditionally, another way of splitting a sentence was dividing it into phrasal units called *bunsetsu*, but there have been efforts to streamline text processing across languages, such as the Universal Dependencies (UD) project.[1] Shown in Figure 2.1 is an example of the multiple ways in which a sentence can be tokenized.

2.3 AI in the language education domain

Because in this work we try to deal with many challenges related to material for language learning, we present a brief overview of how AI concepts can be integrated into the language learning research field.

Intelligent Computer Assisted Language Learning (ICALL) emerged as a sub-field of Computer Assisted Language Learning (CALL). ICALL focuses on applying AI concepts and technologies to CALL, especially natural language processing (NLP) and computational linguistics, user modeling, expert systems, and intelligent tutoring systems (Woo and Choi, 2021).

For an example of a successful and comprehensive ICALL application we can look at Duolingo, a vocabulary and grammar learning system which uses AI to incorporate a spaced repetition component for vocabulary acquisition.[2]

[1] https://universaldependencies.org/
[2] https://en.duolingo.com/

Many ICALL systems are specific to a certain aspect of language learning, such as word processors with spelling and grammatical error checkers, or applications that employ speech recognition technologies to improve pronunciation (Ward, 2017).

Focusing on Large Language Models (LLMs), Caines et al. (2023) note that LLMs could be promising in many areas of interest for ICALL. In content generation, LLMs have been employed in question generation for reading comprehension, prompt generation for writing and speaking exercises, and text simplification. Aside from generation, LLMs have been used for calibrating the difficulty of learning material, for example by assessing text difficulty. In addition, LLM-based systems have been studied for grading essays, and giving personalized feedback in response to learners' grammar mistakes.

However, assessing the effectiveness and utility of these systems, particularly in evaluating open-ended text generation remains an open area of study.

2.4 Related Work

In the following we discuss the related work in the main relevant areas to this work, namely retrieving and generating example sentences, and estimating sentence difficulty.

2.4.1 Example selection

Similarly to Tolmachev and Kurohashi (2017), we seek to provide high-quality and diverse example Japanese sentences. They propose a thorough retrieval approach based on quality and diversity scoring using a Determinantal Point Process, and carry out an evaluation with L2 learners and a teacher. Our work differs from theirs in that we focus on selecting sentences for sense similarity given a target word in context, instead of many possible senses for a word in isolation. Furthermore, we evaluate more aspects of the systems, in particular

their capacity to adapt their outputs to learner proficiency levels. We also employ a language model in the evaluation.

Many other works deal with the task of example sentence selection from a corpus, focusing on dictionary examples (Kilgarriff et al., 2008; de Melo and Weikum, 2009; Hazelbeck and Saito, 2009; Pilán et al., 2013) for languages such as English, Japanese and Swedish. Additionally, Shinnou and Sasaki (2008), Kathuria and Shirai (2012) and Cheng et al. (2018) leverage parallel corpora to extract disambiguated sentences. In our case, we limit our experiments to the monolingual setting.

2.4.2 Example generation

There is a lot of research on controllable text generation approaches (Zhang et al., 2023). Possible targets for generation are definitions for a given term (Zhang et al., 2023; Gardner et al., 2022), as well as example sentences. When it comes to example generation, researchers have shown that generated sentences can improve performance in Word Sense Disambiguation tasks in a supervised (Barba et al., 2021) or unsupervised way (He and Yiu, 2022). Focusing on L2 learners, Harvill et al. (2023) consider lexical complexity and sentence length to generate example sentences of controllable difficulty.

In our case, we opt not to rely on fixed sense inventories, primarily due to the scarcity of available sense-tagged corpora for Japanese. However, we believe that assigning dictionary definitions to words could prove beneficial to learners in future research.

2.4.3 Sentence difficulty estimation

Determining the level of difficulty of text is a key challenge in educational NLP, as vocabulary and grammatical structure of languages interact in a complex way (Collins-Thompson, 2014). To estimate the difficulty of Japanese sentences, Nakamachi et al. (2022) show that a BERT-based classifier trained

on labeled examples can achieve good performance, surpassing existing readability metrics[3] and approaches based on word frequencies. Liu and Matsumoto (2017) focus on estimating Japanese text difficulty for learners with pre-existing knowledge of Chinese characters. In that case, the main source of difficulty is not vocabulary, but grammar and functional expressions. In our work, due to lacking training data from official JLPT material, we resort to training a similar classifier to Nakamachi et al. (2022) with different data.

[3] https://jreadability.net/sys/en

Chapter 3

Methodology

We describe the task, and then present the baselines and datasets.

3.1 Task formulation

We define the L2 contextualized example suggestion task as

$$M(w, s_0, d) = \{s_1, s_2, \ldots, s_i, \ldots, s_K\} \tag{3.1}$$

Given a target word w, a context sentence s_0 and a target difficulty level d, we want to obtain a list of K good example sentences from a model M.

To expand more on what makes a good example, Kilgarriff et al. (2008) suggest that such examples should possess the following characteristics: represent typical usage, be informative and understandable to learners. Building upon the discussion presented by Tolmachev et al. (2022), we aim to obtain multiple examples with diverse syntactic patterns and lexical collocations because learners preferred them.

In Figure 3.1 we show a concrete example, using outputs from the retrieval approach.

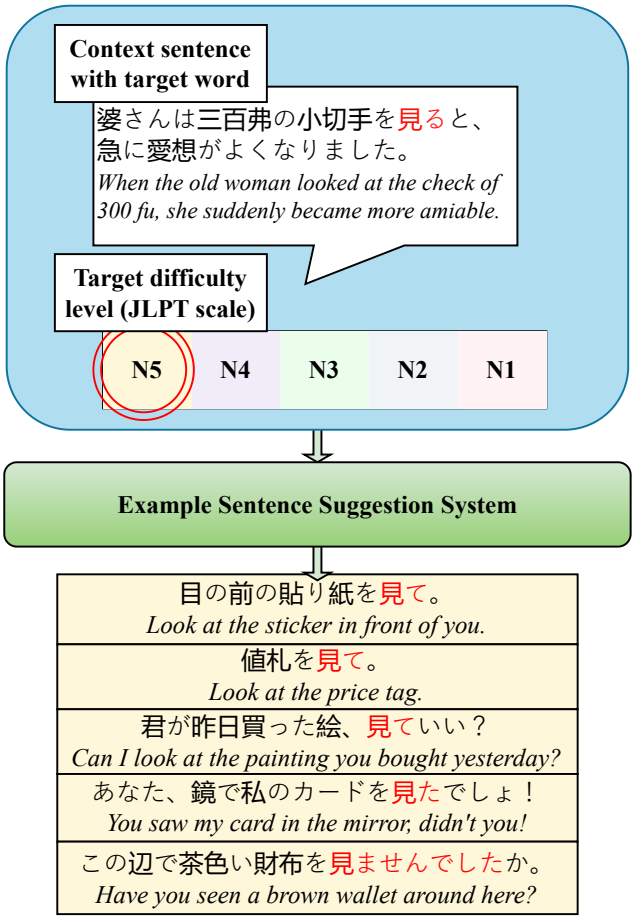

Figure 3.1: Overview of the task. Given a word in context and a difficulty level, the system will suggest diverse and level-appropriate example sentences for that word. In the example, the word is 見る (*miru*, to see).

3.2 Retrieval method

We present our methodology for building a system for example sentence selection.

We design a retrieval model that given a query, will select candidate sentences containing a target word from a corpus (for more details on it, see Section 4.1). Candidate sentences are ranked by how closely they match the target difficulty level and the semantic similarity of the target word in both the suggested and context sentences. Finally, the model selects a subset of sentences considering the total diversity of the list (syntactic and lexical).

We devise a model to quantify automatically for a sentence s_i:

1. how adequate s_i is with respect to the target difficulty level d (Sec. 3.2.1).
2. if s_i contains the target word w and it is used in the same sense as the target word of the context sentence (Sec. 3.2.2).
3. the diversity of the suggested sentences $[s_0, s_1, s_2, \ldots, s_i, \ldots, s_K]$ on vocabulary and syntax (Sec. 3.2.3).

3.2.1 Quality: difficulty

The Japanese Language Proficiency Test (JLPT) proposes a proficiency scale similar to the Common European Framework of Reference for Languages (CEFR). The JLPT levels are, from easier to harder, N5, N4, N3, N2, N1. Our classifier will therefore assign a JLPT level to input sentences.

Then it will be translated into a score between 1 and 0, where 1 means the difficulty is adequate, and 0 means that it is not (too easy or too hard). We formulate this score as

$$max\Big(0, 1 - \text{penalty}_{\text{diff}} * (d - d_i)\Big) \tag{3.2}$$

where d and d_i are the target difficulty level and difficulty label of sentence i.

