Evaluation of Empathy Detection Performance of Deep Learning Models using Theory-of-Mind levels and COVID-19 Emotion-Diary Corpus

Yoon Kyung Lee*, Inju Lee, Jae Eun Park, Sowon Hahn

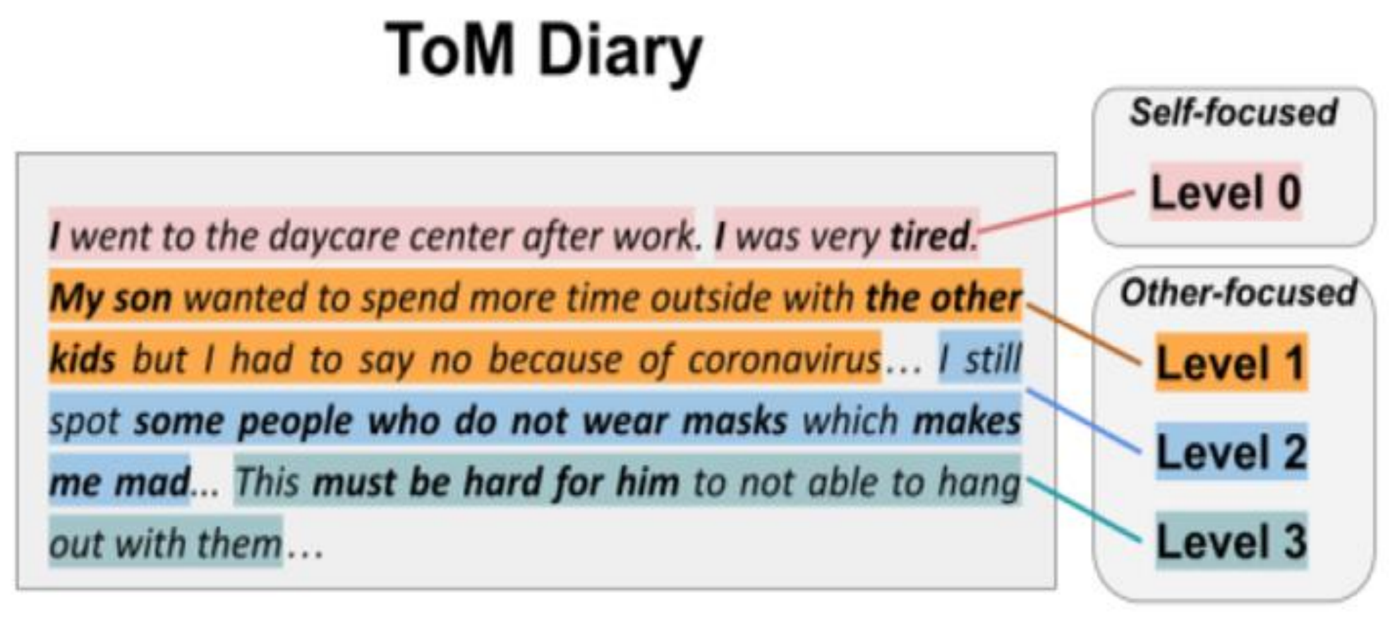
Human Factors Psychology Lab, Dept. of Psychology, Seoul National University

How to create empathetic AIs?

Background

- Theory of Mind is an essential cognitive skill that precedes empathy
- We trained deep learning models to test a ToM skill of a pre-trained deep learning model by fine-tuning a ToM labeled corpus
- We also created a ToM-Diary dataset, an emotion diary corpus labeled with 4 ToM levels.

