

LLM-Assisted Annotation of Japanese Grammatical Patterns and Their Senses for Corpus Construction

Kamila Szeląg-Levine¹, Steven Coyne^{1,2}, Tomotaka Kubota^{1,2}, Shinnosuke Nozue¹, Hong Hai Ngo¹, Machi Shimmei^{1,2}, Yuichiroh Matsubayashi^{1,2}, Keisuke Sakaguchi^{1,2}

kamila.szelaglevine@dc.tohoku.ac.jp

(¹ Tohoku University, ² RIKEN)

Summary

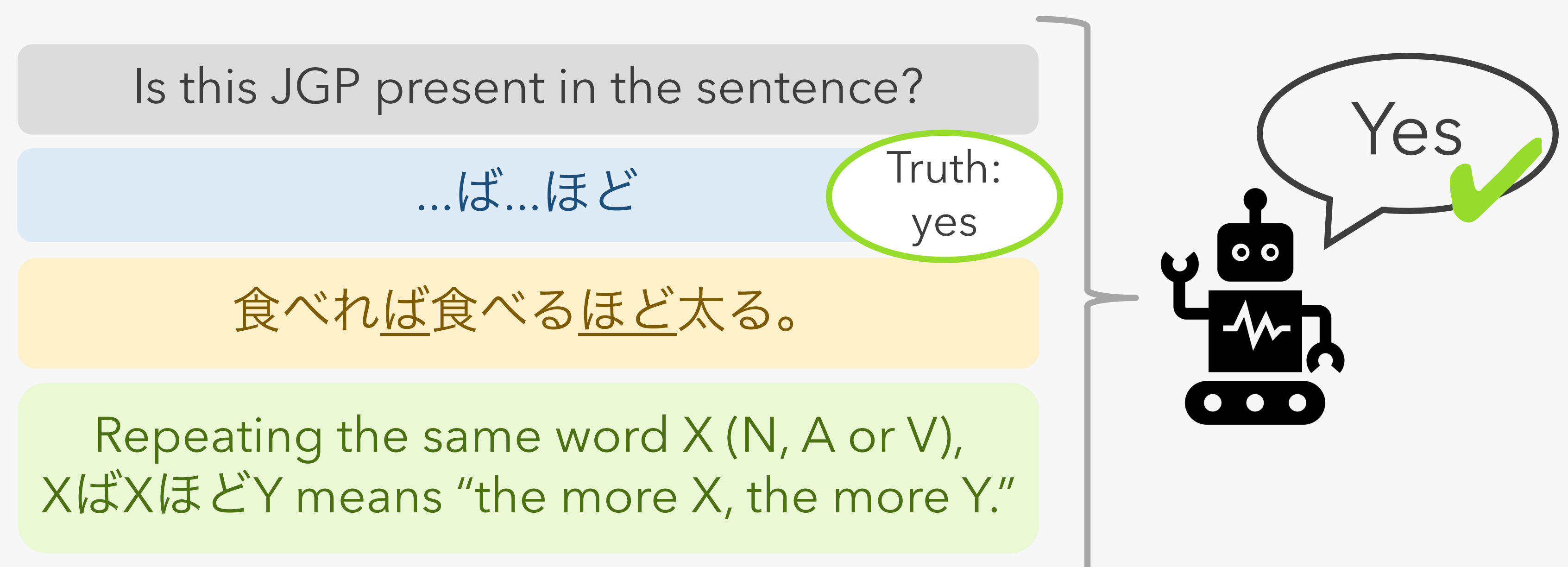
- We aim to build an LLM-assisted annotation pipeline of Japanese grammatical patterns (JGPs). As a prerequisite, we must ensure that LLMs can detect and annotate JGPs correctly.
- We investigate LLMs' ability to **(1) identify JGPs** and **(2) annotate their senses** in sentences.
- Results show that **LLMs can reliably detect simple JGPs but may struggle with nuanced ones.**

Motivation

- Japanese grammar patterns (JGPs)** can have several senses each — learning them can be difficult:
 - E.g. V/A+ことになる is used for a ① decision; ② restatement; ③ unfortunate outcome.
- Data-driven learning helps users learn senses from context by finding linguistic patterns in a corpus.
- No such system exists for JGPs, and creating one requires a **corpus of annotated sentences**.
- Manual annotation is resource-intensive — we develop an **LLM-assisted annotation pipeline**.