In this experiment, we apply a higher penalty if a sentence is labeled as harder than the target level, reasoning that for L2 learners it may be preferable to have easier sentences when there is a discrepancy.

3.2.2 Quality: sense similarity

Anderson and Camacho-collados (2022) and Pilehvar and Camacho-Collados (2019) propose Words in Context (WiC), a different declination of Word Sense Disambiguation. WiC is formulated as a binary classification problem: given a target word and two contexts, the model has to predict whether the target word is used with the same meaning.

To tackle the WiC problem in our use case, we turn to MirrorWiC, an unsupervised fine-tuning method for obtaining contextualized word sense embeddings (Liu et al., 2021). We fine-tune a PLM with MirrorWiC and use the resulting model to extract a vector representation for the target words. We then assign a sense similarity score based on cosine similarity between s_0, the context sentence, and s_i.

3.2.3 Diversity: lexical

As a lexical diversity score, we simply adopt the average of the percentage of unique 1-2-3-4-grams in a sentence list, also considering the context sentence.

3.2.4 Diversity: syntactic

Inspired by the way Tolmachev and Kurohashi (2017) measure syntax diversity around the target words, we opt for a simpler approach supported by other works on syntactic similarity (Chen et al., 2023; Kanagawa and Okadome, 2016).

We compute dependency trees of two sentences and partially generalize their labels, then apply a Label-based Tree Kernel Similarity method to obtain a diversity measure (Chen et al., 2023; Moschitti, 2006; Boghrati et al., 2018).

To compute the syntactic diversity of a list of sentences, we take the average of pairwise scores.

3.2.5 Ranking and Greedy Selection

As the number of candidates can be very high, we build a set of K final sentences with a greedy algorithm. First, we sort the candidate sentences in order of difficulty and sense scores. Then, within a window, we add iteratively the sentence which achieves the highest diversity score, until the list is complete.

3.3 PLM generation method

Considering the PLM baselines, we prompt them with the query, expressed in English. We share the prompt used in Appendix C.

As initial experiments revealed that complying with the query in zero-shot manner was quite difficult, we prompt the PLM multiple times, concatenate the generation outputs and filter out sentences without the target word, to get the required number of output sentences. In the majority of cases, twice was enough.

Chapter 4

Experimental setup

4.1 Dataset: WJTSentDiL Corpus

We build a Japanese sentence corpus (**W**ikipedia, **J**pWaC and **T**atoeba **Sent**ences with **Di**fficulty **L**evel) by merging together three public data sources, which we describe below. We also perform additional filtering to remove spurious sentences.

- Tatoeba[1] is a collaborative online platform where users can share sentences and translations. We select only Japanese sentences and fix errors where entries are made from multiple sentences.
- jpWaC (Sangawa et al., 2010) is a curated corpus of sentences automatically collected from Japanese web domains. We include subsets L0 to L4 of the corpus.
- Wikipedia is a free online encyclopedia. We process raw article text from the Japanese part of the website, more specifically the "jawiki dump" from December 2023.

We use the spaCy[2] and Ginza[3] Python libraries to split raw text into sentences, tokenize them, and assign part-of-speech (POS) tags.

[1] https://tatoeba.org/en/
[2] https://github.com/explosion/spaCy, version 3.7.2
[3] https://github.com/megagonlabs/ginza, version 5.1.3, 'ja-ginza' model.

To keep well-formed sentences, we apply filters following heuristics similar to Kilgarriff et al. (2008) and Sangawa et al. (2010). Namely, we keep sentences that:

- have a length between 5 and 50 tokens.
- have less than 20% of punctuation or numerals.
- do not contain tokens from the Latin, Cyrillic and Arabic scripts.
- end in a predicate and punctuation.
- are not duplicates.

Wikipedia sentences are what makes up the most of the corpus. The final composition and corpus statistics are shown in Table 4.1 and Figure 4.1.

Corpus	Sentences	Tokens	Kanji
jpWaC	152,751	13.01	0.27
Tatoeba	245,793	11.07	0.27
Wikipedia	12,306,416	26.39	0.37
WJTSentDiL	12,704,960	25.93	0.36

Table 4.1: Statistics of WJTSentDiL by source. "Tokens" is the average token count, from Ginza's tokenizer. "Kanji" reports the proportion between Chinese characters and the rest.

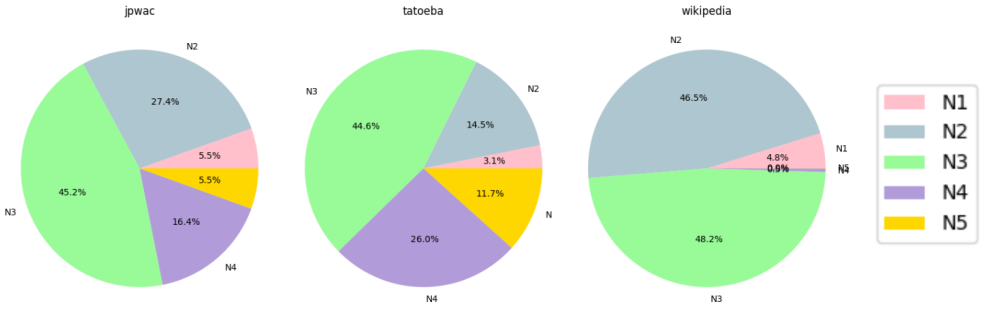

Figure 4.1: JLPT labels distribution on the original sentence sources, as assigned by our difficulty classifier.

4.2 Retrieval method details

4.2.1 Inverted index

The retrieval model uses an inverted index mapping words to the sentences they appear in. The keys are lemmas or "dictionary forms" of words and compounds. The candidate sentences are retrieved using the index by lemmatizing the target word. For example, the target word "食べた" (past form of *to eat*) is lemmatized as "食べる＋た" (*to eat* ＋ past tense auxiliary verb).

4.2.2 Difficulty classifier

The JLPT difficulty classifier is a BERT model (Devlin et al., 2019) pretrained on texts in the Japanese language by Tohoku University,[4] and finetuned on around 5,000 sentences from online Japanese language learning websites.[5]

For more details on the training and performance of the classifier, see Appendix A. Its performance is very good on in-distribution data (i.e. the validation split), but it worsens on a different test set composed of official JLPT past exam sentences. Our hypothesis is that the latter test set contains very long sentences composed of many relative clauses which are very different from the sentences that were used for training. However, during internal testing it was found to work well enough. That was partially confirmed by the raters' evaluation of difficulty for the retrieval system. Nevertheless, we explore possible improvements in the conclusion (Section 7.2).

4.2.3 Sense embeddings

We apply MirrorWiC (Liu et al., 2021) on various baseline PLMs, using as finetuning data 10,000 randomly selected sentences from our corpus. To guide

[4] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

[5] We can provide the test dataset and model weights, but not the training data because of the websites' copyright policy.

model selection, we compare performance on two WiC tasks, XL-WiC (Raganato et al., 2020) and AM2iCo (Liu et al., 2021). As highlighted in Table 4.2, MirrorWiC fine-tuning shows improvement on both tasks for the BERT model pre-trained with Japanese text, over both the not fine-tuned base model and a Japanese Sentence Transformer.[6]

Task→	XL-WiC		AM2iCo	
Model↓, **Metric** →	Accuracy	AUC	Accuracy	AUC
cambridgeltl/mirrorwic-bert-base-uncased	0.541	0.573	0.504	0.516
cl-tohoku/bert-base-japanese-v3	0.635	0.691	0.599	0.638
mirrorwic-cl-tohoku-bert-base-japanese-v3	**0.640**	**0.709**	**0.643**	**0.687**
sonoisa/sentence-bert-base-ja-mean-tokens-v2	0.598	0.654	0.592	0.633

Table 4.2: Word Embedding model Accuracy and Area Under Curve (AUC) on different Japanese WiC tasks. Best results are in bold.

To obtain the embeddings, we average the last 4 layers of the embedding model, and across the sub-tokens that make up the target word, following Liu et al. (2021).

4.2.4 Syntax diversity

To obtain a diversity score, we employ the methodology described in Section 3.2.4.

We use SpaCy to compute dependency parse trees for sentences and substitute the labels with POS and dependency labels.

Then we compute syntax similarity with FastKASSIM. More in detail, the parse trees of a pair of sentences are computed, along with the number of shared subset trees. This is normalized by dividing by the square root of the product of the number of subset trees in each parse tree (Chen et al., 2023).

[6]https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

Chapter 5

Evaluation

5.1 Goals of the evaluation

We outline the core research questions that guide our investigation.

Q0: The capabilities of Large Language Models (LLMs) such as GPT-4 in rating text have been explored (Chen et al., 2023). Therefore, can GPT-4 evaluate the quality of Japanese sentences from the perspective of L2 learners, and how do its assessments compare to those given by humans?

