Topicalization in Language Models: A Case Study on Japanese

Riki Fujihara [1], Tatsuki Kuribayashi [1,2], Kaori Abe [1], Ryoko Tokuhisa [1], Kentaro Inui [1,3]

[1] Tohoku University, [2] Langsmith Inc., [3] RIKEN

Overview

- Probed discourse-level linguistic knowledge of neural language models (LMs), focusing on **topicalization**
- Experiments showed **non-human-like**, **context-independent** behaviors of LMs on topicalization judgment

Topicalization: mark a concern of the message as topic

Topic is typically selected depending on context.

**wa* is the topic marker indicating topic

English

Context: *I broke a vase yesterday.*

A. *The vase was in the room.* more plausible!

B. *There was a vase in the room.*

Japanese (topic-prominent)

Context: *Kabin-o kinou wat-ta.*

A. *Kabin-wa* heya-ni at-ta.*

B. *Kabin-ga heya-ni at-ta.*

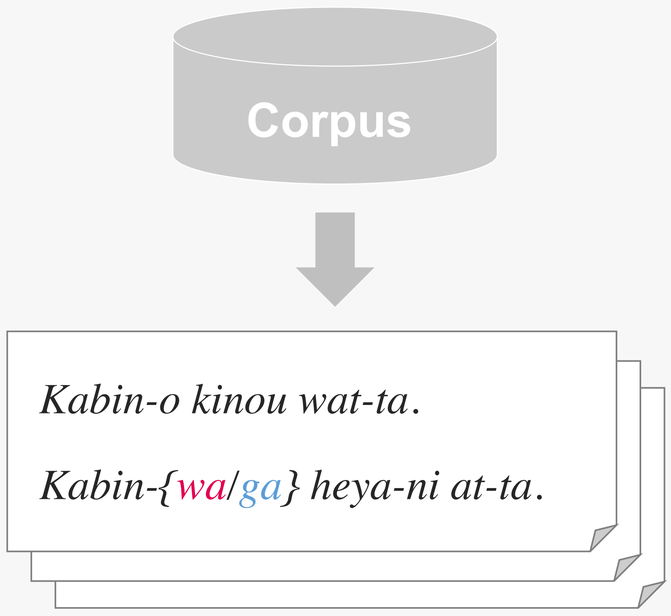
We used topic-prominent language for designing probing task

This enables creating minimally different text pairs w.r.t. topicalization easier

Do LMs capture such discourse-level behaviors?

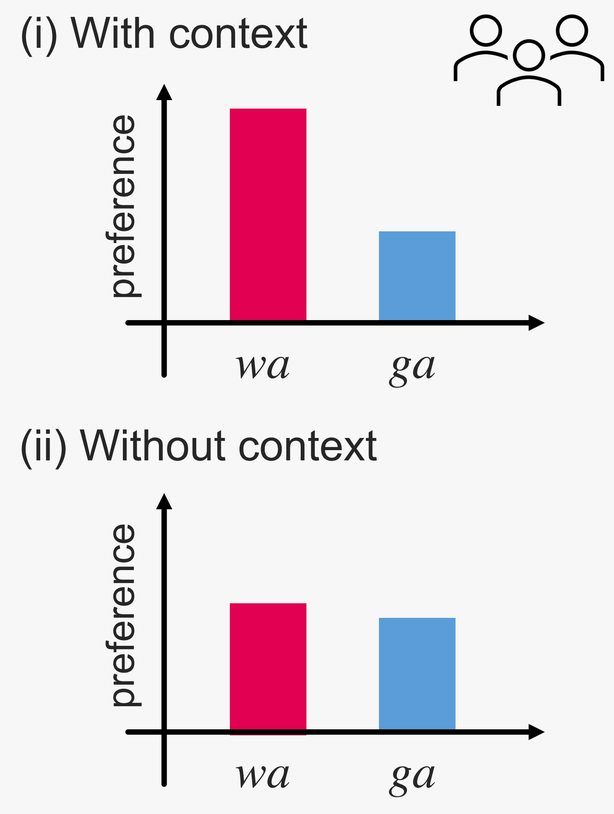
Probing paradigm

1. Create probing task

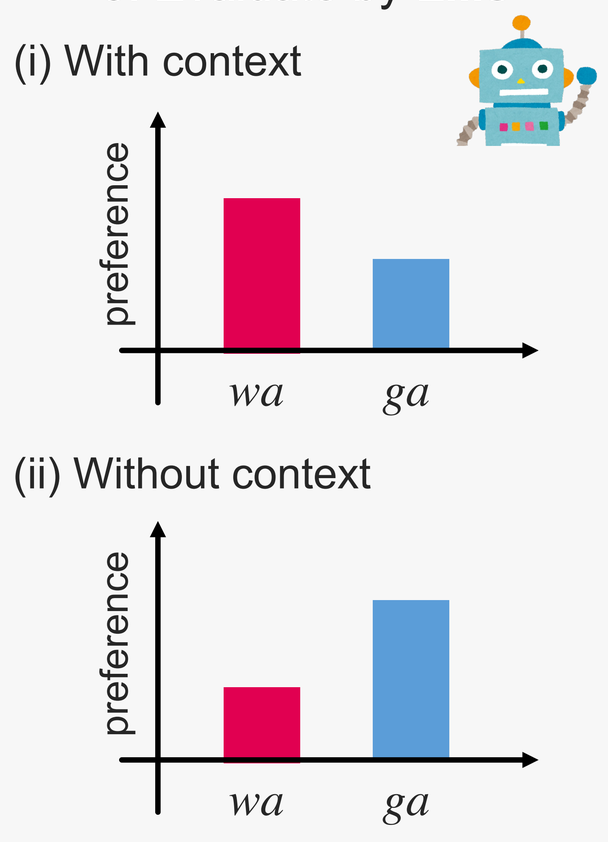


Question: should topicalize *Kabin*?
(Which is more plausible, *wa* or *ga*?)

2. Evaluate by **Humans**



3. Evaluate by **LMs**



4. Compare preference of Humans and LMs

Experiments

Language models

- TRANS-L (Transformer-based, 400M params)
- LSTM (LSTM-based, 55M params)

These models were trained with Japanese newspapers and Wikipedia

Scores

- Corr. r: correlation of topicalization preference in each setting
- Corr. Δ: correlation of context-dependent changes in topicalization preference by showing or not showing the context
- Macro F1: macro-averaged F1 score on selecting *wa* or *ga*

Model	Setting	Corr. r	Corr. Δ	Macro F1
Human	with context	-	-	(100)
	without	-		81.1
TRANS-L	with context	0.67	−0.12	83.5
	without	0.60		81.7
LSTM	with context	0.69	−0.20	81.9
	without	0.62		82.3

😀 Human showed context-dependent trends

😕 LMs showed **context-independent** trends
F1 score changed a little even in without setting

😕 LMs have **non-human-like** generalization
Corr. Δ was slightly negative