Polysemy interpretation by using similarity based estimation

Seongmin Mun

Chosun University

18th June 2022

The Linguistic Society of Korea
한국언어학회

Outline

Personal Profile

Introduction
- Polysemy in Korean
- Distributional semantic models (DSMs)

Methods

Result & Discussion

Appendix

Personal Profile

Seongmin Mun

Experience & Education

PostDoc – Chosun University

- NLP
- Web / server development
- Deep learning
- Image processing

Ph.D. – Université Paris Nanterre

- NLP
- Data visualization
- Neural network
- Linguistics
- Statistics
- Web-based system
- Language models
- Machine learning

M.S. – Ajou University

- Data visualization
- Machine learning
- Web-based system
- Statistics

https://seongminmun.com/

Seongmin Mun

Skills & Endorsements

Research Knowledge

- NLP
- Linguistics
- Data Visualization
- Data analysis
- Machine Learning
- Web Development

Computer Language

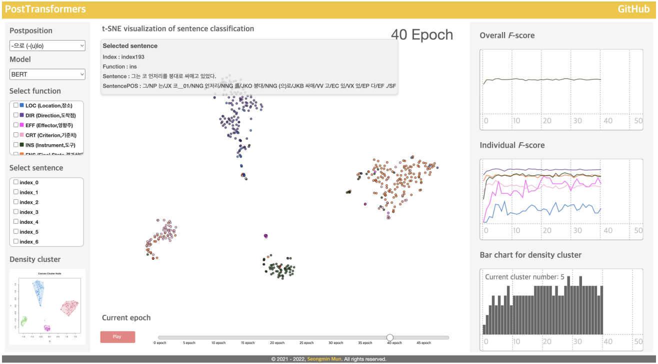
- Java
- JavaScript
- HTML/CSS
- Python
- SQL
- PHP
- R

Statistics Software

- R
- MATLAB
- SAS
- SPSS

https://seongminmun.com/

Mun, 2021



Selected sentence

Index : index193

Function : ins

Sentence : 그는 코 언저리를 붕대로 싸매고 있었다.

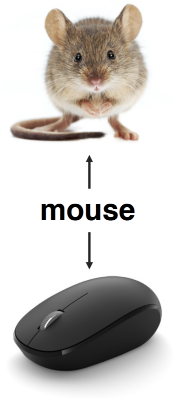
SentencePOS : 그/NP 는/JX 코__01/NNG 언저리/NNG 를/JKO 붕대/NNG (으)로/JKB 싸매/VV 고/EC 있/VX 었/EP 다/EF ./SF

Mun, S. (2021). Polysemy resolution with word embedding models and data visualization: the case of adverbial postpositions -ey, -eyse, and -(u)lo in Korean. presented at IMPRS2020 (MaxPlanck), ICCG11, and ACL 2022

Introduction

Polysemy

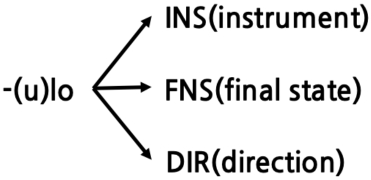
Polysemy, one type of ambiguity, occurs when one form delivers multiple meanings/functions (Glynn and Robinson, 2014).



mouse

Korean language

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



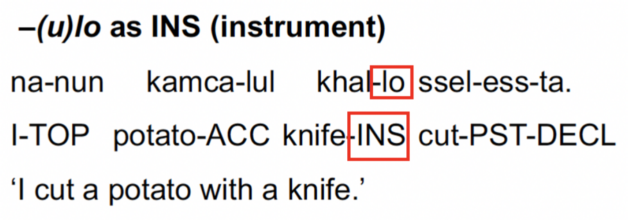
-(u)lo

INS(instrument)

FNS(final state)

DIR(direction)

Polysemy in Korean adverbial postposition



–*(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.’