Q1: How do the automated quality metrics we used to guide the development of the retrieval approach compare with human judgment?

Q2: How good are PLMs at following instructions for this complex task?

Q3: Is text retrieved from a corpus (assumed to be human-authored) preferred to generated text?

Q4: What do humans think of their output?

We try to answer those questions by asking volunteer L2 learners and Japanese native speakers to manually rate and rank systems outputs.

5.2 Baselines

The systems we consider are the retrieval, described in Section 3.2; LLM-jp, a Japanese PLM,[1] and ChatGPT-3.5-turbo.

5.3 Data preparation

We build a set of target words from those used in the human evaluation of Tolmachev and Kurohashi (2017) and also add words from a Word Sense Disambiguation work by Okumura et al. (2011). The former paper used 14 target words, the latter 50, but they had one word in common, so the final count is 63. We randomly split those target words in 53 for validation and experiments, and 10 for testing and use in the human evaluation. We fix a POS count for the randomly selected test words as 3 nouns, 4 verbs, 2 adjectives and 1 adverb.

In addition, for every target word, we obtain a context sentence by randomly selecting sentences from *yourei* and *gogo*,[2] websites which provide a search engine for snippets of text content.

5.4 Human evaluation protocol

We consider as a query the input for the task (Equation 3.1), namely the selected word for human evaluation, along with their associated context sentence and target level. In this experiment we target levels N1, N3 and N5. The system outputs are randomly ordered and presented with the query, forming an "annotation block". Each baseline provides $K = 5$ sentences. This results in 30 blocks (10 queries $\times$ 3 levels) and 150 sentences from each system (30 blocks $\times$ 5 output sentences). We include an annotation block example in Appendix B.

We ask evaluators to rate:

[1] llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0
[2] https://yourei.jp, https://dictionary.goo.ne.jp

1. **Difficulty level**, by rating the difficulty of each sentence on the JLPT scale.This is to see how closely systems match the target difficulty.
2. **Sense similarity**, by evaluating whether the usage of the target word in each sentence aligns with its sense in the original context.This is to see whether the proposed sentences retain the use of the word in a similar sense, and to see whether different raters tend to give different responses.
3. **Rejection**: sentences should be marked for rejection if they are deemed not useful (e.g., unnatural usage) or confusing (e.g., grammatical errors, unclear phrasing, segmentation errors) for language learners.
4. **Syntactic Diversity**, by examining the variety in sentence structure and the different grammatical constructions used to incorporate the target word.
5. **System Ranking**: after rating each system's outputs, rank them from best to worst. The ranking should consider the overall utility of the examples for language learners at the specified target level.

We perform an initial demonstration where we present the task and explain the evaluation guidelines.

The participants are 3 native Japanese speakers and 2 learners of proficiency N1-N2. All are also proficient in English.

5.5 PLM evaluation protocol

We feed GPT-4-turbo (OpenAI, 2023) a modified version of the evaluation guidelines, the system outputs, and ask it to give the same ratings as human evaluators.

Empirically, we noticed that ratings for the same prompt sometimes were different, even when trying to reduce variability. So, we query GPT-4 three times, and also obtain its majority vote. We note that in some cases this could still result in an unclear rating.

Chapter 6

Results and discussion

In this section, we present the results from the human evaluation and the evaluation carried with GPT-4. We will discuss the results and the research questions in three main parts: agreement between annotators, system comparison, and comment and error analysis.

6.1 Q1: Agreement of ratings

The Intraclass correlation coefficient (ICC) is a widely used statistical measure for reliability, that reflects the degree of correlation and agreement between ratings (Koo and Li, 2016). The reason for choosing this metric is that it takes into account the magnitude of the differences between scores. As an example, in our setting, it is important that a sentence rated as N1 by one person and N5 by another is seen as a larger disagreement than a sentence rated N1 and N2.

Calculating the agreement on one rated quality at a time does not take into account the fact that while rating sentences, evaluators could be influenced by the other previously given ratings. To compute the metric with the pingouin Python library,[1] we convert ratings from ordinal labels into numbers, mapping them in a scale where the relative distances are the same among labels. Following Hackl et al. (2023), who studied the reliability of GPT-4 in a similar

[1] https://pingouin-stats.org/build/html/index.html

experiment as ours, we use a specific setting for the ICC based on a two-way mixed effect model. It is also known as ICC(3,1) according to the naming convention of Shrout and Fleiss (1979).

6.1.1 GPT-4 rating consistency

In Table 6.1, we report ICC values for the different ratings and for different groups of raters. We included in this computation only raters who compiled at least half of the blocks for each target level, in order to have an idea of the agreement generalizable to all difficulty levels.

For GPT-4, despite setting its behavior to be nearly deterministic and obtaining ratings on the same day, we observed that the consistency of its ratings varies by type. The model shows excellent agreement in assessing JLPT levels and good consistency in rejecting sentences. However, its consistency is lower for other evaluation areas like sense similarity, syntax diversity, and model ranking. Using a mean combination of ratings improves consistency, but comes at the cost of more forward passes on the same long inputs. A way to further mitigate this might be to improve the prompt.

Rater group→	GPT-4 ($N = 3$)		Human ($N = 3$)		All ($N = 4$)	
Rated item↓	**ICC(3,1)**	**95% CI**	**ICC(3,1)**	**95% CI**	**ICC(3,1)**	**95% CI**
Level	0.941	[0.93, 0.95]	0.681	[0.63, 0.73]	0.673	[0.63, 0.72]
Sense	0.640	[0.59, 0.68]	0.258	[0.18, 0.33]	0.108	[0.06, 0.17]
Reject	0.861	[0.84, 0.88]	0.238	[0.18, 0.30]	0.244	[0.20, 0.30]
Syn. diversity	0.778	[0.70, 0.84]	0.214	[0.08, 0.36]	0.236	[0.13, 0.36]
Ranking	0.694	[0.60, 0.78]	0.218	[0.09, 0.36]	0.218	[0.12, 0.34]

Table 6.1: ICC estimates and their 95% confidence intervals (CI) for different groups. N indicates the number of raters in the group. In the last group, we consider the ratings from the human group and a single evaluation from GPT-4, obtained by selecting the majority rating from the original 3.

6.1.2 Agreement among groups

Focusing on human raters, it seems that agreement on qualities except on difficulty level is quite low (Table 6.1). One reason for this could be that the

Rater↓→	GPT-4_{majority}	GPT-4_1	GPT-4_2	GPT-4_3	HL 1	HL 2	HN 1	HN 2	HN 3
GPT-4_{majority}	1	0.80*	0.78*	0.93*	0.37*	0.22*	0.37*	0.05	0.20
GPT-4_1	0.80*	1	0.55*	0.72*	0.33*	0.17	0.35*	0.02	0.11
GPT-4_2	0.78*	0.55*	1	0.82*	0.29*	0.17	0.45*	0.13	0.28*
GPT-4_3	0.93*	0.72*	0.82*	1	0.37*	0.21*	0.28*	-0.03	0.20
HL 1	0.37*	0.33*	0.29*	0.37*	1	0.29*	0.46*	0.13	0.68*
HL 2	0.22*	0.17	0.17	0.21*	0.29*	1	0.22*	0.14	0.47*
HN 1	0.37*	0.35*	0.45*	0.28*	0.46*	0.22*	1	0.30*	0.42*
HN 2	0.05	0.02	0.13	-0.03	0.13	0.14	0.30*	1	0.42*
HN 3	0.20	0.11	0.28*	0.20	0.68*	0.47*	0.42*	0.42*	1

Table 6.2: Pairwise agreement matrix of ICC(3,1) scores on **ranking preferences**. "HL" refers to a human learner, while "HN" to a human native speaker. *: P-value $< .05$.

guidelines for other metrics are too generic, which causes more variability in the ratings, although it was expected that language learners and native speakers of Japanese may not have the same rating patterns. Additionally, since we required many ratings at once, there could be some additional effects (e.g. fatigue, bias from the order of annotation).

6.1.3 Pairwise agreement on ranking

To further investigate whether a LLM such as GPT-4 ranks similarly to humans, in Table 6.2 we report the pairwise agreement for the preferred system ranking from all annotators. Inter-rater agreement scores between GPT-4 and humans are generally lower than those among humans of different groups. This suggests that humans, regardless of whether they are native speakers or not, have more similar ranking preferences compared to the AI models. However, there are also outliers, such as HN 2, who has a way of ranking that shows no agreement with many other raters.

The experiment showed the difficulty of making AI evaluations match human preferences and confirms that even among humans, there is an inherent amount of disagreement on judgments assessing the suitability of learning material.

6.2 Q2-Q3-Q4: Quantitative analysis of ratings

After the agreement analysis, we discuss how the raters evaluated the systems. For the qualities other than difficulty label and ranking preference, we report for brevity only the main findings. The full data can be found in Appendix E.

