Polysemy resolution with word embedding models and data visualization

the case of adverbial postpositions *-ey*, *-eyse*, and *-(u)lo* in Korean

Seongmin Mun

Université Paris Nanterre & UMR 7114, MoDyCo

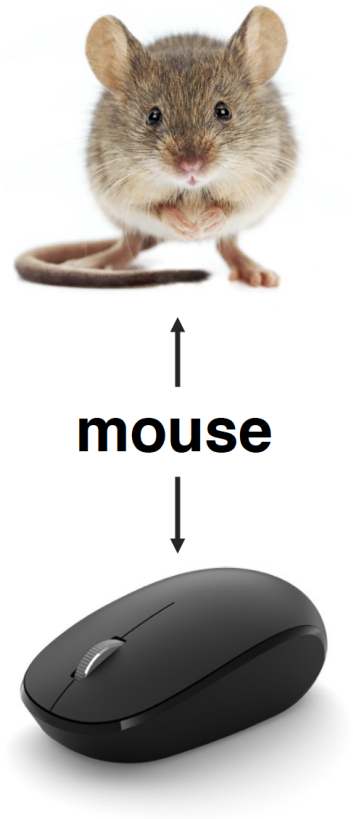
18 June 2021

Outline

Introduction

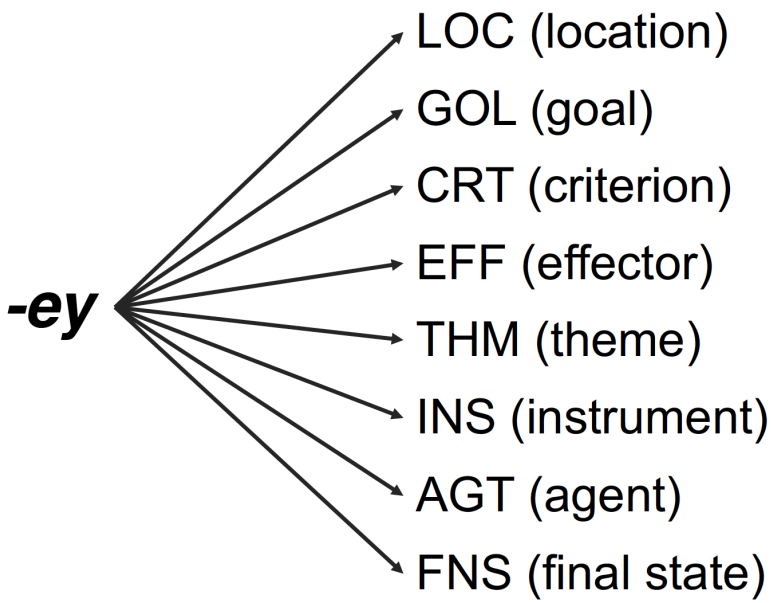
Polysemy

Polysemy, one type of ambiguity, occurs when one form delivers multiple meanings/functions (Glynn and Robinson, 2014).



Korean language

Korean is a Subject-Object-Verb language, which marks grammatical information with dedicated postpositions (Sohn, 1999).



Polysemy in Korean adverbial postposition

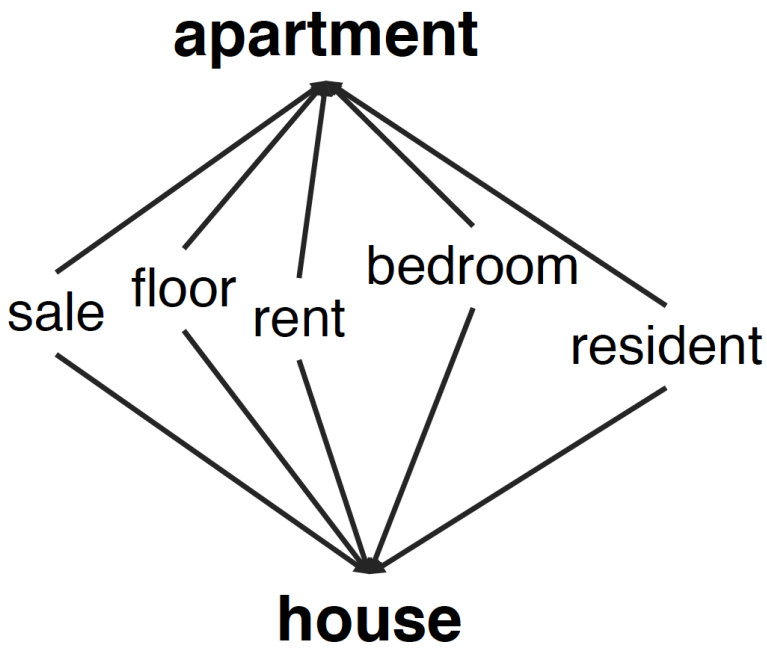
지붕 위에 구멍이 났다.
cipung wi-ey kwumeng-i na-ss-ta.
Roof top-LOC hole-NOM appear-PST-DECL
'There is a hole on the top of the roof.'

