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

Seongmin Mun & Guillaume Desagulier

Chosun University & Paris VIII University

27th May 2022

Outline

Introduction
- Polysemy in Korean

Corpus
- Sejong corpus
- Creation of a hand-coded corpus

Classification models: BERT & GPT-2

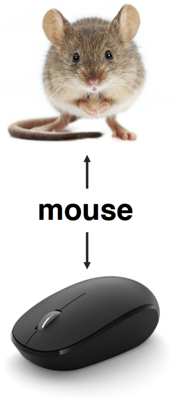
Visualization: PostTransformers

Discussion & Conclusion

Introduction

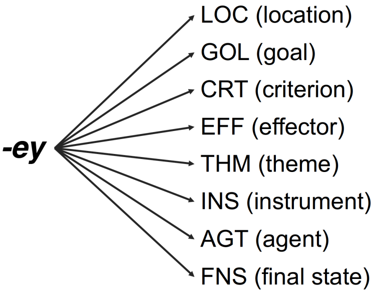
Polysemy

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

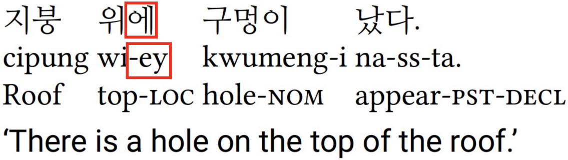


Korean language

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



Polysemy in Korean adverbial postposition



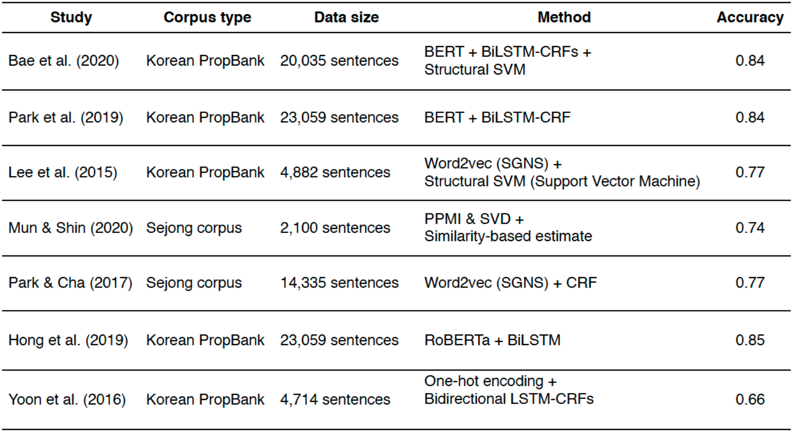
지붕	위에	구멍이	났다.
cipung	wi-ey	kwumeng-i	na-ss-ta.
Roof	top-LOC	hole-NOM	appear-PST-DECL

‘There is a hole on the top of the roof.’

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

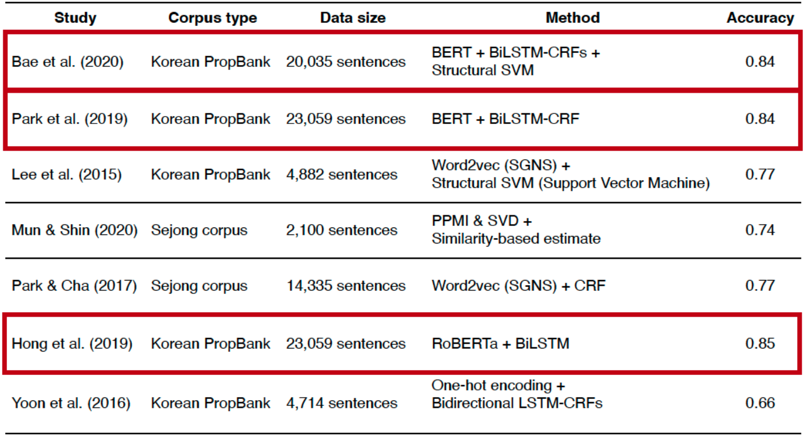
Question: How a speaker can understand the function of postposition?