6.2.1 Difficulty level ratings

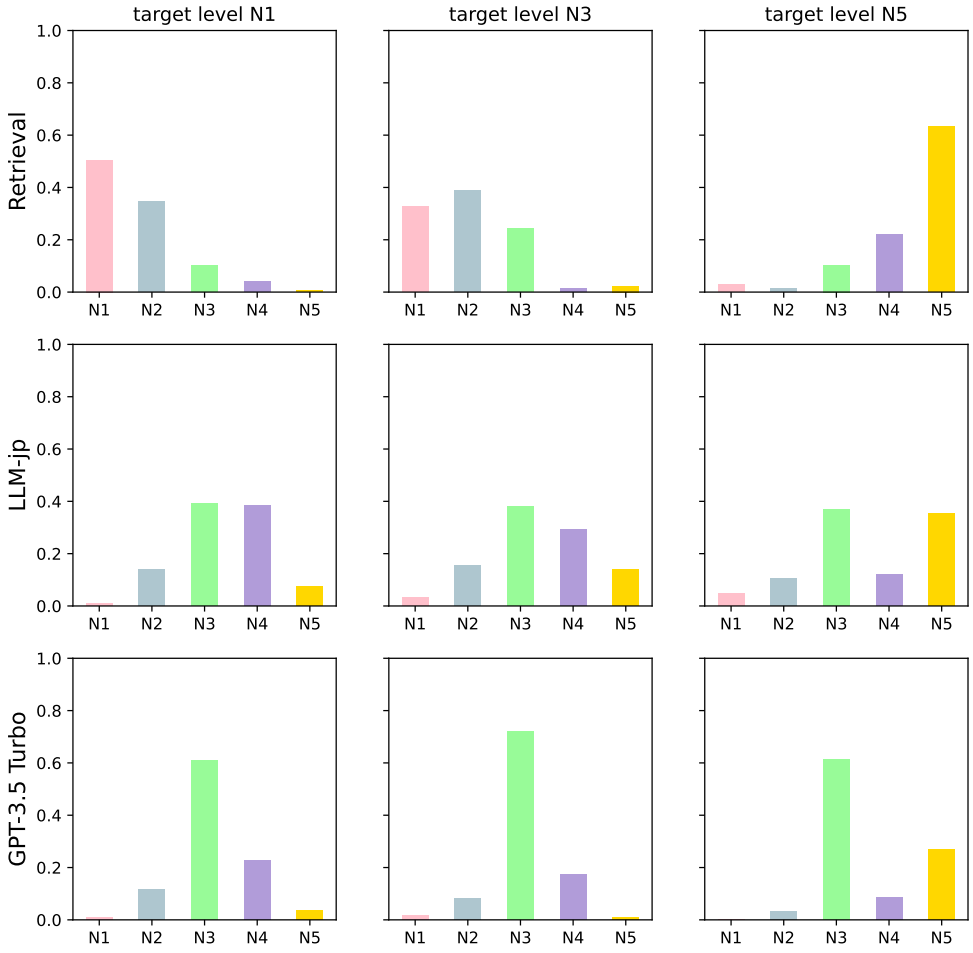

Figure 6.1: Evaluators' ratings on difficulty. Each row presents the proportions of JLPT labels assigned by humans for one system, across the three target difficulty levels set for the evaluation.

Figure 6.1 shows the proportion of human-assigned JLPT difficulty labels for each baseline, grouped by target level.

When considering how close the difficulty of proposed sentences is to the target level, our retrieval approach is markedly better for N1 and N5, while for N3, it produced a significant proportion of harder sentences. ChatGPT seems better for N3. Being so consistent with the difficulty of sentences is not always an advantage because it makes it difficult to adapt to different requirements, for example when requesting advanced sentences. This problem is shared with the other generative approach, although LLM-jp had more difficulties in following the prompt, i.e. repetitions, sentences without the target word, and incoherent text.

6.2.2 Sense similarity ratings

When the raters indicated whether the target word in each sentence had a similar meaning as the one in the context, a vast majority indicated that the sense was the same. The percentage of sentences rated as "not similar" was only about 2% for the retrieval, and 13% for both the generative baselines. This shows that the systems generally succeed in producing example sentences with the same sense.

6.2.3 Rejection ratings

According to the evaluation guidelines, unnatural sentences and those with confusing errors should be marked. On average, 8% of sentences suggested by the retrieval were rejected, while for LLM-jp it was 13%, and 16% for ChatGPT-3.5. Checking raters' comments confirmed that there were some segmentation errors in retrieval and generation baselines, such as sentences starting with punctuation, or with a fragment. It seems that generative models are more prone to errors, while the retrieved sentences are better in this aspect "by design". Still, careful pre-processing and post-processing of text are needed as errors can be confusing especially for beginner learners. We discuss more about this in Section 6.3.

6.2.4 Diversity ratings

Considering the syntax diversity of the list of sentences, the retrieval method earned the most "high" ratings across all target levels. ChatGPT-3.5 received mostly "medium" votes, and LLM-jp got the lowest ratings. The latter model often produced very repetitive sentences, where only one or two words were different. This highlights a problem in zero-shot text generation, that it is difficult to obtain both diversity and adherence to the prompt.

6.2.5 System preference ratings

Table 6.3 presents the votes on "which one is the best system?" by all human participants, and GPT-4's majority ratings. The sentence lists produced by the retrieval system were considered better for all raters when considering the total vote count. Except for HL2 and HN3, the retrieval system was rated best in over 50% of cases. When considering target levels, it also markedly wins in suggesting lists for advanced and beginner target difficulty levels, while it is not rated best as often for the intermediate level. It appears that the sentences suggested by the retrieval system for the N3 level are often on the more difficult end, as shown in Figure 6.1.

System→	Retrieval				LLM-jp				ChatGPT-3.5			
Rater↓, **Target**→	N1	N3	N5	Tot.	N1	N3	N5	Tot.	N1	N3	N5	Tot.
GPT-4_{majority}	**7**	**5**	**5**	**17**	2	2	2	6	1	3	3	7
HL 1†	**5**	**4**	–	**9**	0	0	–	0	0	2	–	2
HL 2	**4**	3	**6**	**13**	2	2	2	6	**4**	**4**	2	10
HN 1	**10**	**4**	**10**	**24**	0	2	0	2	0	**4**	0	4
HN 2	**7**	1	**8**	**16**	2	**5**	1	8	1	4	1	6
HN 3†	**7**	1	–	**8**	1	2	–	3	0	**6**	–	6

Table 6.3: Number of annotation blocks in which the considered baseline is rated first in overall quality, by target difficulty level. †: The participant mostly rated blocks with target level N1 and N3 only, because of time constraints.

6.3 Q5: Qualitative analysis of participants' comments

Comments on errors There were segmentation errors in retrieval and generation baselines, such as sentences starting with punctuation, or with an ungrammatical fragment.

For an example, we can consider one error from the retrieval and one from LLM-jp:

- と怒るのだが、毎回カメラのスイッチングのタイミングが合わず、タイミングが合ったら合ったでピンボケを起こしている, translating to **(と-is-angry-but), every time the timing of the camera switching is off, and when the timing is right, it ends up being out of focus*. と怒るのだが is a segmentation error because it starts with a quoting particle.

- favorite dish is sushi.1. 右手で持っていたスプーンを左手でも持てるようになったんだ. The Japanese part translates to *I've become able to hold the spoon with my left hand, which I had been holding with my right hand*. There is a segmentation error in the first part, due to the model including English characters and punctuation, which was not fixed by the post-processing function at the time.

Careful pre-processing and post-processing of text is needed before actually presenting the sentences, especially to beginner learners, as these errors could be confusing.

Comments on sense According to the participating native speakers, there were also some unnatural-sounding sentences among the suggestions and also in the randomly chosen contexts.

As a pointer towards saying that the sense similarity notion we used is too

general, a native speaker noted that in a context sentence and in some suggested sentences, the target word has the same writing form but different readings depending on nuances of meaning and context. For example,「開く」can be read as either *aku* or *hiraku*. Their meaning is not so different, but there are some differences.

Apart from pointing out segmentation errors, a native Japanese speaker commented on a target word in the evaluation (全然, *zenzen*). It is commonly used in negative statements, to mean "not at all" (Sawada, 2007). Using it in positive statements would be considered "slightly broken" in a formal situation, but it was correct a hundred years ago, and it is used in today's slang. A generation system produced a similar sentence as the context in which the usage was "uncommon". Indeed, the context sentence chosen from the web was an excerpt from a collection published in 1938 by Osamu Dazai, a famous Japanese writer. This should prompt thinking about what actually makes a sentence correct.

This usage ties into a linguistic principle known as polarity, a concept found across all human languages (Löbner, 2000). When a word typically associated with negative contexts is used in a positive statement, it can sound odd, similar to "I ran at all" in English.