Figure: An example sentence involving the postposition *-ey* as a function of LOC (location)

Question: How a speaker can understand the function of postposition?

Concept of DSMs

The concept of distributional semantic models (DSMs) is that **a word meaning is closely tied to a context** that is created by a group of neighborhood words, dubbed the distributional hypothesis (Firth, 1957; Harris,1954).



Previous studies on adverbial postpositions

Study	Corpus type	Data size	Method	Accuracy
Bae et al. (2015)	Korean PropBank	4,882 sentences	One-hot encoding + Structural SVM & FFNN (Feed-Forward Neural Network)	0.75
Kim & Ock (2016)	Sejong corpus	59.220 sentences	One-hot encoding + CRF (Conditional Random Fields Model)	0.83
Lee et al. (2015)	Korean PropBank	4,882 sentences	Word2vec (SGNS) + Structural SVM (Support Vector Machine)	0.77
Mun & Shin (2020)	Sejong corpus	2,100 sentences	PPMI & SVD + Similarity-based estimate	0.74
Park & Cha (2017)	Sejong corpus	14,335 sentences	Word2vec (SGNS) + CRF	0.77
Shin et al. (2005)	Sejong corpus	4,355 sentences	Word token-based embedding + SVM	0.71
Yoon et al. (2016)	Korean PropBank	4,714 sentences	One-hot encoding + Bidirectional LSTM-CRFs	0.66

DSMs that I used

- Word-level embedding model
  - Count-based model: a combination of Positive Pointwise Mutual Information (PPMI; Church and Hanks, 1989) and Singular Value Decomposition (SVD; Eckart and Young, 1936)
  - Prediction-based model: Skip-Gram and Negative Sampling (SGNS; Mikolov et al., 2013)

DSMs that I used

- Word-level embedding model
  - Count-based model: a combination of Positive Pointwise Mutual Information (PPMI; Church and Hanks, 1989) and Singular Value Decomposition (SVD; Eckart and Young, 1936)
  - Prediction-based model: Skip-Gram and Negative Sampling (SGNS; Mikolov et al., 2013)
- Sentence-level embedding model
  - *Contextualized* word embedding model: Bidirectional Encoder Representations from Transformer (BERT; Devlin et al., 2018)

Corpus

What is Sejong corpus?

- Sejong corpus was created by the 21st Century Sejong Project, a ten-year-long project that was launched in 1998.
- Sejong corpus is a representative large-scale corpus in Korean (Kim, 2006; Shin, 2008).
- Previous studies often used this corpus as a linguistic resource (e.g., Kim & Ock, 2016; Park & Cha, 2017; Shin et al., 2005).

What is Sejong corpus?

Table 1: *Primary corpus*

Corpus type	Corpus size(eojul)
Raw corpus	63,899,412
Grammatically tagged corpus	15,226,186
Parsed corpus	570,064
Semantically Tagged corpus	10,132,348
Sum	**89,830,015**

Table 2: *Plan for construction of raw corpus*

Field	Portion
Newspaper	20%
Magazine	10%
Academic works	35%
Literary works	20%
Quasi-spoken data	10%
The others	5%
Sum	100%

The eojul is defined as a morpheme or combination of several morphemes serving as the minimal unit of sentential components in Korean.

What is the Sejong corpus?

Table 1: *Primary corpus*

Corpus type	Corpus size(eojul)
Raw corpus	63,899,412
Grammatically tagged corpus	15,226,186
Parsed corpus	570,064
Semantically Tagged corpus	10,132,348
Sum	**89,830,015**

Table 2: *Plan for construction of raw corpus*

Field	Portion
Newspaper	20%
Magazine	10%
Academic works	35%
Literary works	20%
Quasi-spoken data	10%
The others	5%
Sum	100%

The eojul is defined as a morpheme or combination of several morphemes serving as the minimal unit of sentential components in Korean.