Previous studies on adverbial postpositions



Study	Corpus type	Data size	Method	Accuracy
Bae et al. (2020)	Korean PropBank	20,035 sentences	BERT + BiLSTM-CRFs + Structural SVM	0.84
Park et al. (2019)	Korean PropBank	23,059 sentences	BERT + BiLSTM-CRF	0.84
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
Hong et al. (2019)	Korean PropBank	23,059 sentences	RoBERTa + BiLSTM	0.85
Yoon et al. (2016)	Korean PropBank	4,714 sentences	One-hot encoding + Bidirectional LSTM-CRFs	0.66

Previous studies on adverbial postpositions



Study	Corpus type	Data size	Method	Accuracy
Bae et al. (2020)	Korean PropBank	20,035 sentences	BERT + BiLSTM-CRFs + Structural SVM	0.84
Park et al. (2019)	Korean PropBank	23,059 sentences	BERT + BiLSTM-CRF	0.84
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
Hong et al. (2019)	Korean PropBank	23,059 sentences	RoBERTa + BiLSTM	0.85
Yoon et al. (2016)	Korean PropBank	4,714 sentences	One-hot encoding + Bidirectional LSTM-CRFs	0.66

Transformer-architecture models that we used

- *Contextualized* word embedding model
  - Bidirectional Encoder Representations from Transformer (BERT; Devlin et al., 2018)
  - Generative Pre-Training 2 (GPT-2; Radford et al., 2019)

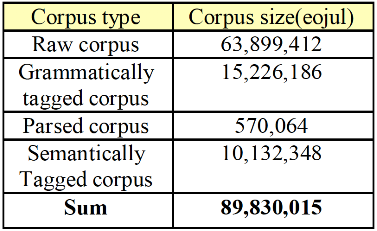
Corpus

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 (Shin, 2008).
- Previous studies often used this corpus as a linguistic resource (e.g., Kim & Ock, 2016; Park & Cha, 2017; Shin et al., 2005).

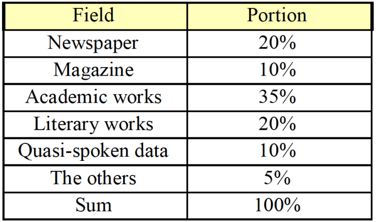
What is Sejong corpus?

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*

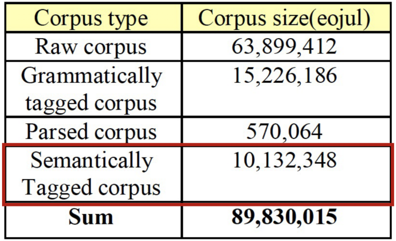


Field	Portion
Newspaper	20%
Magazine	10%
Academic works	35%
Literary works	20%
Quasi-spoken data	10%
The others	5%
Sum	100%

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

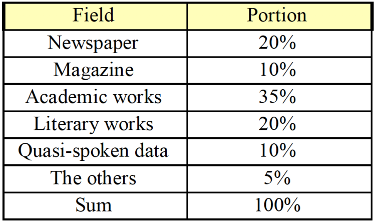
What is the Sejong corpus?

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**

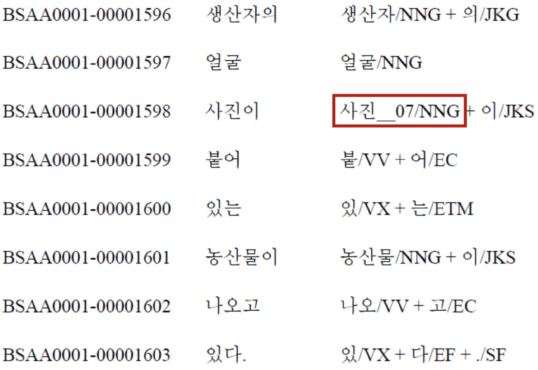
Table 2: *Plan for construction of raw corpus*