Comments on difficulty Language learners noted that many sentences contained one or two difficult kanji, that are encountered at higher proficiency levels, even though the overall sentence structure is more straightforward to understand. This happened mostly with the retrieval approach, which used a text classifier for difficulty, which did not take word difficulty explicitly into account.

Chapter 7

Conclusions

7.1 Limitations

In our work, the retrieval approach dealt with scoring sentences using mainly unsupervised approaches and PLMs. The corpus we build is not as large as other corpora. In our comparisons, for LLMs we explored only basic prompting strategies without fine-tuning, wanting to investigate approaches in a setting without labeled data.

As for the evaluation, the number of volunteers who participated in the study was quite limited. Additionally, comparing our baselines with the approach of Tolmachev and Kurohashi (2017) would have been insightful. However, due to the absence of a practical implementation and limited resources for human evaluation, we opted for PLM baselines.

Regarding the limitations on content checking, sentences generated or retrieved using these approaches could reflect negative biases that could impact or influence negatively the model of language that is internalized by the learners. It poses an increased risk when there are not enough sources of information, or limited sharing of ideas and communication with other learners and native speakers of the foreign language that can more effectively teach distinguishing polite and casual register and other aspects of pragmatics, other than just word usage.

7.2 Conclusion and future directions

This thesis presents a methodology for example sentence suggestion for learners of Japanese (it can also be applied to multiple languages, with small changes in the implementation). The baselines we considered highlight the many possible roles of Pretrained Language Models: assessing difficulty, providing semantic representations, directly producing sentences and evaluating their quality.

From the feedback and data collected from the evaluation with volunteer human learners and native speakers, we can point out that, even though the retrieval methodology was considered to be the best in terms of adherence to difficulty level and diversity, there is potential for improvement and for combining these systems to balance their shortcomings.

Evaluating generated text is a longstanding challenge, and in the area of language learning there are no well-established automated metrics, so we investigated the capabilities and responses of a state-of-the-art LLM in rating text. In our opinion, it is promising because the model seems to have the ability to evaluate linguistic features of sentences. While a general agreement in rating text difficulty could be found, since each person can make different assessments, finding a way to make models take that variability into account could be useful.

It could be studied whether using word-level features can prevent unknown kanji from appearing in example sentences. Such features could be a JLPT label or the Japanese school grade level they are taught in. Another interesting area of study is estimating the actual vocabulary known by the learner, modeling the process of second language acquisition (Settles et al., 2018; Cui and Sachan, 2023). Also regarding personalization, there is potential for suggestion and generation of L2 material based on each learner's interests. A direction to explore further is to experiment with more advanced LLM prompting strategies, such as Chain of Thought or Reinforcement Learning, to iteratively

refine outputs for better adaptation to learners' preferences. A retrieval approach like ours could serve as a starting point.

We hope that our findings and collected feedback could prove helpful in developing systems for obtaining better L2 learning material automatically, in a way that benefits language learners along their path to proficiency.

References

[1] Mark Anderson and Jose Camacho-collados. 2022. Assessing the limits of the distributional hypothesis in semantic spaces: Trait-based relational knowledge and the impact of co-occurrences. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 173–185, Seattle, Washington. Association for Computational Linguistics.

[2] Tim Andersson and Pablo Picazo-Sanchez. 2023. Closing the gap: Automated distractor generation in japanese language testing. *Education Sciences*, 13(12).

[3] Zhang Baicheng. 2009. Do example sentences work in direct vocabulary learning? *Issues in Educational Research*, 19.

[4] Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. Exemplification Modeling: Can You Give Me an Example, Please? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3779–3785, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

[5] Reihane Boghrati, Joe Hoover, Kate M Johnson, Justin Garten, and Morteza Dehghani. 2018. Conversation level syntax similarity metric. *Behavior research methods*, 50(3):1055–1073.

[6] Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. 2023. On the application of large language models for language teaching and assessment technology.

[7] Maximillian Chen, Caitlyn Chen, Xiao Yu, and Zhou Yu. 2023. Fastkassim: A fast tree kernel-based syntactic similarity metric. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

[8] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

[9] Shang-Chien Cheng, Jhih-Jie Chen, Chingyu Yang, and Jason Chang. 2018. LanguageNet: Learning to Find Sense Relevant Example Sentences. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 99–102, Santa Fe, New Mexico. Association for Computational Linguistics.

[10] Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

[11] Peng Cui and Mrinmaya Sachan. 2023. Adaptive and personalized exercise generation for online language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 10184 – 10198, Stroudsburg, PA. Association for Computational Linguistics. 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023); Conference Location: Toronto, Canada; Conference Date: July 9-14, 2023.

[12] Gerard de Melo and Gerhard Weikum. 2009. Extracting sense-disambiguated example sentences from parallel corpora. In *Proceedings of the 1st Workshop on Definition Extraction*, pages 40–46, Borovets, Bulgaria. Association for Computational Linguistics.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[14] V. Fromkin, R. Rodman, and N. Hyams. 2013. *An Introduction to Language*. Cengage Learning.

[15] Takuro Fujii, Koki Shibata, Atsuki Yamaguchi, Terufumi Morishita, and Yasuhiro Sogawa. 2023. How do different tokenizers perform on downstream tasks in scriptio continua languages?: A case study in Japanese. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 39–49, Toronto, Canada. Association for Computational Linguistics.

[16] Noah Gardner, Hafiz Khan, and Chih-Cheng Hung. 2022. Definition modeling: literature review and dataset analysis. *Applied Computing and Intelligence*.

[17] Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education*, 8.

[18] John Harvill, Mark Hasegawa-Johnson, Hee Suk Yoon, Chang D. Yoo, and Eunseop Yoon. 2023. One-Shot Exemplification Modeling via Latent Sense Representations. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 303–314, Toronto, Canada. Association for Computational Linguistics.

[19] Gregory Hazelbeck and Hiroaki Saito. 2009. A Corpus-based E-learning System for Japanese Vocabulary. *Journal of Natural Language Processing*, 16(4):3–27.

[20] Xingwei He and Siu Ming Yiu. 2022. Controllable Dictionary Example Generation: Generating Example Sentences for Specific Targeted Audiences. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 610–627, Dublin, Ireland. Association for Computational Linguistics.

[21] Eriko Kanagawa and Takeshi Okadome. 2016. Syntactic characteristics and similarities of japanese authors' writing styles: A kernel-based approach. In *2016 International Conference on Asian Language Processing (IALP)*, pages 59–62.

[22] Pulkit Kathuria and Kiyoaki Shirai. 2012. Word Sense Disambiguation Based on Example Sentences in Dictionary and Automatically Acquired from Parallel Corpus. In *Advances in Natural Language Processing*, Lecture Notes in Computer Science, pages 210–221, Berlin, Heidelberg. Springer.

[23] Adam Kilgarriff, Milos Husák, Katie McAdam, Michael Rundell, and P. Rychlý. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress*, Barcelona, Spain. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

[24] Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163.

[25] Yoko Kubota. 1989. *A Brief History of Time: From the Big Bang to Black Holes*. Libreria Editrice Cafoscarina, London.

[26] Jun Liu, Fei Cheng, Yiran Wang, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Automatic Error Correction on Japanese Functional Expressions Using Character-based Neural Machine Translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

[27] Jun Liu and Yuji Matsumoto. 2016. Simplification of Example Sentences for Learners of Japanese Functional Expressions. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 1–5, Osaka, Japan. The COLING 2016 Organizing Committee.

[28] Jun Liu and Yuji Matsumoto. 2017. Sentence complexity estimation for Chinese-speaking learners of Japanese. In *Proceedings of the 31st Pacific*

Asia Conference on Language, Information and Computation, pages 296–302. The National University (Phillippines).

[29] Jun Liu, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Sentence Suggestion of Japanese Functional Expressions for Chinese-speaking Learners. In *Proceedings of ACL 2018, System Demonstrations*, pages 56–61, Melbourne, Australia. Association for Computational Linguistics.

[30] Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2021. MirrorWiC: On eliciting word-in-context representations from pretrained language models. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 562–574, Online. Association for Computational Linguistics.

[31] Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2021. MirrorWiC: On Eliciting Word-in-Context Representations from Pretrained Language Models. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 562–574, Online. Association for Computational Linguistics.

[32] Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[33] Sebastian Löbner. 2000. Polarity in Natural Language: Predication, Quantification and Negation in Particular and Characterizing Sentences. *Linguistics and Philosophy*, 23(3):213–308.

[34] Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *11th conference of the European Chapter of the Association for Computational Linguistics*, pages 113–120.

[35] Yugo Murawaki. 2019. On the definition of japanese word.

[36] Nakamachi, Toshinori, Nishiuchi, Masayu, and Oku. 2022. Estimation of japanese text difficulty based on the japanese language proficiency test.

