How do Transformer-Architecture Models Address Polysemy of Korean Adverbial Postpositions?

Seongmin Mun[1] & Guillaume Desagulier[2]
Ajou University[1], Paris VIII University & UMR 7114, MoDyCo[2]

Introduction

- **A Korean postposition** normally involves many-to-many associations between form and function. As such a postposition is polysemous. For example, an adverbial postposition *-(u)lo* (-ulo after a consonant) is interpreted as six major functions: criterion (CRT), direction (DIR), effector (EFF), final state (FNS), instrument (INS), and location (LOC) (Shin, 2008). For instance, the following sentence involving the postposition *-(u)lo* as a marker of INS (instrument) as in (1).

(1) ***-(u)lo* as INS (instrument)**

na-nun	kamca-lul	khal-lo	ssel-ess-ta.
I-TOP	potato-ACC	knife-INS	cut-PST-DECL

'I cut a potato with a knife.'

- **Contextualized word-embedding model:** The model considers neighborhood information about a polysemous word on the basis of sequences of words around the target word.

- **Background:** Several studies have used transformer-architecture models to address the word-level polysemy of Korean adverbial postpositions (e.g., Bae et al., 2020). Notably, the particular reason for the transformer architecture's superior performance over the others is somewhat unclear.

- **Question:** *How do Transformer-Architecture Models Address Polysemy of Korean Adverbial Postpositions?*

Methods

- **Input:** A portion of Sejong corpus (Shin, 2008), with semantic annotations of postpositions *-ey, -eyes,* and *-(u)lo* cross-verified by three native speakers of Korean (κ = 0.948 (*-ey*), 0.928 (*-eyse*), and 0.947 (*-(u)lo*)).

-ey		*-eyse*		*-(u)lo*	
Function	Frequency	Function	Frequency	Function	Frequency
LOC	1,780	LOC	4,206	FNS	1,681
CRT	1,516	SRC	647	DIR	1,449
THM	448			INS	739
GOL	441			CRT	593
FNS	216			LOC	158
EFF	198			EFF	88
INS	69				
AGT	47				
Total	4,715	Total	4,853	Total	4,708

- **Classification models:** We devised a classification model by employing Bidirectional Encoder Representations from Transformer (BERT; Devlin et al., 2018) and Generative Pre-Training 2 (GPT-2; Radford et al., 2018)

- **Visualization system:** In order to better understand how BERT and GPT-2 recognize the word-level polysemy, we developed a visualization system by using the test set under the two-dimensional distribution (i.e., t-distributed Stochastic Neighbor Embedding; Maaten and Hinton, 2008).