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

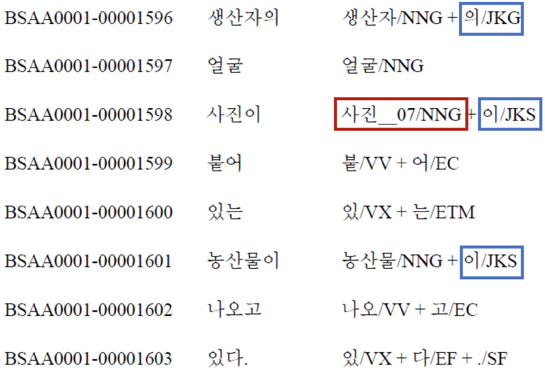
The eojul is defined as a morpheme or combination of several morphemes serving as the minimal unit of sentential components in Korean.

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

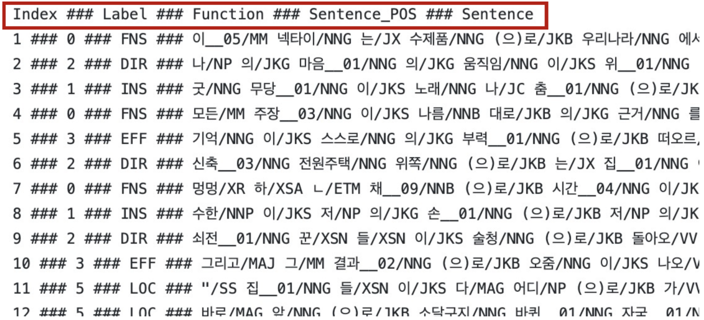
Description for annotation

- 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

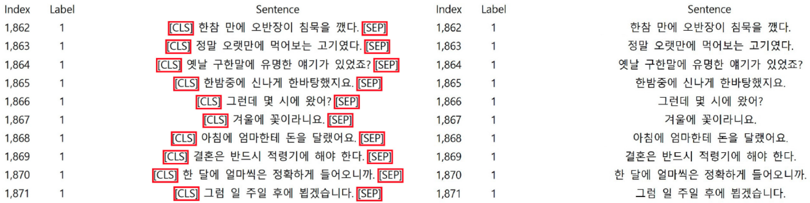


```
Index ### Label ### Function ### Sentence_POS ### Sentence
1 ### 0 ### FNS ### 이__05/MM 넥타이/NNG 는/JX 수제품/NNG (으)로/JKB 우리나라/NNG 에
2 ### 2 ### DIR ### 나/NP 의/JKG 마음__01/NNG 의/JKG 움직임/NNG 이/JKS 위__01/NNG
3 ### 1 ### INS ### 굿/NNG 무당__01/NNG 이/JKS 노래/NNG 나/JC 춤__01/NNG (으)로/JK
4 ### 0 ### FNS ### 모든/MM 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/V
12 ### 5 ### LOC ### 바로/MAG 앞/NNG (으)로/JKB 소닥구지/NNG 바퀴__01/NNG 자국__01/N
```

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

Classification models: BERT & GPT-2

Creating training and test sets

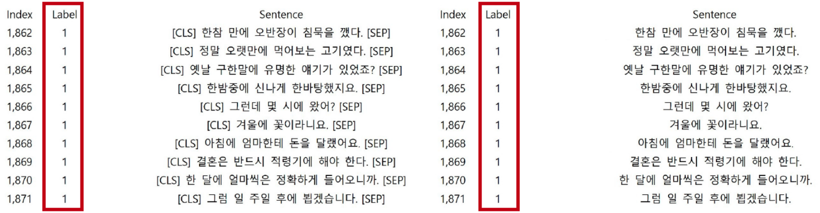


Index	Label	Sentence
1,862	1	[CLS] 한참 만에 오반장이 침묵을 깼다. [SEP]
1,863	1	[CLS] 정말 오랫만에 먹어보는 고기였다. [SEP]
1,864	1	[CLS] 옛날 구한말에 유명한 얘기가 있었죠? [SEP]
1,865	1	[CLS] 한밤중에 신나게 한바탕했지요. [SEP]
1,866	1	[CLS] 그런데 몇 시에 왔어? [SEP]
1,867	1	[CLS] 겨울에 꽃이라니요. [SEP]
1,868	1	[CLS] 아침에 엄마한테 돈을 달랬어요. [SEP]
1,869	1	[CLS] 결혼은 반드시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