[37] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2011. On SemEval-2010 Japanese WSD Task. *Journal of Natural Language Processing*, 18(3):293–307.

[38] OpenAI. 2023. Gpt-4 technical report.

[39] Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

[40] Ildikó Pilán, Elena Volodina, and Richard Johansson. 2013. Automatic selection of suitable sentences for language learning exercises.

[41] Ildikó Pilán, Elena Volodina, and Richard Johansson. 2013. Automatic Selection of Suitable Sentences for Language Learning Exercises. In *20 Years of EUROCALL: Learning from the Past, Looking to the Future*, pages 218–225. Research-publishing.net.

[42] Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

[43] Kristina Hmeljak Sangawa, T. Erjavec, and Yoshiko Kawamura. 2010. Automated collection of japanese word usage examples from a parallel and a monolingual corpus.

[44] Osamu Sawada. 2007. Two types of adverbial polarity items in japanese: absolute and relative. In *Proceedings of the 10th Conference of the Pragmatics Society of Japan*.

[45] Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

[46] Hiroyuki Shinnou and Minoru Sasaki. 2008. Division of example sentences based on the meaning of a target word using semi-supervised clustering. In *International Conference on Language Resources and Evaluation*.

[47] Mitchell Shortt. 2021. Synthesizing a Japanese-language functional expression learning system with Chinese-speaking learners'cultural interests and backgrounds. *Educational Technology Research and Development*, 69(1):319–322.

[48] Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological bulletin*, 86(2):420–428.

[49] Arseny Tolmachev and Sadao Kurohashi. 2017. Automatic extraction of high-quality example sentences for word learning using a determinantal point process. In *Proceedings of the 12th Workshop on Innovative Use of*

NLP for Building Educational Applications, pages 133–142, Copenhagen, Denmark. Association for Computational Linguistics.

[50] Arseny Tolmachev, Sadao Kurohashi, and Daisuke Kawahara. 2022. Automatic japanese example extraction for flashcard-based foreign language learning. *Journal of Information Processing*, 30:315–330.

[51] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2023. Pre-trained language models and their applications. *Engineering*, 25:51–65.

[52] Monica Ward. 2017. *ICALL's relevance to CALL*, pages 328–332. Research-publishing.net.

[53] Jin-Ha Woo and Heeyoul Choi. 2021. Systematic review for ai-based language learning tools. *ArXiv*, abs/2111.04455.

[54] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3).

[55] Hengyuan Zhang, Dawei Li, Yanran Li, Chenming Shang, Chufan Shi, and Yong Jiang. 2023. Assisting language learners: Automated trans-lingual definition generation via contrastive prompt learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 260–274, Toronto, Canada. Association for Computational Linguistics.

Acknowledgements

I am grateful to my family, in particular to my mother, father and grandparents, who supported me a lot even when I was far away. A big thank you to the all the people I had the luck of meeting.

Appendix A

Difficulty classifier training and evaluation

Parameter	**Value**
model	cl-tohoku/bert-base-japanese-v3
tokenizer	model's AutoTokenizer
no. labels	5 *(N1, N2, N3, N4, N5)*
learning rate	2e-5
batch size	8
no. epochs	10
adam β_1	0.9
adam β_2	0.999
adam ϵ	1e-7
weight decay	0.01

Table A.1: Summary of training parameters for the difficulty classifier.

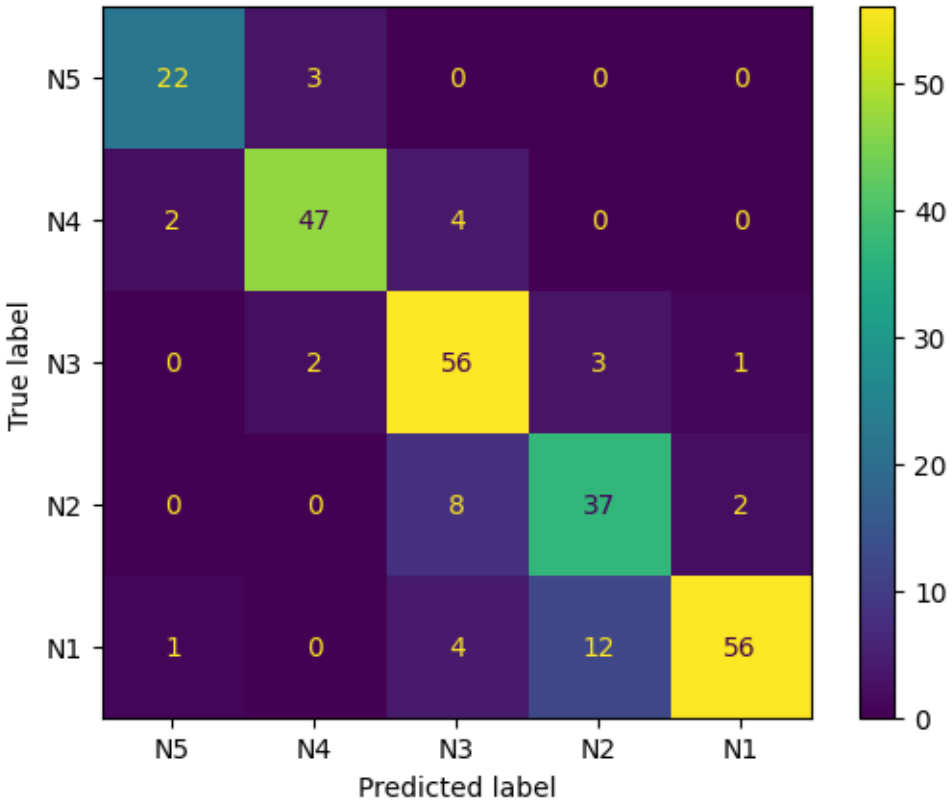

Figure A.1: Confusion matrix for the difficulty classifier, on sentences obtained in the same way as the training data (i.e. distant supervision labeling from language websites).

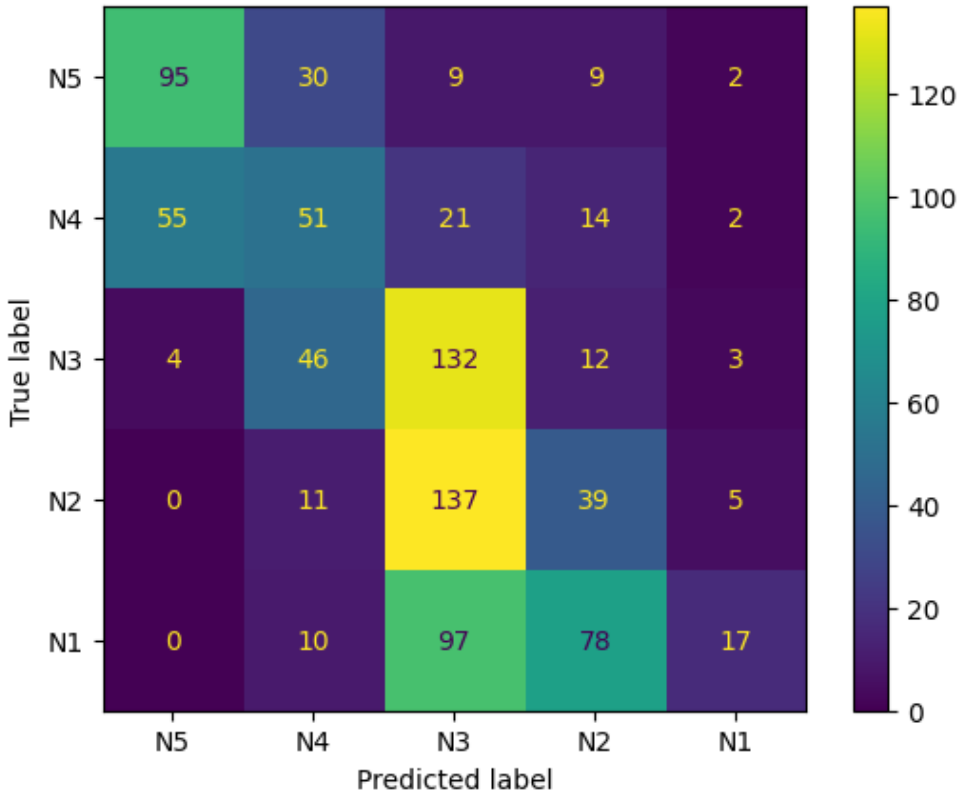

Figure A.2: Confusion matrix for the difficulty classifier, on sentences obtained from a different source (i.e. past exams from the official JLPT website).