Methods

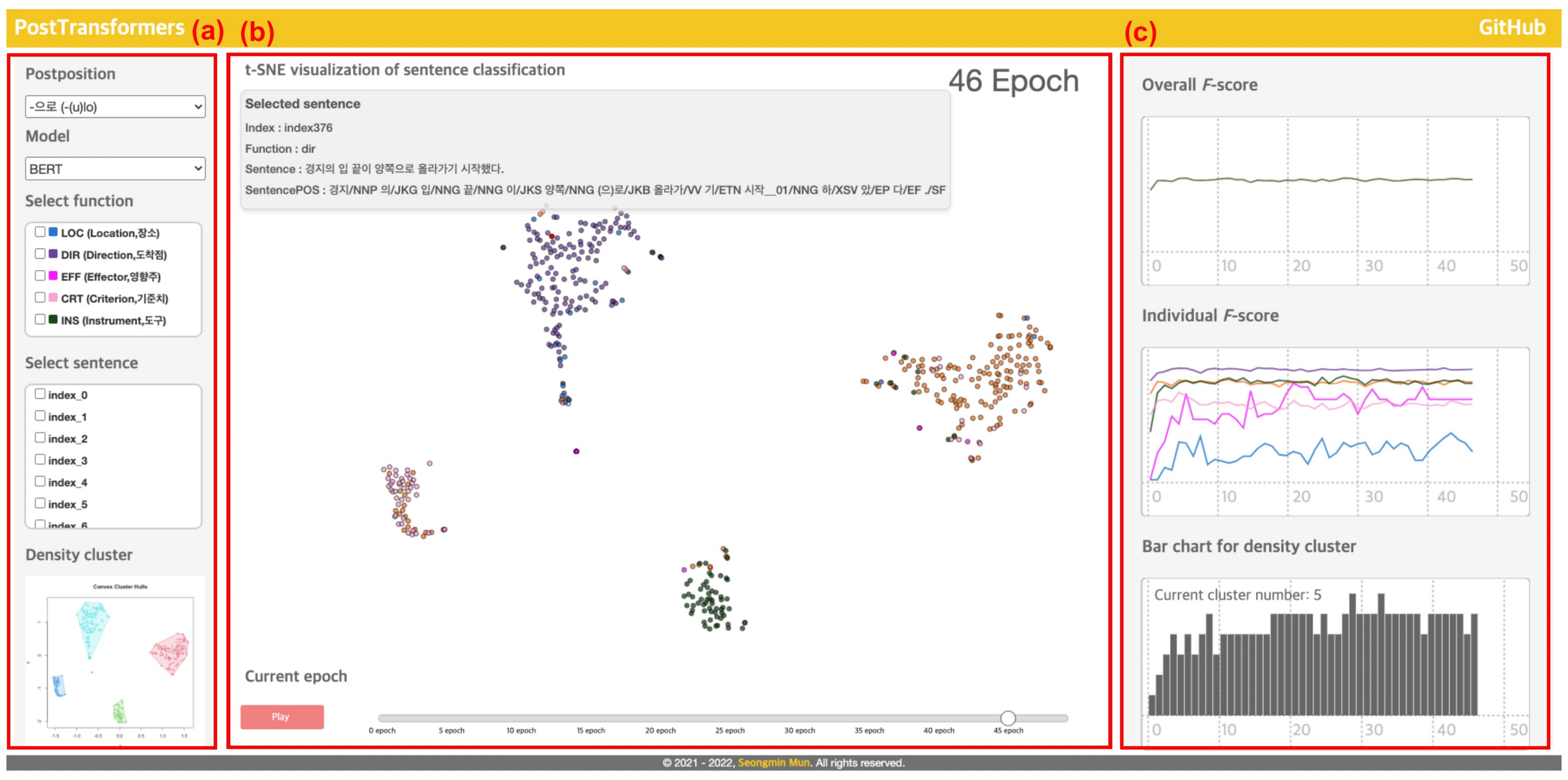

Figure 1: The overview interface of the visualization system (available at: https://seongmin-mun.github.io/Visualization/2022/PostTransformers/index.html)

Results

- **Findings**
  - ✓ First, BERT performed better than GPT-2 in revealing the polysemy of Korean postpositions (BERT: 0.744 for *-ey*, 0.875 for *-eyse*, 0.795 for *-(u)lo;* GPT-2: 0.68 for *-ey*, 0.844 for *-eyse*, 0.676 for *-(u)lo*).
  - ✓ Second, there was an inverse relation between the classification performance and the number of functions of each postposition.
  - ✓ Third, the model was affected by the corpus size of each function.
  - ✓ Fourth, the model was able to identify the intended functions of a postposition as the epoch progressed (see Figure 2).
  - ✓ Fifth, these models were affected by the rarely occurring input and/or semantic closeness between the items, limiting the performance of two models in the given task to some extent.

