Do language models know how to be polite?

Soo-Hwan Lee
Department of Linguistics
New York University
soohwan.lee@nyu.edu

Shaonan Wang
Department of Psychology
New York University
shaonan.wang@nyu.edu

1 Introduction

Politeness is often associated with a degree of formality that the speaker conveys to the addressee of a conversation. There are multiple ways to convey politeness in natural language. Languages such as Korean and Japanese, for instance, have politeness markers that appear in certain positions inside a given sentence. Sometimes, the absence of these politeness markers leads to inappropriateness. This work focuses on a particular case in which a politeness marker can be realized only when its dependency requirement is satisfied. While language model (LM) performance on syntactic dependencies such as filler-gap dependencies (Wilcox et al., 2018), subject-verb agreement (Linzen et al., 2016), anaphor binding (Hu et al., 2020), and control phenomena (Lee and Schuster, 2022) have been explored in recent years, little work has been done on non-syntactic dependencies that reflect politeness or even pragmatic effects in general. The phenomenon at issue is unique in that the dependency is not fulfilled by any of the commonly assumed syntactic disposition or agreement patterns observed elsewhere in the human grammar. For instance, a politeness-denoting possessor pronoun (i.e. *cey* 'my (polite)' or *cehuy* 'our (polite)') can be licensed by the sentence final politeness marker *yo* as shown in (1). Note that this does not resemble subject-verb agreement in that a *non-head* of a noun phrase (e.g. possessor) can participate in the licensing condition. The abbreviations used in this work are given in Appendix A.

(1) **Cey** ai-eykey si-lul ilke-cwu-e-**yo**.
my.POL child-DAT poem-ACC read-give-D-POL
'Please read the poem to my child.'

In the absence of *yo*, the presence of a politeness pronoun is not possible as in (2) (# = inappropriate).

(2) **#Cey** ai-eykey si-lul ilke-cwu-e.
my.POL child-DAT poem-ACC read-give-D
Intended: 'Please read the poem to my child.'

We create minimal pair sentences such as (1) and (2) for our experiment. Our results suggest that the overall performance of the Transformer-based LMs such as GPT-2 and the variants of BERT on this dependency test is unexpected. Since their performance is below or around chance accuracy for the main task of our experiment, we posit that these pretrained LMs fail to fully capture the politeness phenomenon in Korean. The performance of ChatGPT on a related task, however, is significantly better than its predecessors. While it is tempting to conclude that ChatGPT is better suited for capturing this specific phenomenon, we show that the model is right for the wrong reason. We demonstrate that the model merely selects the sentence that ends with the politeness marker *yo*, instead of recognizing the true dependency between the cue (e.g. *cey* 'my (polite)') and the target (*yo*). This may be attributed to the way in which the prompt is addressed to ChatGPT. We use the word 'appropriate' to state the questions in our prompt. It is likely that the LM simply associates the meaning of 'appropriateness' with politeness without taking the linguistic dependency into full consideration. This suggests that ChatGPT does not focus on the dependency as much as it should. Further progress needs to be made in terms of enabling data-driven LMs to pick up on this type of dependency observed in politeness contexts.

2 The dependency of politeness

First person pronouns in Korean are morphologically sensitive to the discourse effect of politeness. For first person singular pronouns, the alternation between the plain form *nay* 'my' and the politeness form *cey* 'my (polite)' is possible. For first person plural pronouns, the alternation between the plain form *wuli* 'our' and the politeness form *cehuy* 'our (polite)' is possible. Similar to *cey* 'my (polite)' in (1) and (2), *cehuy* 'our (polite)' can co-occur with the sentence final politeness marker *yo*. The two

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 375-378.
Amherst, Massachusetts, June 15-17, 2023

first person pronouns pattern alike with respect to the dependency requirement:

(3) **Cehuy** ai-eykey si-lul ilke-cwu-e-**yo**.
our.POL child-DAT poem-ACC read-give-D-POL
'Please read the poem to our child.'

(4) **#Cehuy** ai-eykey si-lul ilke-cwu-e.
our.POL child-DAT poem-ACC read-give-D
Intended: 'Please read the poem to our child.'

In (3), the co-occurrence of *cehuy* and *yo* makes the sentence appropriate. In (4), the presence of *cehuy* in the absence of *yo* makes the sentence inappropriate. (3) and (4) together show that the politeness pronoun *cehuy* is linguistically dependent on the sentence final marker *yo*.

3 Stimuli & design

Using both of the politeness-sensitive pronouns *cey* and *cehuy*, we generate minimal pairs for our experiment. We adapted Hu et al.'s (2020) and Lee and Schuster's (2022) experimental design for testing LMs on linguistic dependencies. One of our minimal pairs is provided below. (5) is the most appropriate sentence in (5)–(7) (*la* in (7) indicates strong imperative):

(5) **Cey** ai-eykey si-lul ilke-cwu-e-**yo**.
my.POL child-DAT poem-ACC read-give-D-POL
'Please read the poem to my child.'