Example of the semantically tagged corpus

BSAA0001-00001596	생산자의	생산자/NNG + 의/JKG
BSAA0001-00001597	얼굴	얼굴/NNG
BSAA0001-00001598	사진이	사진__07/NNG + 이/JKS
BSAA0001-00001599	붙어	붙/VV + 어/EC
BSAA0001-00001600	있는	있/VX + 는/ETM
BSAA0001-00001601	농산물이	농산물/NNG + 이/JKS
BSAA0001-00001602	나오고	나오/VV + 고/EC
BSAA0001-00001603	있다.	있/VX + 다/EF + ./SF

Example of the semantically tagged corpus

BSAA0001-00001596	생산자의	생산자/NNG + 의/JKG
BSAA0001-00001597	얼굴	얼굴/NNG
BSAA0001-00001598	사진이	사진__07/NNG + 이/JKS
BSAA0001-00001599	붙어	붙/VV + 어/EC
BSAA0001-00001600	있는	있/VX + 는/ETM
BSAA0001-00001601	농산물이	농산물/NNG + 이/JKS
BSAA0001-00001602	나오고	나오/VV + 고/EC
BSAA0001-00001603	있다.	있/VX + 다/EF + ./SF

Description for annotation

- Annotators: three native speakers of Korean.
- Data: 15,000 sentences (*-ey*: 5,000; *-eyse*: 5,000; *-(u)lo*: 5,000)
- Functions: select the most frequent functions based on the Sejong Electronic Dictionary and the previous studies on adverbial postpositions.
  - *-ey*: Location, Goal, Effector, Criterion, Theme, Instrument, Agent, Final state
  - *-eyse*: Source, Location
  - *-(u)lo*: Final state, Instrument, Direction, Effector, Criterion, Location
- Fleiss's Kappa: *-ey*: 0.948; *-eyse*: 0.928; *-(u)lo*: 0.947

A hand-coded corpus

-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

A hand-coded corpus

```
Index ### Label ### Function ### Sentence_POS ### Sentence
1 ### 0 ### FNS ### 이__05/MM 넥타이/NNG 는/JX 수제품/NNG (으)로/JKB 우리나라/NNG 에서
2 ### 2 ### DIR ### 나/NP 의/JKG 마음__01/NNG 의/JKG 움직임/NNG 이/JKS 위__01/NNG
3 ### 1 ### INS ### 굿/NNG 무당__01/NNG 이/JKS 노래/NNG 나/JC 춤__01/NNG (으)로/JK
4 ### 0 ### FNS ### 모든/MM 주장__03/NNG 이/JKS 나름/NNB 대로/JKB 의/JKG 근거/NNG 를
5 ### 3 ### EFF ### 기억/NNG 이/JKS 스스로/NNG 의/JKG 부력__01/NNG (으)로/JKB 떠오르,
6 ### 2 ### DIR ### 신축__03/NNG 전원주택/NNG 위쪽/NNG (으)로/JKB 는/JX 집__01/NNG
7 ### 0 ### FNS ### 멍멍/XR 하/XSA ㄴ/ETM 채__09/NNB (으)로/JKB 시간__04/NNG 이/JK
8 ### 1 ### INS ### 수한/NNP 이/JKS 저/NP 의/JKG 손__01/NNG (으)로/JKB 저/NP 의/JK
9 ### 2 ### DIR ### 쇠전__01/NNG 꾼/XSN 들/XSN 이/JKS 술청/NNG (으)로/JKB 돌아오/VV
10 ### 3 ### EFF ### 그리고/MAJ 그/MM 결과__02/NNG (으)로/JKB 오줌/NNG 이/JKS 나오/V
11 ### 5 ### LOC ### "/SS 집__01/NNG 들/XSN 이/JKS 다/MAG 어디/NP (으)로/JKB 가/VV
12 ### 5 ### LOC ### 바로/MAG 앞/NNG (으)로/JKB 소달구지/NNG 바퀴__01/NNG 자국__01/N
```

Available at: https://github.com/seongmin-mun/Corpora/tree/main/APIK

Part 1: word-level embedding

Creating training and test sets

Training set

집__01/NNG 에서/JKB_LOC 동화책/NNG 을/JKO 보/VV 았/EP 다/EF ./SF
집__01/NNG 에서/JKB_SRC 다시/MAG 오/VV 았/EP 다/EF ./SF

Test set
그/MM 와/JKB 집__01/NNG 에서/JKB 만나/VV 았/EP 다/EF ./SF

Figure: Example sentences used in the model training and testing (*-eyse*)