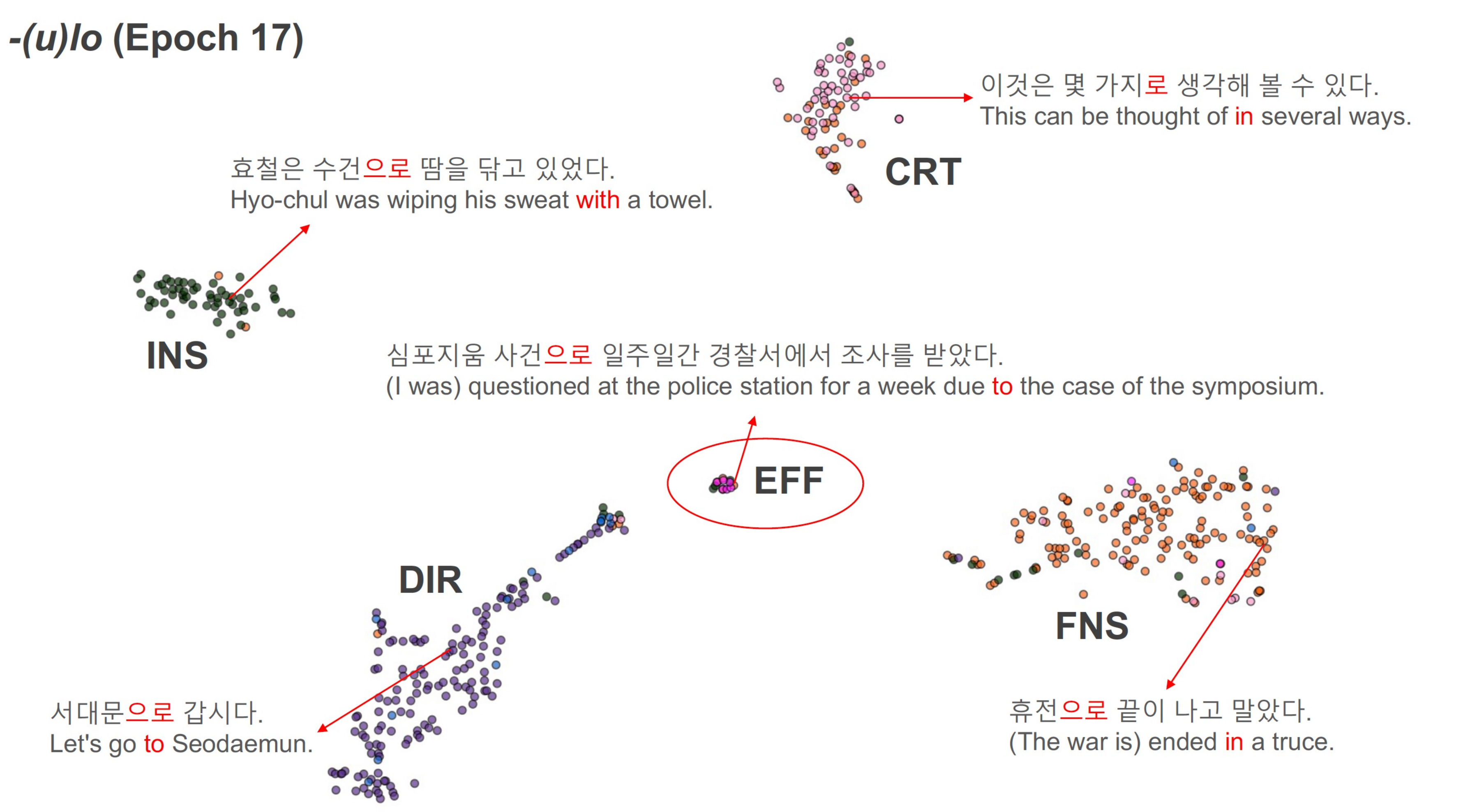

Figure 2: The t-SNE outcome of BERT model for *-(u)lo* in Epoch 17

REFERENCES

- Jangseong Bae, Changki Lee, Soojong Lim, and Hyunki Kim. 2020b. Korean semantic role labeling with bert. The Korean Institute of Information Scientists and Engineers, 47(11):1021–1026.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(Nov):2579–2605.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.
- Hyo-pil Shin. 2008. The 21st sejong project : with a focus on selk (sejong electronic lexicon of korean) and the knc (korean national corpus). In In The 3rd International Joint Conference on Natural Langauge Processing.

Presented at the ALPS 2023 (16-20 JAN 2023), Corresponding Author: Seongmin Mun (*stat34@ajou.ac.kr*)
This work was supported by Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022S1A5C2A02090368)