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

Figure: Example sentences used in the training for BERT (left) and GPT-2 (Right)

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]

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

Figure: Example sentences used in the training for BERT (left) and GPT-2 (Right)

Model specification: BERT

- Bidirectional Encoder Representations from Transformer (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 (.00002); Batch (16); Sequence length (128); Seed (42); Epsilon (.00000001)
  - Dimension reduction: *t*-SNE (Maaten and Hinton, 2008)

Model specification: GPT-2

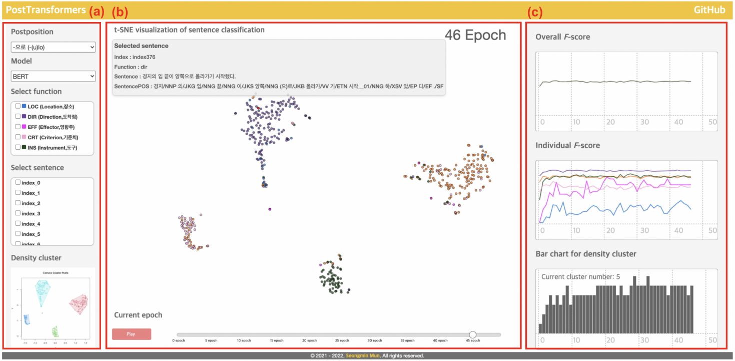
- Generative Pre-Training 2 (GPT-2; Radford et al., 2019)
  - Package used: *Transformer*
  - Pre-trained model: KoGPT2 (Jeon et al., 2021)
  - Tokenizer: GPT2 tokenizer (Jeon et al., 2019)
  - Epoch: from one to 50
  - Other parameters: Learning rate (.00002); Batch (16); Sequence length (128); Seed (42); Epsilon (.00000001)
  - Dimension reduction: *t*-SNE (Maaten and Hinton, 2008)

Model performance: Classification

- BERT performed better than GPT-2 in revealing the polysemy of Korean postpositions.
  - BERT: *-ey*: 0.744, *-eyse*: 0.875, *-(u)lo*: 0.795
  - GPT-2: *-ey*: 0.68, *-eyse*: 0.844, *-(u)lo*: 0.676
- The model performance increased as the epoch progressed.

Visualization: PostTransformers

Visualization: PostTransformers



Postposition: -으로 (-(u)lo)

Model: BERT

Select function

- □ LOC (Location,장소)
- □ DIR (Direction,도착점)
- □ EFF (Effector,영향주)
- □ CRT (Criterion,기준치)
- □ INS (Instrument,도구)

Select sentence

- □ index_0
- □ index_1
- □ index_2
- □ index_3
- □ index_4
- □ index_5
- □ index_6

Selected sentence

Index : index376

Function : dir

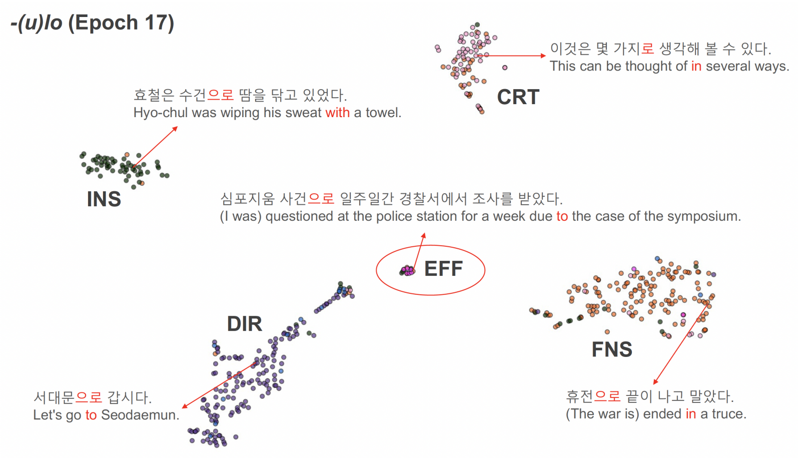
Sentence : 경지의 입 끝이 양쪽으로 올라가기 시작했다.