Class	Precision	Recall	F1-score	Support
N5	0.88	0.88	0.88	25
N4	0.90	0.89	0.90	53
N3	0.78	0.90	0.84	62
N2	0.71	0.79	0.75	47
N1	0.95	0.77	0.85	73
Macro Avg	0.84	0.84	0.84	260
Weighted Avg	0.85	0.84	0.84	260
Accuracy		0.84		260

Table A.2: Metrics on data from the test split from the same data distribution for the difficulty classifier.

Class	Precision	Recall	F1-score	Support
N5	0.62	0.66	0.64	145
N4	0.34	0.36	0.35	143
N3	0.33	0.67	0.45	197
N2	0.26	0.20	0.23	192
N1	0.59	0.08	0.15	202
Macro Avg	0.43	0.39	0.36	879
Weighted Avg	0.43	0.39	0.36	879
Accuracy		0.38		879

Table A.3: Metrics on a test set of sentences from the official JLPT exams for the difficulty classifier.

Appendix B

Human evaluation form - Example of an evaluation block

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				**System 2**				**System 3**			
Suggested Sentences	**Difficulty rating**	**Sense rating**	**Reject**	**Suggested Sentences**	**Difficulty rating**	**Sense rating**	**Reject**	**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	☐	外交経験が無い素人2人組の外交は半年が空費され、相手から一方的に条件を呑まされる寸前になり失敗に終わった。	N1	Similar	☐	彼は相手の発言に敏感に反応し、即座に的を得た反論を返した。	N3	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	☐	ドイツの戦略爆撃機とイギリス、アメリカの戦略爆撃機の合計の多い国家が少ない国家（ドイツ又はイギリス）を攻撃することになり、相手国家の工業力を低下させる。	N1	Similar	☐	相手の心情を察しつつ、適切なアドバイスを提供することが大切です。	N3	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	☑	これはその言葉を発した側が、その発言を持って相手を貶めようとしているためである。	N2	Similar	☐	ディベートでは相手の弱点を見つけ、巧妙に攻撃することが求められます。	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	☐	カトコフの主張は、一般的に穏健なものではあったが、ひとたびや筆を執るや否や、痛烈に相手を批判せずにはいられなかった。	N1	Similar	☐	相手の意図を読み取りながら、戦術的な効果的な問いかけをする必要があります。	N3	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	☐	また、両派ともに相手の絶滅を主張し、小型の出刃包丁やハンマーなどを使用した襲撃を続けたため逮捕者や離脱者が多数出て、組織の維持すら危うくなった。	N1	Similar	☐	相手の動きを見極めて、適切なタイミングで反応することが勝利の鍵です。	N3	Similar	☐
Syntactic Diversity				**Syntactic Diversity**				**Syntactic Diversity**			
Medium				High				High			
System ranking				**System ranking**				**System ranking**			
3rd				1st				2nd			

Appendix C

LLM baselines prompts

We share the prompts, obtained with manual testing and trial and error. We found that the models responded in a satisfactory way also to prompts where the request was formulated in plain English, as well as in Japanese.

For LLM-jp, this was the prompt used to obtain the final outputs:

> write *k target level* example sentences in japanese, that must contain the word ”*target word*” used in a similar sense as ”*context sentence*”. following are *k* diverse sentences that must use ”*target word*”:

For ChatGPT-3.5-turbo, we used the same prompt as the other LLM, and only appended the following instruction to reduce verbosity.

> Provide sentences in Japanese in a numbered list, without any translation or romaji.

Appendix D

GPT-4 evaluation prompt

We present the prompt given to GPT-4-turbo when rating evaluation blocks with the baselines outputs:

> This evaluation aims to rate and compare three systems in providing good example sentences for learners of Japanese at different proficiency levels. An annotation block consists of proposed sentences by 3 systems for a target word, a context sentence and a target difficulty level. The lists of sentences are supposed to help language learners to see diverse examples of a target word in context.
>
> Difficulty: Rate the difficulty of each sentence according to the JLPT (Japanese Language Proficiency Test) scale, where N1 is the most difficult and N5 is the easiest. Indicate which level a sentence belongs to (one of N1, N2, N3, N4, N5). It is possible that for the target level, the system proposes a sentence that is of a different level (higher or lower). Below is a summary of the proficiency levels.[1]

[1] Taken from https://www.jlpt.jp/e/about/levelsummary.html. The description are put into a table for readability.

Level	Description
N1	Complex and abstract Japanese across various contexts.
N2	Everyday Japanese in varied situations, with clear materials on different topics.
N3	Japanese in common everyday situations.
N4	Basic Japanese understanding, including familiar topics, basic vocabulary, and kanji.
N5	Fundamental Japanese, including hiragana, katakana, and basic kanji.

Sense Similarity: Indicate if the target word in each sentence maintains a close sense as in the original context. Possible values: ”similar”, ”not similar”. Think broadly and intuitively, rather than strictly by dictionary definitions.

Reject: For each sentence, indicate ”Reject” if you think the sentence is not good or useful (for example because it does not reflect natural use).

Sentence diversity: For each system output list, rate the sentences' diversity, focusing on the number of different uses of syntax and structure. Possible values: ”Low”, ”Medium”, ”High”.

System ranking: Rank the systems' outputs from best to worst, considering the overall usefulness of the example sentences for that word, for a language learner of that proficiency level.

Comment: Leave a short comment.

Appendix E

Additional rating statistics

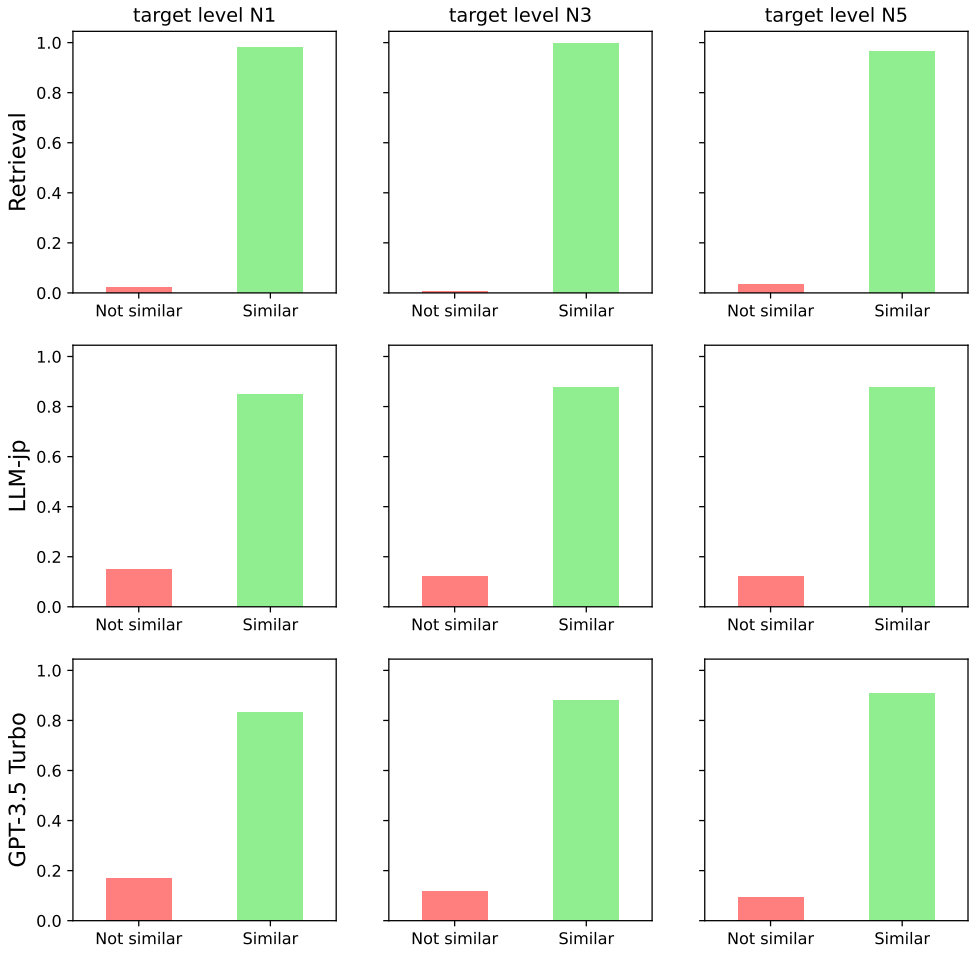

Figure E.1: Ratings on sense similarity of proposed sentences.