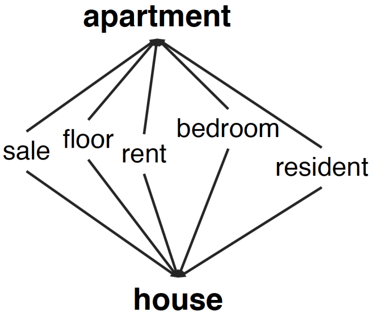
Figure: An example sentence involving the postposition -*(u)lo* as a function of INS (instrument)

Polysemy in Korean

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

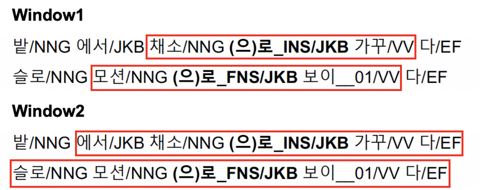
Concept of DSMs

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).



Context window

A range of words surrounding a target word, affecting the determination of its characteristics (Mun, 2021)



Window1

밭/NNG 에서/JKB 채소/NNG **(으)로_INS/JKB** 가꾸/VV 다/EF

슬로/NNG 모션/NNG **(으)로_FNS/JKB** 보이__01/VV 다/EF

Window2

밭/NNG 에서/JKB 채소/NNG **(으)로_INS/JKB** 가꾸/VV 다/EF

슬로/NNG 모션/NNG **(으)로_FNS/JKB** 보이__01/VV 다/EF

Word-level embedding model

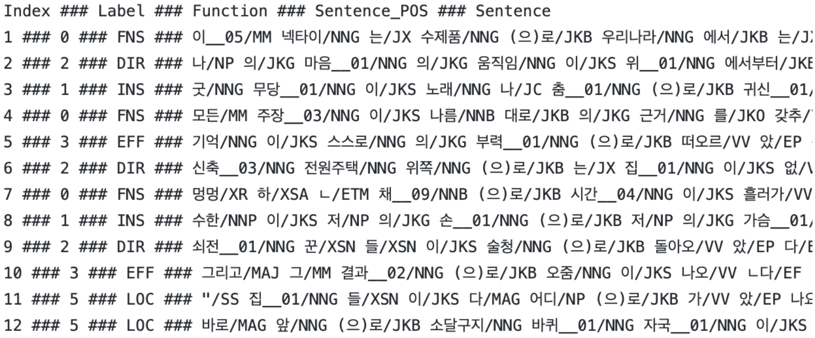
- Model training: Positive Pointwise Mutual Information (PPMI; Church and Hanks, 1989) and Singular Value Decomposition (SVD; Eckart and Young, 1936).
- Classification model: similarity-based estimate (Dagan et al., 1993) by calculating cosine similarity scores between *-(u)lo* and its co-occurring content words.

Methods

Corpus: Adverbial Postpositions In Korean (APIK)

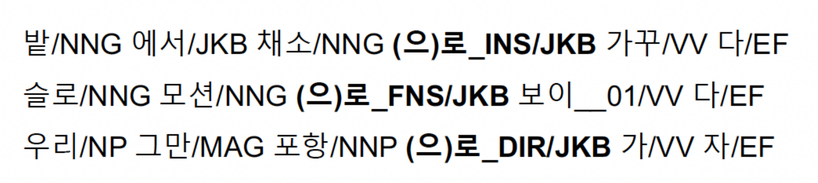
- Sejong corpus, with semantic annotations of three adverbial postpositions *-ey*, *-eyse*, and *-(u)lo* cross-verified by three native speakers of Korean (Mun & Desagulier, 2022)
- Available at: https://github.com/seongmin-mun/Corpora/tree/main/APIK