Methods



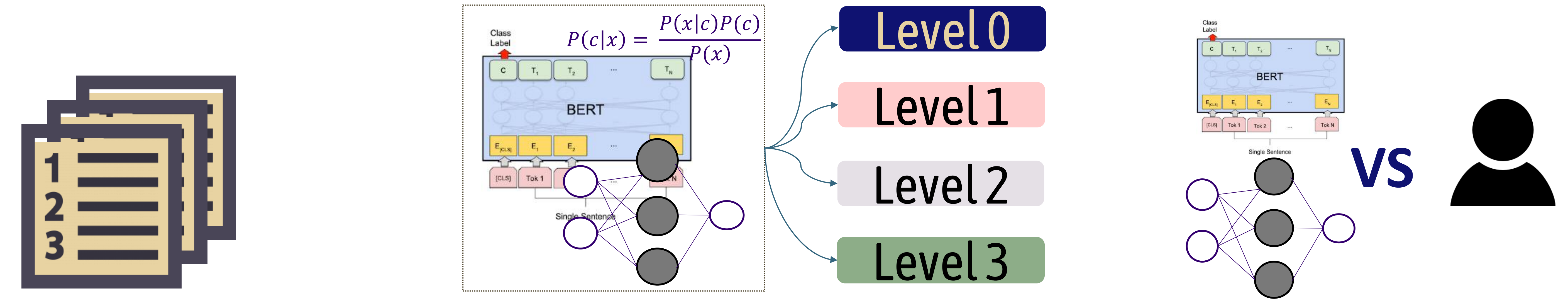
ToM Diary

I went to the daycare center after work. I was very tired. My son wanted to spend more time outside with the other kids but I had to say no because of coronavirus… I still spot some people who do not wear masks which makes me mad… This must be hard for him to not able to hang out with them…

Diary Entry → Annotate → Review

Thirty psychology students annotated the data and five psychologists reviewed the labeled data (N_{sent} =74,014). Annotators and reviewers showed substantial agreement (Cohen's kappa = .7). The average number of labels per diary was 2.94 (SD=2.15, $N_{diaries}$= 19,205).

Procedure



Train

- **ToM-annotated sentences (19k)**
  - Balanced vs Imbalanced set
  - Train/Validation/Test = 9:1:1
  - Pre-trained Word2Vec (N_{docs}= 19,205) for MNB, FFNN, Bi-LSTM
  - All POS Tags vs Core POS tags only

ToM Prediction

- **Machine Learning Classifiers (4)**
  - Multinomial Naïve Bayes (MNB)
  - Feedforward Neural Network (FFNN)
  - Bidirectional LSTM (Bi-LSTM)
  - Bidirectional Encoder Representations from Transformers (BERT)