Creating training and test sets

Training set

집__01/NNG 에서/JKB_LOC 동화책/NNG 을/JKO 보/VV 았/EP 다/EF ./SF
집__01/NNG 에서/JKB_SRC 다시/MAG 오/VV 았/EP 다/EF ./SF

Test set
그/MM 와/JKB 집__01/NNG 에서/JKB 만나/VV 았/EP 다/EF ./SF

Figure: Example sentences used in the model training and testing (*-eyse*)

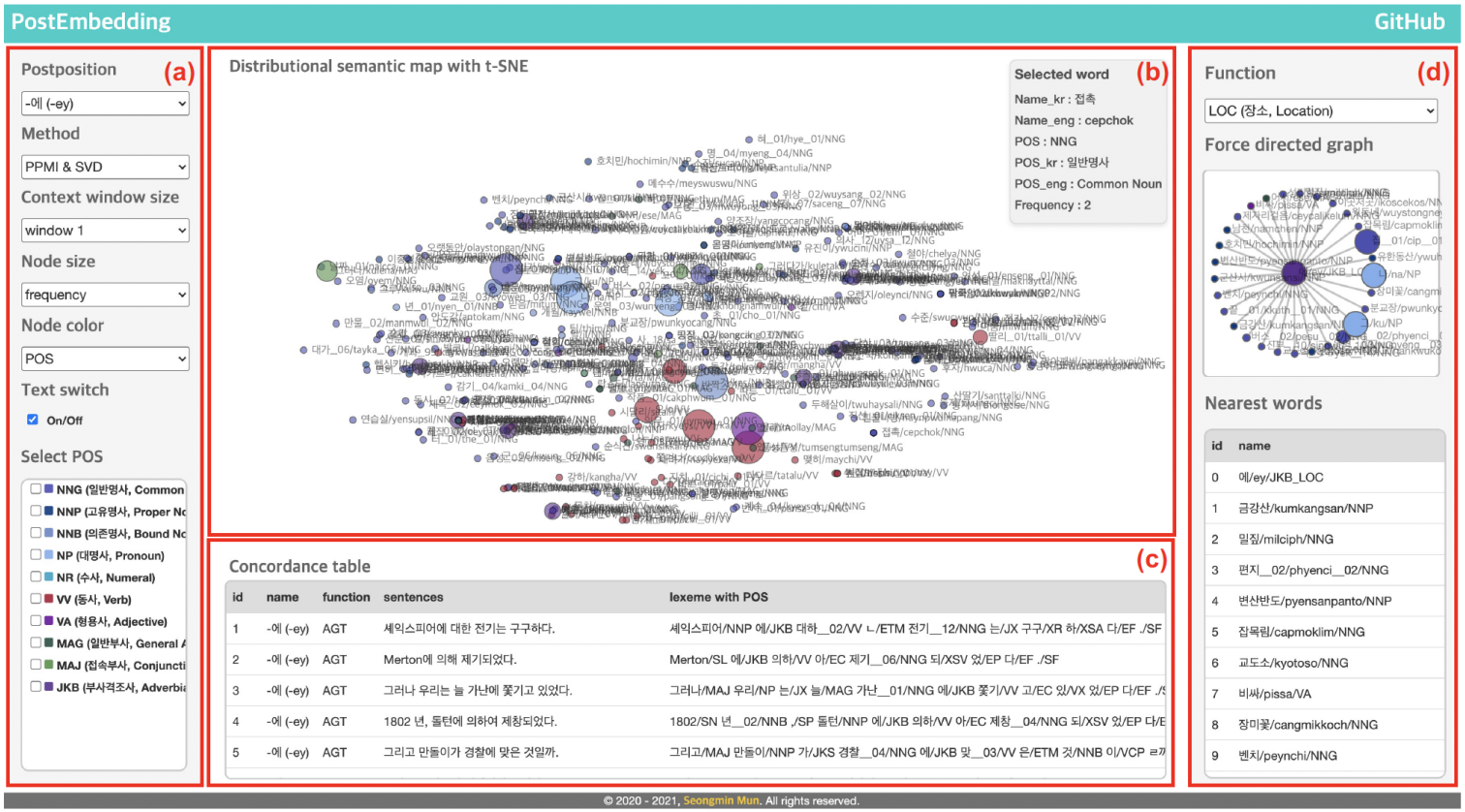
Model specification: PPMI-SVD and SGNS

- General
  - Normalization: 10-fold cross-validation
  - Context window size: from one to ten
  - Embedding size: 500
  - Dimension reduction: *t*-SNE (Maaten and Hinton, 2008)
  - Classification model: Similarity-based estimation (Dagan et al., 1995)
  - Measurement method: accuracy (e.g., Levy et al., 2015; Riedl Biemann, 2017; Warstadt et al., 2019)
- Architecture specific
  - Count-based model: PPMI-SVD
    - Package used: *Linalg* from the *scipy* package
  - Prediction-based model: SGNS
    - Package used: *Word2Vec* from the *gensim* package