Corpus: Adverbial Postpositions In Korean (APIK)



```
Index ### Label ### Function ### Sentence_POS ### Sentence
1 ### 0 ### FNS ### 이__05/MM 넥타이/NNG 는/JX 수제품/NNG (으)로/JKB 우리나라/NNG 에서/JKB 는/J
2 ### 2 ### DIR ### 나/NP 의/JKG 마음__01/NNG 의/JKG 움직임/NNG 이/JKS 위__01/NNG 에서부터/JK
3 ### 1 ### INS ### 굿/NNG 무당__01/NNG 이/JKS 노래/NNG 나/JC 춤__01/NNG (으)로/JKB 귀신__01
4 ### 0 ### FNS ### 모든/MM 주장__03/NNG 이/JKS 나름/NNB 대로/JKB 의/JKG 근거/NNG 를/JKO 갖추
5 ### 3 ### EFF ### 기억/NNG 이/JKS 스스로/NNG 의/JKG 부력__01/NNG (으)로/JKB 떠오르/VV 았/EP
6 ### 2 ### DIR ### 신축__03/NNG 전원주택/NNG 위쪽/NNG (으)로/JKB 는/JX 집__01/NNG 이/JKS 없/V
7 ### 0 ### FNS ### 멍멍/XR 하/XSA ㄴ/ETM 채__09/NNB (으)로/JKB 시간__04/NNG 이/JKS 흘러가/VV
8 ### 1 ### INS ### 수한/NNP 이/JKS 저/NP 의/JKG 손__01/NNG (으)로/JKB 저/NP 의/JKG 가슴__01
9 ### 2 ### DIR ### 쇠전__01/NNG 꾼/XSN 들/XSN 이/JKS 술청/NNG (으)로/JKB 돌아오/VV 았/EP 다/E
10 ### 3 ### EFF ### 그리고/MAJ 그/MM 결과__02/NNG (으)로/JKB 오줌/NNG 이/JKS 나오/VV ㄴ다/EF
11 ### 5 ### LOC ### "/SS 집__01/NNG 들/XSN 이/JKS 다/MAG 어디/NP (으)로/JKB 가/VV 았/EP 나오
12 ### 5 ### LOC ### 바로/MAG 앞/NNG (으)로/JKB 소달구지/NNG 바퀴__01/NNG 자국__01/NNG 이/JKS
```

Corpus: Adverbial Postpositions In Korean (APIK)



밭/NNG 에서/JKB 채소/NNG **(으)로_INS/JKB** 가꾸/VV 다/EF

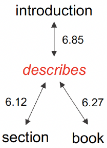
슬로/NNG 모션/NNG **(으)로_FNS/JKB** 보이__01/VV 다/EF

우리/NP 그만/MAG 포항/NNP **(으)로_DIR/JKB** 가/VV 자/EF

ACC = accusative case marker; DAT = dative marker; DECL = declarative; EF = final ending; JKB = adverbial case marker; MAG = general adverb; NNG = common noun; NNP = proper noun; NOM = nominative case marker; NP = pronoun; PST = past tense marker; TOP = topic; VV = verb

Similarity-based estimate (Dagan et al., 1993)

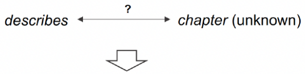
Network from the training set

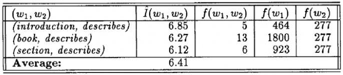


Input as a test item

[introduction,
chapter (unknown),
book,
section]

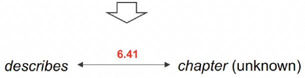
Q: How to calculate the similarity score between '*describes*' and '*chapter*'?