Experimental setup

- Selected 103 JGPs and 418 example sentences from a dictionary of Japanese grammar¹.
- Experiment 1: Ask the LLM to determine the presence of a JGP.**
 - Supplied with **sentence-JGP pairs** + pedagogical **description** of the JGP.



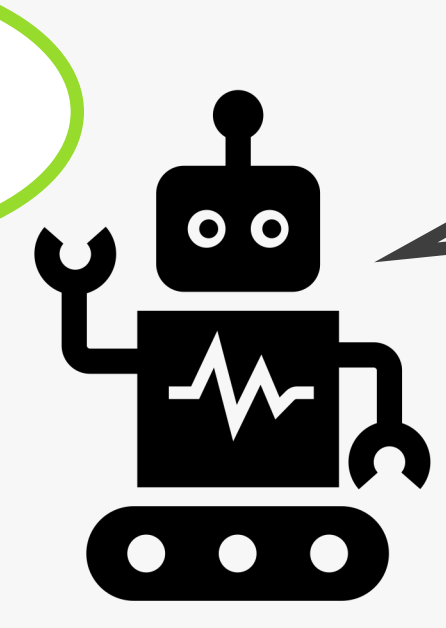
Preliminary results

- Tested cheaper models for viability at scale (gemini-3-flash, gpt-5.4-nano, gemma-4, gpt-oss, phi-4). Most performed very well (over 99% accuracy), apart from gpt-oss (92%) and phi-4 (84%).
- If models failed:
 - Medium/large models** → usually “overthought” the sense but detected the plain form correctly.
 - Small models** → also mismatched the output with their answer or provided no reasoning.

Nによって 〈agent in a passive sentence〉

その村の家の多くは洪水によって押し流された。
(Many of the houses in that village were swept away by the flood.)

Truth: yes



gemma-4

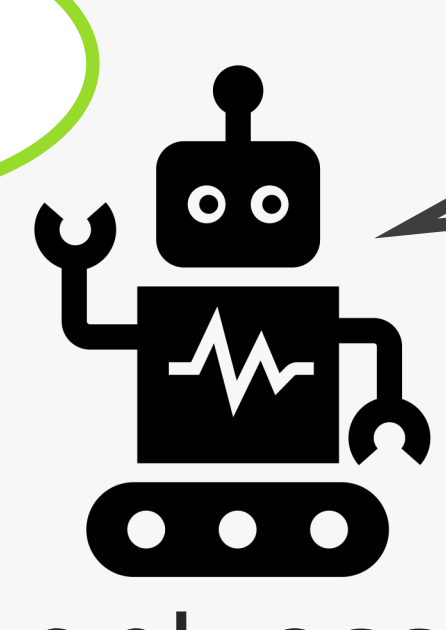
“In this sentence, “洪水” (flood) is the cause of the action, not the intentional agent...Does it act as the agent?...A “flood” is a natural phenomenon/cause, not a sentient agent performing the act of washing away houses...Wait, let me re-read carefully...”

“Overthinking”

あえて...ない

みんなに嫌がられてまで、あえて自分の方針を押し通すこともないじゃないか。
(There’s no need to risk pushing your own policy through when it would only make everyone dislike you, is there?)

Truth: yes



gpt-oss

“That is “あえて ... ない”. Thus output “Yes”.”

Output doesn’t match reasoning

Discussion

- Senses can be difficult even for humans, so a <1% error is comparable to human indecision.
- Most JGPs in this experiment were simple, perhaps inflating performance. We need **further testing with nuanced JGPs**, but medium-sized models can be **suitable for detecting simple ones**.
- Next, we will investigate LLMs' ability to span-annotate JGPs to create a corpus (Experiment 2).

¹ “A Handbook of Japanese Grammar Patterns for Teachers and Learners” (Group Jamassy, 2015)