(6) **#Cey** ai-eykey si-lul ilke-cwu-e.
my.POL child-DAT poem-ACC read-give-D
Intended: 'Please read the poem to my child.'

(7) **#Cey** ai-eykey si-lul ilke-cwu-e-la.
my.POL child-DAT poem-ACC read-give-D-IMP
Intended: 'Please go read the poem to my child.'

We consider the word containing *yo* as the target word. We expect that an LM would assign a higher probability to the target word in (5) than to its counterparts in (6) and (7). Note that neither (6) nor (7) is the most appropriate sentence. Randomly assigning the highest probability to any one of the three target words results in chance accuracy which is 0.33. This is the case for all of the minimal pairs in our main condition.

Since LM performance on long-distance dependencies has been considered to be mixed (Marvin and Linzen, 2018; Futrell et al., 2019; Hu et al., 2020), we also examine simple cases where the cue word and the target word are only separated by a single word. In the minimal pair below, (8) is the most appropriate sentence in (8)–(10). Once again, chance accuracy is 0.33 for each minimal pair in the simple condition.

(8) Ai-eykey **cey** si-lul ilke-cwu-e-**yo**.
child-DAT my.POL poem-ACC read-give-D-POL
'Please read my poem to the child.'

(9) #Ai-eykey **cey** si-lul ilke-cwu-e.
child-DAT my.POL poem-ACC read-give-D
Intended: 'Please read my poem to the child.'

(10) #Ai-eykey **cey** si-lul ilke-cwu-e-la.
child-DAT my.POL poem-ACC read-give-D-IMP
Intended: 'Please go read my poem to the child.'

15,360 minimal pairs are generated for the simple and main (non-simple) conditions, respectively. There are three sentences in each minimal pair and only the sentence that hosts the target word *yo* is appropriate (chance accuracy = 0.33). If an LM assigns the highest probability to the appropriate target word with *yo*, then it gets a score of 1. If an LM assigns the highest probability to either one of the two inappropriate target words without *yo*, then it gets a score of 0. The accuracy of an LM is calculated by the score it receives divided by the total number of minimal pairs.

For ChatGPT, we randomly sample 120 minimal pairs for the main and simple conditions respectively and tested the model performance with the following prompt: '다음 중 가장 적절한 문장은 어떤 것인가? 왜 그 문장이 가장 적절한가?' (English translation: Which sentence is the most appropriate? Why is that sentence the most appropriate?). We use the word 'appropriate' instead of 'grammatical', since politeness has not been characterized as a purely grammatical phenomenon such as subject-verb agreement. The accuracy of ChatGPT is calculated by dividing the number of its responses specifying that the sentence with *yo* is the most appropriate by the total number of minimal pairs used in the experiment.

4 Language models

We evaluate how well the LMs perform on our main and simple conditions. Specifically, we examine the latest Transformer-based LMs trained on Korean, including the left-to-right generative language model GPT-2 and bidirectional encoders BERT, ALBERT, and RoBERTa. In addition to the uni/bidirectionality of how the probability of a word is assigned, one major difference between these LMs is their training objectives. GPT-2 is trained using an objective of predicting the next word in a sequence. BERT is trained using an objective of predicting masked words in a sentence. ALBERT is a memory-efficient and faster version of BERT using parameter-sharing

techniques. RoBERTa uses an extra training technique called "dynamic masking" to improve the model's ability to handle out-of-vocabulary words. All of these language models are pre-trained on a 70GB corpus including Korean Wikipedia, commerce reviews, blog websites, etc. They can be downloaded from the 'Pretrained Language Models For Korean' project at https://github.com/kiyoungkim1/LMkor. In the experiment, we used these pre-trained models to calculate the probability of the target words in our stimuli. Codes to utilize these models and generate the probability of the target word can be found at https://github.com/wangshaonan/Korean_politeness.

We also evaluate ChatGPT's performance. ChatGPT is built based on the same transformer architecture as GPT-2, which is a neural network that uses self-attention mechanisms to process natural language input and generate text output. However, ChatGPT incorporates additional improvements and is based on a larger and more diverse training dataset, enabling it to generate more sophisticated and contextually-appropriate responses than its predecessor. We used the web version of ChatGPT (January 9 and January 30 Version in 2023) at https://chat.openai.com/.

5 Results

5.1 GPT-2 and BERT-based models

Figure 1 shows that all LMs achieve below or around chance level accuracy (0.33) on our main condition. Only BERT and ALBERT achieve significantly better mean accuracy on the simple condition than on the main condition. These results indicate that these LMs fail to fully capture the politeness phenomenon in Korean. Moreover, this suggests that the degree of linear adjacency plays a role on LM performance.