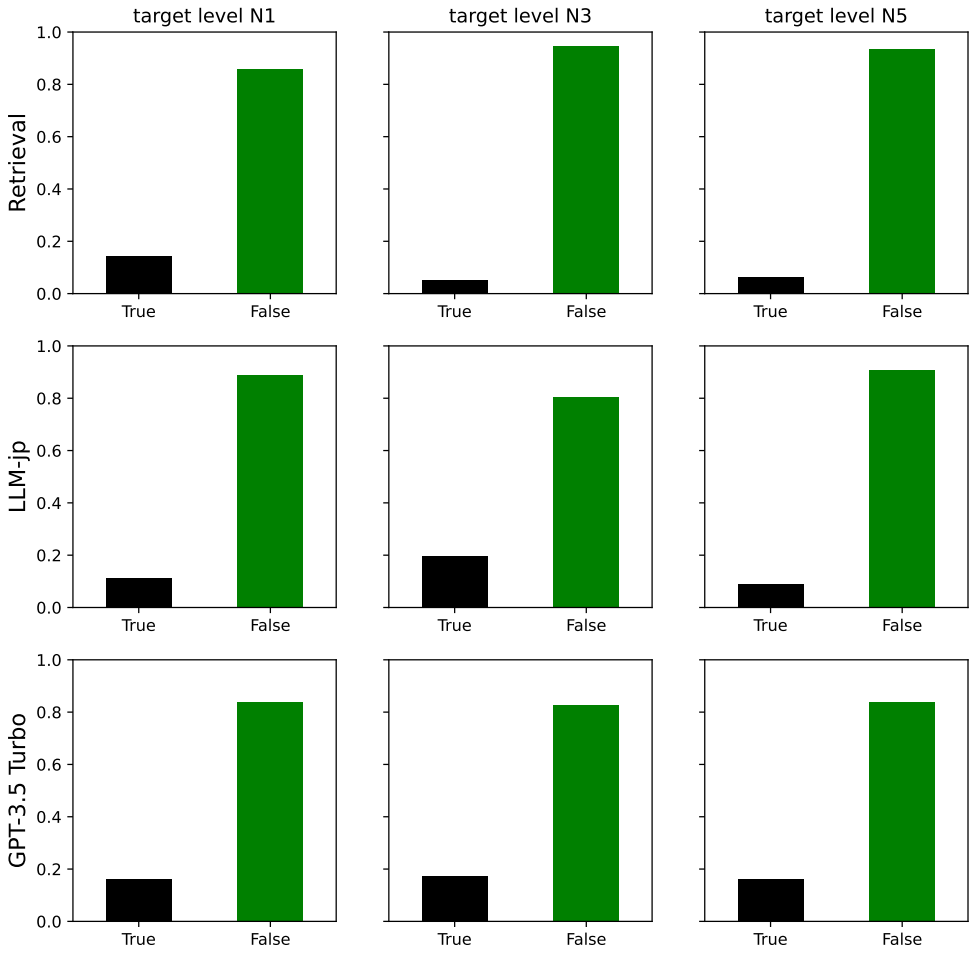

Figure E.2: Proportion of rejected proposed sentences.

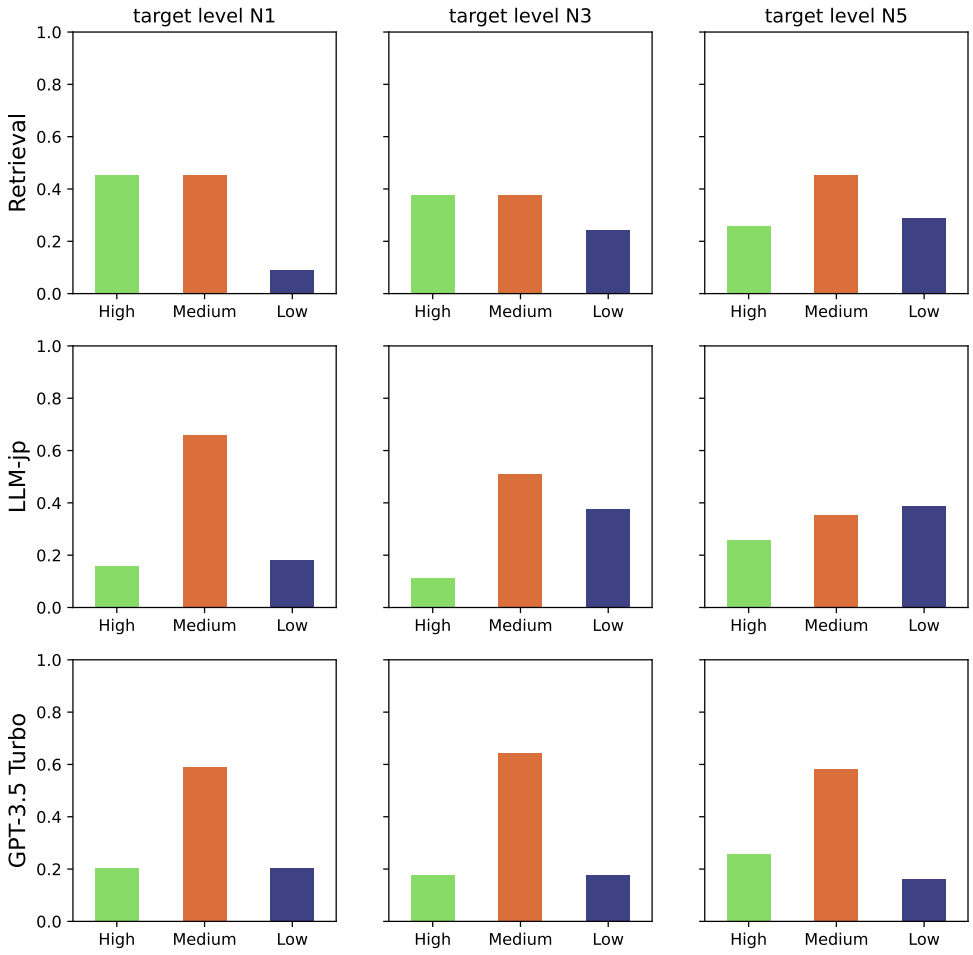

Figure E.3: Ratings on syntax diversity of proposed sentences.

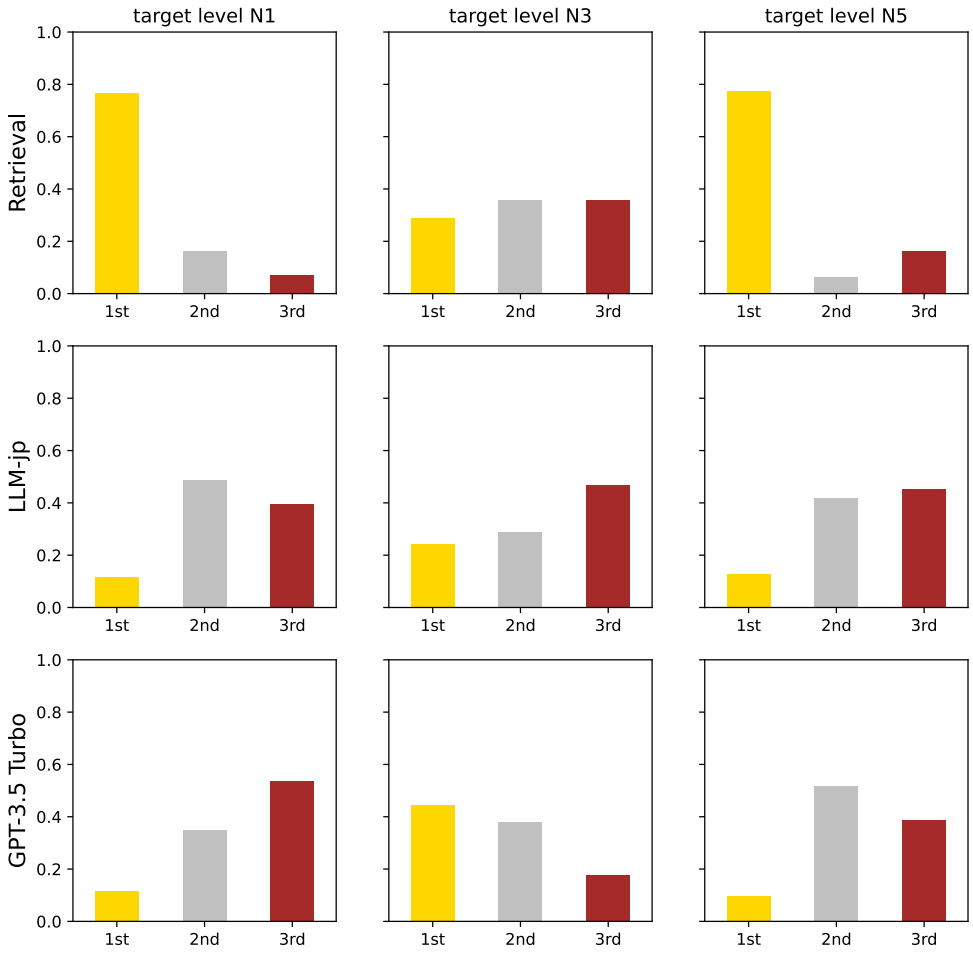

Figure E.4: Rankings (first, second, third place) for each system.