Model performance: Classification

- The fewer functions the postposition had, the higher the classification accuracy was obtained.
  - PPMI-SVD: *-ey*: 0.534, *-eyse*: 0.773, *-(u)lo*: 0.567
  - SGNS: *-ey*: 0.204, *-eyse*: 0.693, *-(u)lo*: 0.368
- The PPMI-SVD obtained high classification accuracy at a larger window size.
- They performed well only when the target functions occurred very frequently in the corpus.

Visualization: PostEmbedding



Available at: https://seongmin-mun.github.io/VisualSystem/Major/PostEmbedding/index.html

Visualization: clusters and co-occurring words

Index	Function	Sentence
517	CRT	근디/MAJ 내/NP 가/JKS 아침/NNG 에/JKB 미역국/NNG 도/JX 못/MAG 묵__03/VV 었/EP 다/EF ./SF
789	GOL	그/MM 강물/NNG 속__01/NNG 에/JKB 당신/NP 과/JC 내/NP 가/JKS 떠가/VV ㅂ니다/EF ./SF
1306	CRT	잠깐/MAG 내/NP 가/JKS 안/MAG 보__01/VV 는/ETM 사이__01/NNG 에/JKB ./SF
1487	LOC	쇼윈도/NNG 저편/NP 에/JKB 내/NP 가/JKS 있/VV 었/EP 다/EF ./SF
1496	EFF	내/NP 가/JKS 저/MM 녀석/NNB 때문/NNB 에/JKB 푹/MAG 늙/VV 었/EP 어/EF ./SF
1763	CRT	에어/NNG 는/JX 어젯밤/NNG 에/JKB 내/NP 가/JKS 빼__01/VV 었/EP 어요/EF ./SF

.....

Visualization: clusters and co-occurring words

Index	Function	Sentence
1483	EFF	바람__01/NNG 에/JKB 흔들리/VV 는/ETM 나뭇잎/NNG ./SF
1505	EFF	바람__01/NNG 에/JKB 흔들리/VV 는/ETM 갈대/NNG 이/VCP 냐/EF ./SF
1514	EFF	바람__01/NNG 때문/NNB 에/JKB 흔들리/VV 는/ETM 것/NNB 은/JX 아니/VCN 었/EP 어/EF ./SF
1597	EFF	바람__01/NNG 에/JKB 흔들리/VV 는/ETM 갈대/NNG 이/VCP 냐/EF ./SF
1946	EFF	잔잔/XR 하/XSA ㄴ/ETM 바람결/NNG 에/JKB 난실난실/MAG 흔들리/VV 었/EP 습니다/EF ./SF
3612	EFF	바람__01/NNG 에/JKB 흔들리/VV 는/ETM 갈대/NNG 의/JKG 순정__03/NNG .../SE .../SE

Discussion: word-level embedding models

- The two models have shown unsatisfactory classification performances.
- They performed well only when the target functions occurred very frequently in the corpus.
- The cluster was not changed much by the environments of word-level embedding (2 models $*$ 3 postpositions $*$ 10 window sizes).

Part 2: sentence-level embedding

Creating training and test sets

Index	Label	Sentence
1,862	1	[CLS] 한참 만에 오반장이 침묵을 깼다. [SEP]
1,863	1	[CLS] 정말 오랫만에 먹어보는 고기였다. [SEP]
1,864	1	[CLS] 옛날 구한말에 유명한 얘기가 있었죠? [SEP]
1,865	1	[CLS] 한밤중에 신나게 한바탕했지요. [SEP]
1,866	1	[CLS] 그런데 몇 시에 왔어? [SEP]
1,867	1	[CLS] 겨울에 꽃이라니요. [SEP]
1,868	1	[CLS] 아침에 엄마한테 돈을 달랬어요. [SEP]
1,869	1	[CLS] 결혼은 반드시 적령기에 해야 한다. [SEP]
1,870	1	[CLS] 한 달에 얼마씩은 정확하게 들어오니까. [SEP]
1,871	1	[CLS] 그럼 일 주일 후에 뵙겠습니다. [SEP]

