Enhancing parsing accuracy by fine-tuning language models with honorific awareness

Nayoung Kwon (Univ. of Oregon) & Seongmin Mun (Kyungpook National Univ.)

Pre-trained language models (PLMs), such as BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019), have driven significant advancements in natural language processing. However, languages like Korean, which account for less than 1% of web content, pose unique challenges for PLM training due to limited resources and data availability. This study focuses on fine-tuning Korean-specific BERT (KoBERT) and GPT-2 (KoGPT2) models to enhance their performance in parsing complex sentence structures. Specifically, we utilize subject-verb honorific agreement—a morphosyntactic feature specific in Korean.

Parsing embedded clauses in Korean (1) is particularly challenging because the subject of the embedded verb can be either NP1 or NP2, depending on the complementizers (e.g., subject control, object control, noun complement, etc.). The paucity of explicit morphological agreement further complicates resolution. To address this, we fine-tuned KoBERT and KoGPT2 using subject-verb honorific agreement. In Korean, honorification—a privative feature (Kim & Sells, 2007)—requires an honorific suffix licensed by an honorifiable subject, although the subject itself does not require the suffix (Kwon & Sturt, 2024). The training dataset included a 1:5 frequency ratio of NP1 to NP2 subject interpretations, reflecting the natural distribution of complementizers associated with NP1 and NP2 control.

Methodology: We employed two approaches. First, KoBERT and KoGPT2 were fine-tuned to classify sentences (2) as grammatical or ungrammatical. Training and evaluation were repeated 30 times with randomly sampled datasets. The classification results were compared against human acceptability judgments. Performance analysis using generalized linear mixed-effects models revealed GPT-2's superior adaptability ($\beta = -1.61$, $SE = 0.09$, $t = -16.43$, $p < .0001$), particularly in conditions involving honorific-marked subjects (4). In contrast, BERT did not show a significant improvement in performance ($\beta = 0.04$, $SE = 0.034$, $t = 1.21$, n.s.) and even exhibited a notable decline, particularly in the NH-NH condition.

Second, we examined attention weights to assess how honorific features influenced parsing. Attention weights from 12 heads across 12 layers were aggregated using a method inspired by Vig (2019), aligning tokenized outputs with original words. Sentences were categorized as successfully or unsuccessfully classified by both models. Relative attention rates for NP1 and NP2 were computed before and after fine-tuning. The analysis revealed that, as fine-tuning progressed, GPT-2 exhibited a shift in sensitivity to honorific suffix usage (5). Additionally, its attention focus transitioned from NP1 to NP2, reflecting the frequency distribution in the training dataset. In contrast, BERT showed an overall increase in attention weights after fine-tuning but did not demonstrate heightened sensitivity to honorific suffix usage in specific phrases.

Results and Implications: GPT-2 demonstrated robust improvements across most conditions, particularly in handling honorific agreement and successfully resolving structural ambiguities. In contrast, BERT exhibited limited adaptability, showing sensitivity to honorific features without acquiring the necessary structural rules. These findings highlight the importance of incorporating language-specific morphosyntactic features in fine-tuning pre-trained language models (PLMs) for underrepresented languages like Korean.

(1) NP1 NP2 embedded.verb main.verb

(2) NP1_HON/NON-HON NP2_HON/NON-HON embedded.verb_SUBJ.HON.suffix main.verb

(3) Human acceptability of sentences with embedded verbs marked with the subject honorific suffix

NP1	NP2	NP1 control	NP2 control
H	H	Yes	Yes
	NH	Yes	No
NH	H	No	Yes
	NH	No	No

(4) Mean classification success rates (SD) for BERT and GPT-2 at initial and final training stages

Model	NP1	NP2	NP1 control		NP2 control	
			Epoch 1	Epoch 30	Epoch 1	Epoch 30
BERT	H	H	89.1 (0.31)	86.5 (0.34)	86.5 (0.34)	86.1 (0.35)
		NH	81.3 (0.39)	74.1 (0.44)	76.8 (0.42)	79.2 (0.41)
	NH	H	45.3 (0.5)	61.8 (0.49)	82.8 (0.38)	84.4 (0.36)
		NH	84.2 (0.37)	79.6 (0.4)	84.6 (0.36)	78.5 (0.41)
GPT-2	H	H	93 (0.25)	98.4 (0.13)	94.3 (0.23)	98.8 (0.11)
		NH	69.5 (0.46)	79.6 (0.4)	92.5 (0.26)	97.1 (0.17)
	NH	H	52.4 (0.5)	72.8 (0.45)	86.6 (0.34)	94.4 (0.23)
		NH	98.4 (0.13)	99.4 (0.08)	97 (0.17)	97.8 (0.15)

(5) Attention weight analysis results

Min-Max analysis of attention weights in correctly classified results	Min-Max analysis of attention weights in incorrectly classified results