5.2 ChatGPT

Compared to the other models, significantly higher accuracy is obtained by ChatGPT on a similar task. ChatGPT's accuracy is around 0.96 on the main condition and 0.99 on the simple condition as shown in Figure 2. While the results on the surface seem to suggest that ChatGPT is somehow recognizing the dependency requirement, the results from a non-politeness-related dependency task suggest otherwise. Crucially, in cases where the politeness marker *yo* has to be absent due to a non-politeness cue word (e.g. Siwu-ya 'hey Siwu') as shown in (12) and (13), ChatGPT often selects the sentence ending with *yo* as the most appropriate sentence within a given minimal pair. A minimal pair consists of three sentences as illustrated in (11)–(13).

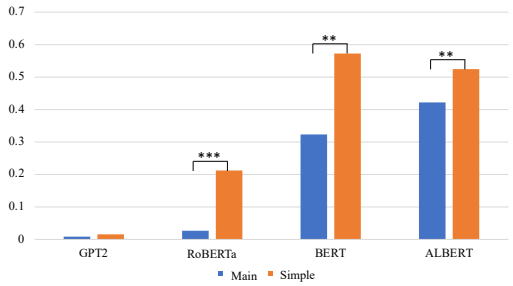

Figure 1: Accuracy of different language models on main and simple conditions. ** indicates p<0.05 and *** indicates p<0.01.

(11) #Siwu-**ya** nay ai-eykey si-lul
Siwu-N.POL my child-DAT poem-ACC
ilke-cwu-e-**yo**.
read-give-D-POL
Intended: 'Hey Siwu, please read the poem to my child.'

(12) Siwu-**ya** nay ai-eykey si-lul
Siwu-N.POL my child-DAT poem-ACC
ilke-cwu-e.
read-give-D
'Hey Siwu, read the poem to my child.'

(13) Siwu-**ya** nay ai-eykey si-lul
Siwu-N.POL my child-DAT poem-ACC
ilke-cwu-e-**la**.
read-give-D-IMP
'Hey Siwu, go read the poem to my child.'

Either (12) or (13) is the most appropriate sentence depending on the context. Crucially note that (11) cannot be the most appropriate due to the anti-dependency between the non-politeness cue word and *yo*. Here, chance accuracy is 0.66. Out of 120 minimal pairs, ChatGPT performed well only on 9 minimal pairs (0.08). ChatGPT's performance is significantly below chance level accuracy as shown in Figure 2.

Overall, the LMs used in our experiments fail to fully recognize the linguistic dependency between the cue word and the target word sensitive to politeness and non-politeness contexts.

6 Discussion

Our results suggest that all LMs used in the study fail to highlight the unique dependency requirement necessary for expressing politeness in Korean. The

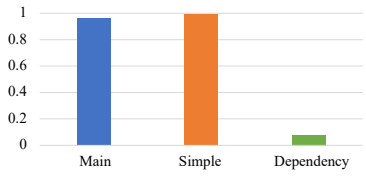

Figure 2: Accuracy of ChatGPT on main and simple conditions plus a non-politeness-related dependency task.

results from ChatGPT suggests that the model performs well for the wrong reason. Based on the responses from ChatGPT, it is plausible to assume that the LM generally associates the meaning of 'appropriateness' with politeness. Note that the non-politeness-related dependency task in our experiment provides us the opportunity to distinguish appropriateness from politeness. In (11)–(13), for instance, the sentence with a politeness marker is inappropriate. While the experimental design is enough to show that ChatGPT does not fully attend to the linguistic dependency at issue, it may be possible for ChatGPT to tune into the dependency based on a different prompt. For future research, it may be worth stating the prompt in slightly different ways. One way to do this is to replace words such as 'appropriate' with some other contentful words. Using different prompts may shed light on where the potential challenges lie. Another point worth mentioning is that the linguistic phenomenon discussed in this work is not observed in all languages. Nevertheless, it would be meaningful to examine the languages that host this type of dependency and to see whether the overall results reported in this work can be replicated and generalized. Overall, future work remains to be done on how data-driven LMs can correctly capture this type of dependency especially in an experimental setting.

Acknowledgements

We would like to thank Alec Marantz and the anonymous reviewers for their thoughtful comments.

References

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Hu, Sherry Yong Chen, and Roger Levy. 2020. A closer look at the performance of neural language models on reflexive anaphor licensing. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 323–333, New York, New York. Association for Computational Linguistics.

Soo-Hwan Lee and Sebastian Schuster. 2022. Can language models capture syntactic associations without surface cues? A case study of reflexive anaphor licensing in English control constructions. *Proceedings of the Society for Computation in Linguistics*, 5(1):206–211.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

A Abbreviations

ACC accusative case	**IMP** imperative
D default	**N.POL** non-polite
DAT dative case	**POL** polite