Figure: Example sentences used in the BERT training (*-ey*, CRT)

Creating training and test sets

Index	Label	Sentence
1,862	1	[CLS] 한참 만에 오반장이 침묵을 깼다. [SEP]
1,863	1	[CLS] 정말 오랫만에 먹어보는 고기였다. [SEP]
1,864	1	[CLS] 옛날 구한말에 유명한 얘기가 있었죠? [SEP]
1,865	1	[CLS] 한밤중에 신나게 한바탕했지요. [SEP]
1,866	1	[CLS] 그런데 몇 시에 왔어? [SEP]
1,867	1	[CLS] 겨울에 꽃이라니요. [SEP]
1,868	1	[CLS] 아침에 엄마한테 돈을 달랬어요. [SEP]
1,869	1	[CLS] 결혼은 반드시 적령기에 해야 한다. [SEP]
1,870	1	[CLS] 한 달에 얼마씩은 정확하게 들어오니까. [SEP]
1,871	1	[CLS] 그럼 일 주일 후에 뵙겠습니다. [SEP]

Figure: Example sentences used in the BERT training (*-ey*, CRT)

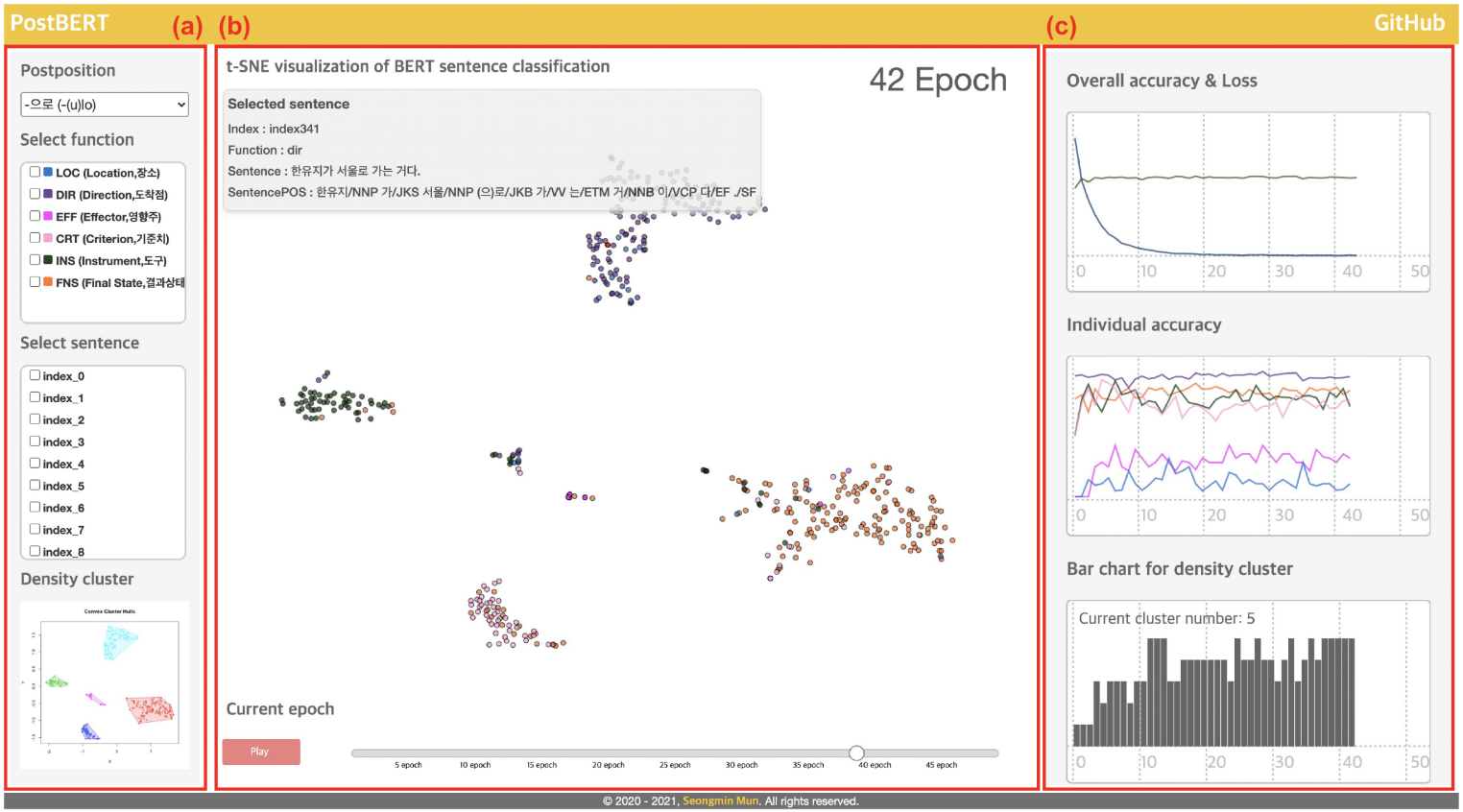
Model specification: BERT