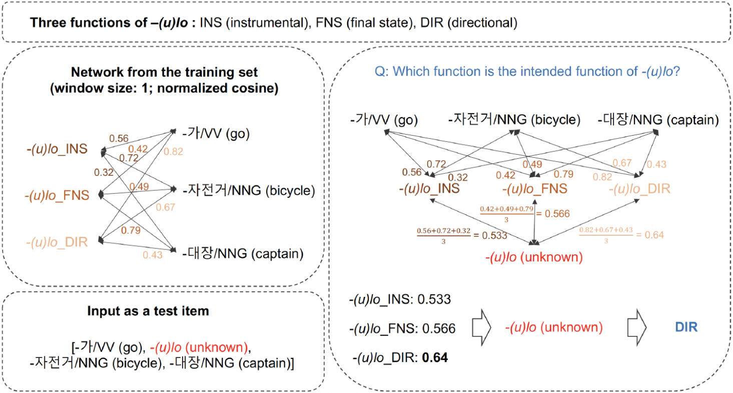


(w_1, w_2)	$\hat{I}(w_1, w_2)$	$f(w_1, w_2)$	$f(w_1)$	$f(w_2)$
(introduction, describes)	6.85	5	464	277
(book, describes)	6.27	13	1800	277
(section, describes)	6.12	6	923	277
Average:	6.41			

Table 1: The similarity based estimate as an average on similar pairs: $\check{I}(chapter, describes) = 6.41$



Approach (adapted from Dagan et al., 1993)



Three functions of *–(u)lo* : INS (instrumental), FNS (final state), DIR (directional)

Network from the training set
(window size: 1; normalized cosine)

Input as a test item

[-가/VV (go), -(u)lo (unknown),
-자전거/NNG (bicycle), -대장/NNG (captain)]

Q: Which function is the intended function of *-(u)lo*?

$\frac{0.42+0.49+0.79}{3} = 0.566$

$\frac{0.56+0.72+0.32}{3} = 0.533$

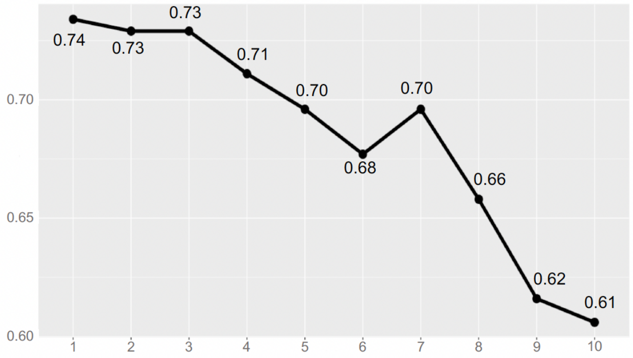
$\frac{0.82+0.67+0.43}{3} = 0.64$

-(u)lo_INS: 0.533
-(u)lo_FNS: 0.566
-(u)lo_DIR: **0.64**

⇨ -(u)lo (unknown) ⇨ **DIR**

Result & Discussion

Result



X-axis: window sizes
Y-axis: accuracy (%)

Our model achieved the highest classification accuracy rate in the window size of one, and the accuracy rates decreased as the window size increased.

Interpretation

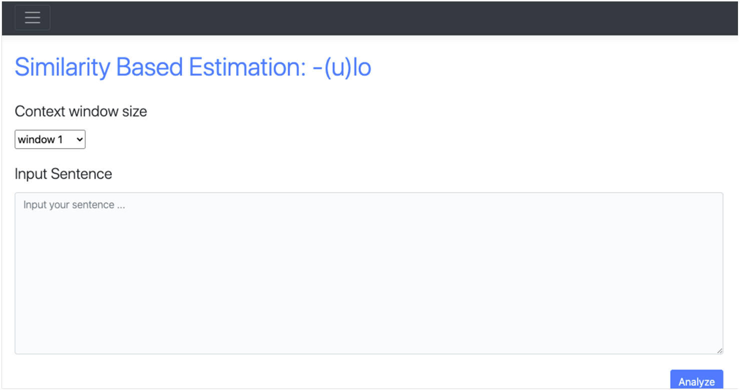
- This trend aligns with advantages of small window sizes (Bullinaria Levy, 2007).
- Considering that a narrower range of context window relates more to syntactic than to semantic information (Patel et al., 1997), our model may have employed structural, more than semantic, characteristics of tri-grams (word-target-word) for the best classification performance.

Appendix

Data processing by using Python

- Colab: Python code

Web-based System



Similarity Based Estimation: -(u)lo

Context window size

window 1

Input Sentence

Input your sentence ...

Thank you for listening.