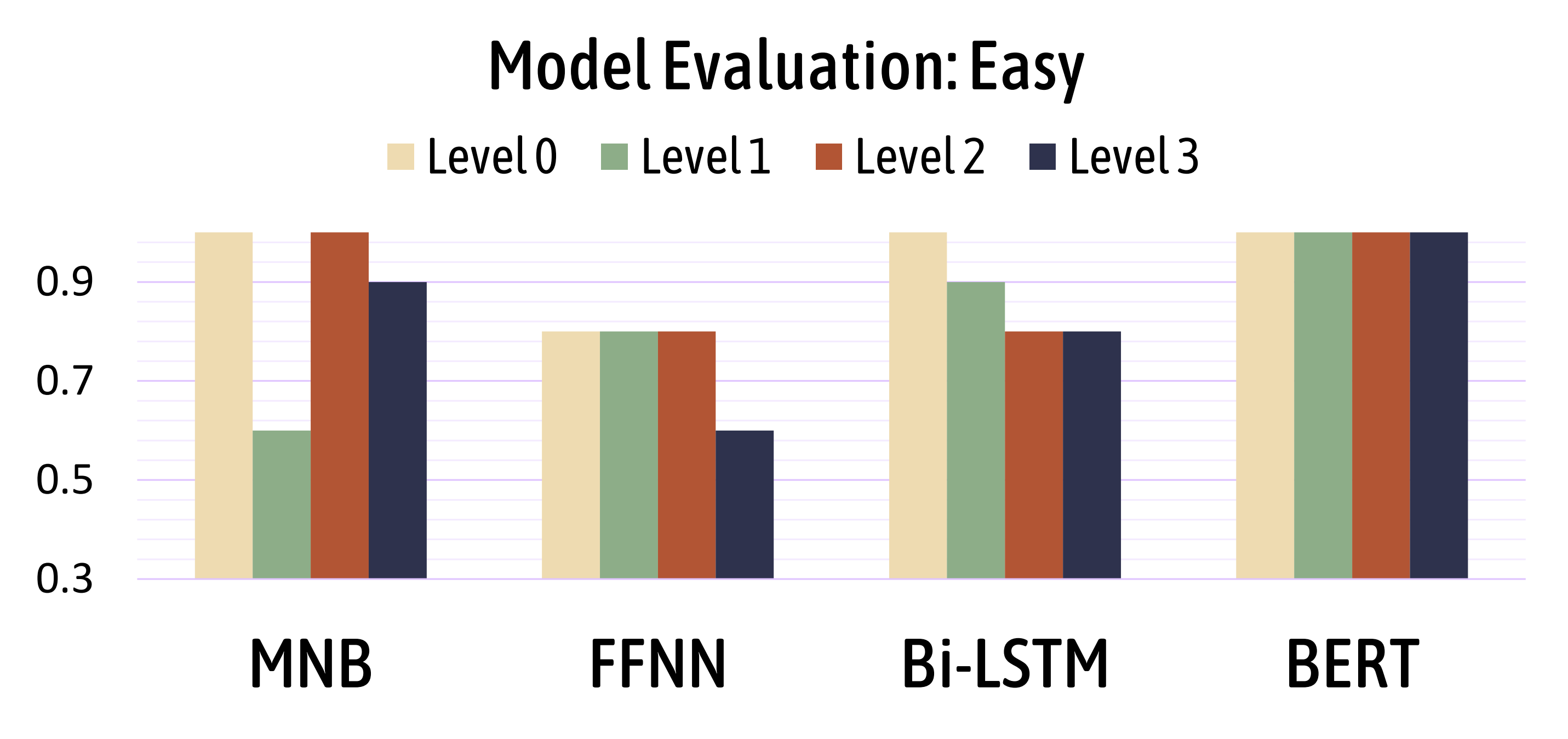
Evaluation

- **Easy vs Hard (N_{sent}= 80)**
  - Syntactic Ambiguity
  - Fictional characters, and animals, "the virus"
  - Intention (e.g., Sarcasm)