- *Contextualized* word embedding model: BERT (Devlin et al., 2018)
  - Package used: *Transformer*
  - Pre-trained model: KoBERT (Jeon et al., 2019)
  - Tokenizer: KoBERT tokenizer (Jeon et al., 2019)
  - Epoch: from one to 50
  - Other parameters: Learning rate (.00001); Batch (32); Sequence length (256); Seed (42); Epsilon (.00000001)
  - Dimension reduction: *t*-SNE (Maaten and Hinton, 2008)

Model performance: Classification

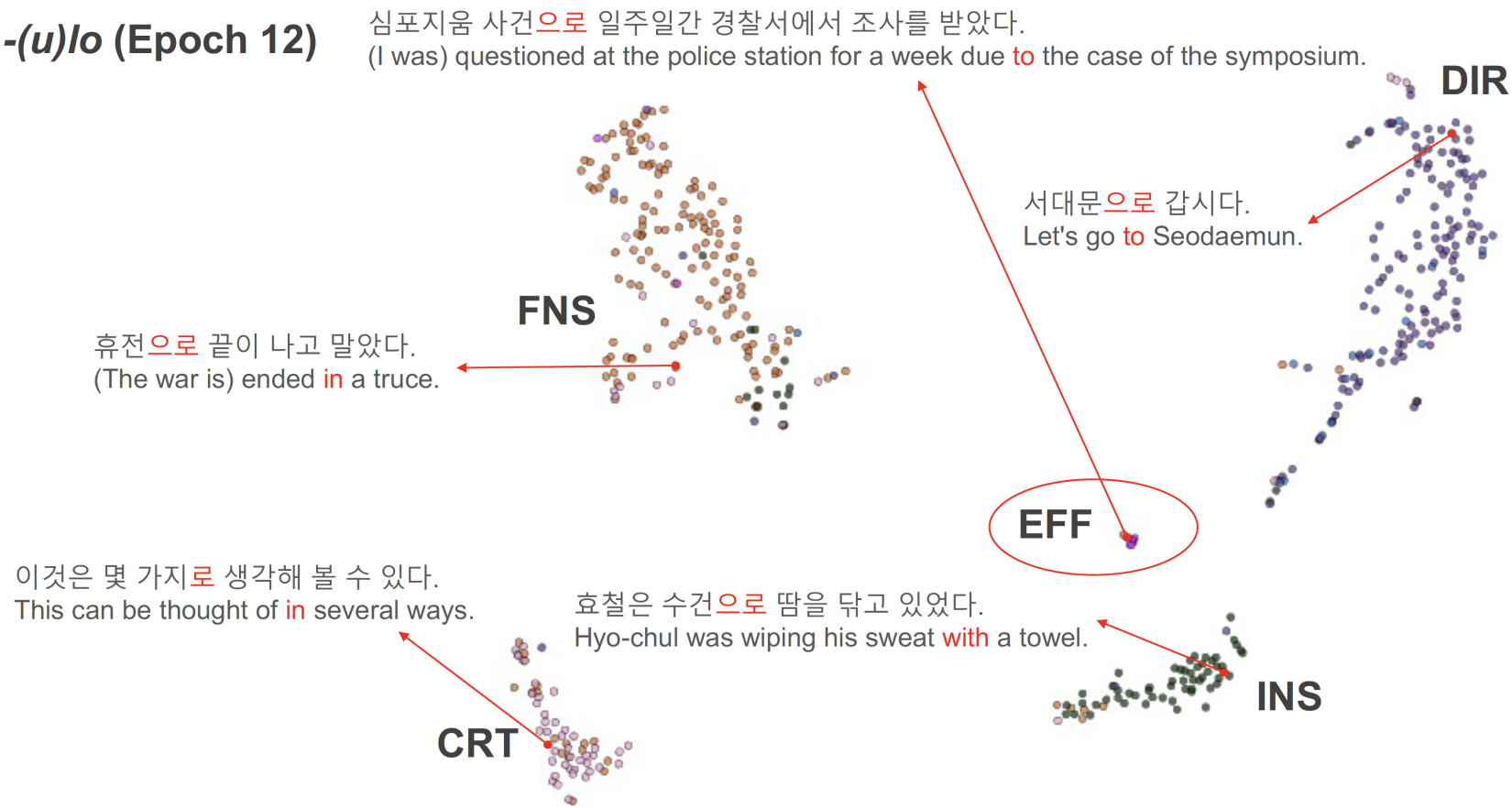
- The higher classification accuracy was obtained when the postposition has a fewer number of functions.
  - BERT: *-ey*: 0.815, *-eyse*: 0.898, *-(u)lo*: 0.813
- The model performance increased as the epoch progressed.