SentencePOS : 경지/NNP 의/JKG 입/NNG 끝/NNG 이/JKS 양쪽/NNG (으)로/JKB 올라가/VV 기/ETN 시작__01/NNG 하/XSV 았/EP 다/EF ./SF

© 2021 - 2022, Seongmin Mun. All rights reserved.

Available at: https://seongmin-mun.github.io/Visualization/2022/PostTransformers/index.html

Visualization: clusters of BERT



-*(u)lo* (Epoch 17)

이것은 몇 가지로 생각해 볼 수 있다.
This can be thought of in several ways.

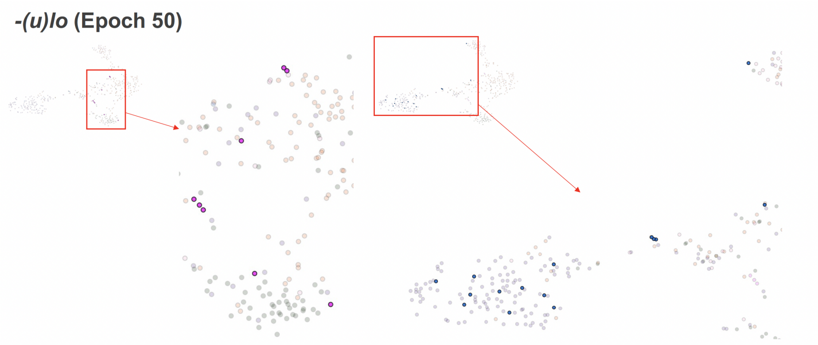
효철은 수건으로 땀을 닦고 있었다.
Hyo-chul was wiping his sweat with a towel.

심포지움 사건으로 일주일간 경찰서에서 조사를 받았다.
(I was) questioned at the police station for a week due to the case of the symposium.

서대문으로 갑시다.
Let's go to Seodaemun.

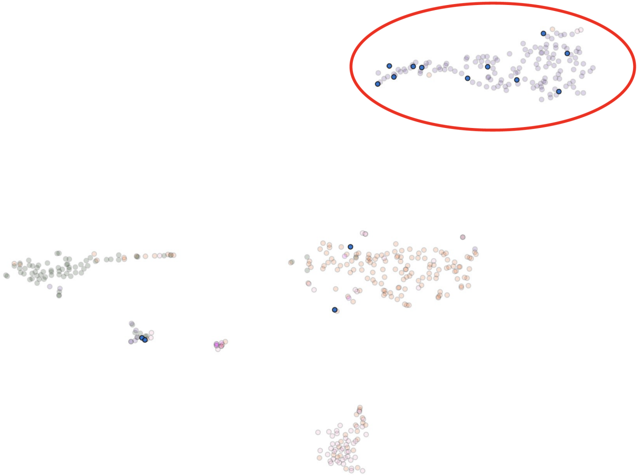
휴전으로 끝이 나고 말았다.
(The war is) ended in a truce.

Visualization: clusters of GPT-2



***-(u)lo* (Epoch 50)**

Visualization: clusters of BERT



Discussion & Conclusion

Discussion

- The BERT model performs in a stable way and simulates how humans recognize the polysemy involving Korean adverbial postpositions better than GPT-2 model does.

Discussion

"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

"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

"**GPT-2's perplexity** is better captured by the considered features and it resulted to be more affected by **lexical parts-of-speech** and features capturing the **vocabulary richness of a sentence**. On the contrary, **BERT's perplexity** seems to be best predicted by **syntactic features** highly sensitive to sentence length."
(Miaschi et al. 2021)

Discussion

- BERT performs better than GPT-2 because the meaning of Korean adverbial postposition is maybe sensitive to syntactic features.
- Perhaps, BERT is a better approach for understanding how humans deal with polysemy.

Conclusion

- To understand word-level polysemy of Korean postposition, at least, we have to use the syntactic 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.

Thank you for listening.