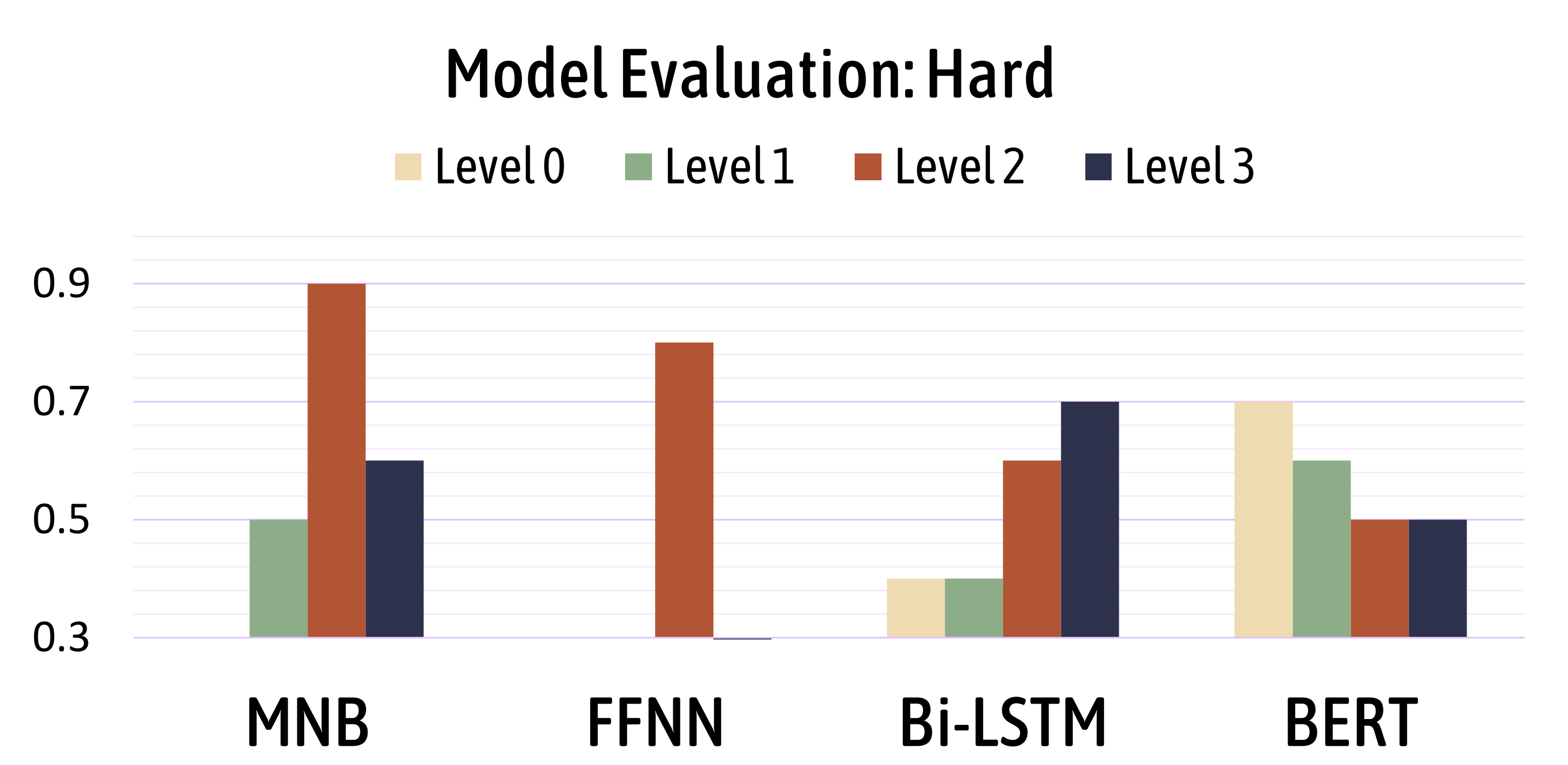
Results

		MNB	FFNN	Bi-LSTM	BERT
Precision		0.64	0.56	0.73	0.78
Recall		0.64	0.56	0.73	0.78
F1 Score		0.64	0.56	0.73	0.78
Acc	Level 0	0.63	0.72	0.85	0.89
	Level 1	0.52	0.47	0.59	0.76
	Level 2	0.83	0.64	0.75	0.75
	Level 3	0.55	0.42	0.73	0.72

- BERT classifier more successfully predicted the ToM level than the other three models.
- Bi-LSTM classifier sometimes classified level 3 when trained with all POS better than BERT. This suggests the overall context of the sentence should be maintained to judge whether the writer tried to infer others' mental state or not.
- Adding train data (up to 8,000 sentences) did not enhance the performance. The delicate nuances of the sentences are crucial for classifying level 3 correctly, and it can be detected by other POS except for core POS, such as postpositions and interjections.



Model Evaluation: Easy



Model Evaluation: Hard

Findings

- Except for sentences with syntactic ambiguity, sarcasm, and non-human subjects, BERT showed the best performance.
- Even when empathizing with others, people use level 2, "refuse to accept other's perspective".

Conclusions

- **People use different levels of ToM in the process of empathizing others.**
- **AIs should learn ToM skills to accurately understand and predict human emotion and intention.**

Resources

Lee, YK., Lee, I., Park, J. E., Jung, Y., Kim, J., & Hahn, S. (2021). A Computational Approach to Measure Empathy and Theory-of-Mind from Written Texts. *arXiv preprint arXiv:2108.11810*. [arXiv]

Lee, YK., Jung, Y., Lee, I., Park, J. E., & Hahn, S. (2021). Building a Psychological Ground Truth Dataset with Empathy and Theory-of-Mind During the COVID-19 Pandemic. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43.* https://escholarship.org/uc/item/950900w7

humanfactorspsych

hfpsych.snu.ac.kr

*email: yoonlee78@snu.ac.kr

2021 The Korean Society for Cognitive Science