Visualization: PostBERT



Available at:
https://seongmin-mun.github.io/VisualSystem/Major/PostBERT/index.html

Visualization: clusters of sentence-level embeddings



Visualization: clusters of sentence-level embeddings

Visualization: clusters of sentence-level embeddings

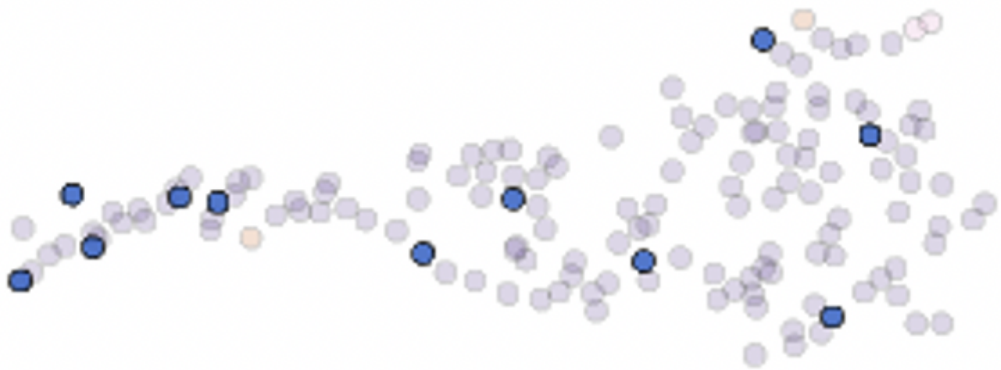


Figure: The DIR cluster in the distributional map for *-(u)lo* (Epoch 46) highlighting the LOC instances.

Discussion: sentence-level embedding model

- The model can identify the intended functions of a postposition as the epoch progresses, even though the corpus size of a function is small.
- The BERT model performs in a stable way and simulates how humans recognize the polysemy involving Korean adverbial postpositions better than the word-level embedding models do.

Discussion: sentence-level embedding model

"These results suggest that it is likely that BERT does acquire **some form of a structural inductive bias** from self-supervised pretraining, at least outside of the NPI domain."
(Warstadt Bowman, 2020)

Discussion: sentence-level embedding model

"One possibility is that the transformer's self-attention mechanism and layer-wise organization improves its **ability to represent lexically specific structures**."
(Hawkins et al., 2020)

Discussion: sentence-level embedding model

"Our results allow us to conclude that BERT does indeed have access to **a significant amount of information**, much of which linguists typically call constructional information."
(Madabushi et al., 2020)

Discussion: sentence-level embedding model

- BERT performs better than word-level embedding models because BERT uses the amount of information not only morphological information but also structural information.
- Perhaps, BERT is a better approach for understanding how humans deal with polysemy.

Conclusion

Summary of major findings

- Word-level embedding models
  - If we only used morphological information, we cannot resolve the word-level polysemy.
  - When the function of the postposition changed, the surrounding words also changed.
  - There is a particular word that has a strong connection with the specific function of the postposition.

Summary of major findings

- Sentence-level embedding model
  - To understand word-level polysemy, at least, we have to use the amount of information not only morphological information but also structural information.
  - If we spend more time learning a language, we can identify the word-level polysemy more clearly.
  - Even if the function of the postposition is used rarely but it can be distinguished from the other functions, we can identify it as a distinguished function.
  - If the functions are semantically similar to each other, it is hard to be distinguished one from the other.

Summary of implications

- It provides the possible ways and limitations of applying three different embedding models to the Korean language (i.e., methodological generalizability).
- It proposes two interactive visualization systems that help to explore the relationships between words or sentences (i.e., model evaluation).

Thank you for listening.