How does context window size address polysemy of adverbial postposition *-(u)lo* in Korean?

Seongmin Mun (Chosun University)
Gyu-ho Shin (Palacky University Olomouc)

20 August 2021

Outline

Introduction

- Polysemy in Korean
- Distributional semantic models (DSMs)

Corpus

- Sejong corpus
- A hand-coded corpus

Methods

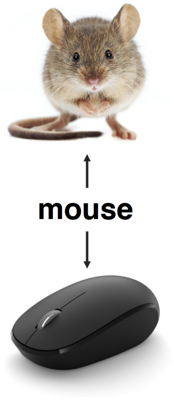
Results

Conclusion

Introduction

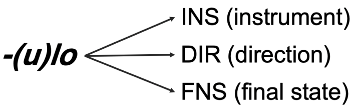
Polysemy

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).



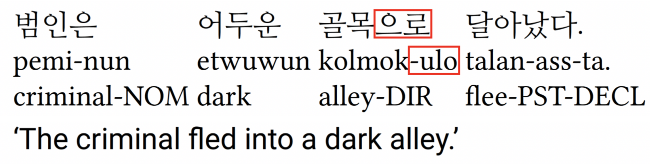
-(u)lo

INS (instrument)

DIR (direction)

FNS (final state)

Polysemy in Korean adverbial postposition



범인은	어두운	골목으로	달아났다.
pemi-nun	etwuwun	kolmok-ulo	talan-ass-ta.
criminal-NOM	dark	alley-DIR	flee-PST-DECL

'The criminal fled into a dark alley.'

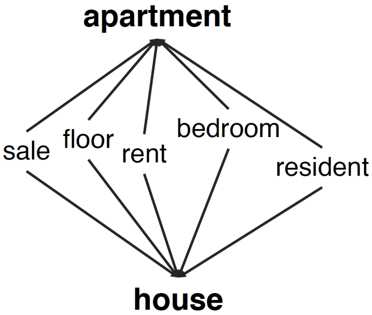
Figure: An example sentence involving the postposition *-(u)lo* as a function of DIR (direction)

Polysemy in Korean

Question: How can a speaker understand the functions of postpositions?

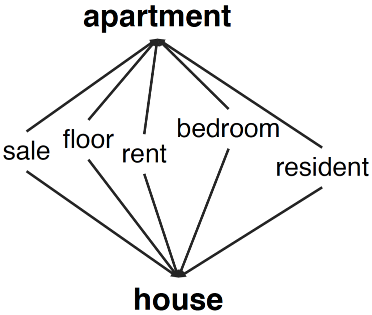
Assumption

Construal of a polysemous word occurs in conjunction with a series of words, delivering various framesemantic meanings (Goldberg, 2006) and yet purporting similar interpretations (Harris, 1954).



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 Korean 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

Distributional semantic models (DSMs)

Context window: a range of words surrounding a target word, affecting the determination of its characteristics (Lison and Kutuzov, 2017).

Distributional semantic models (DSMs)

Question: How does context window address polysemy interpretation in Korean?

Corpus

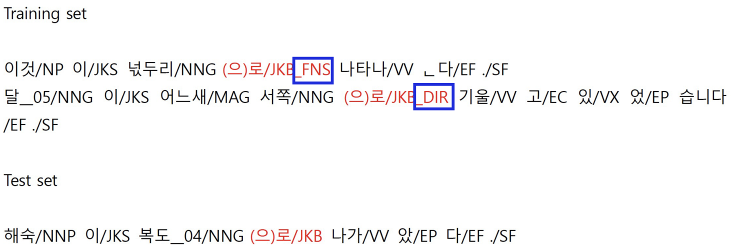
What is Sejong corpus?

- 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).

Description for input

- A portion of Sejong corpus (Shin, 2008), with semantic annotations of –(u)lo cross-verified by three native speakers of Korean (*k*= 0.95).
- Data: 2,100 sentences
  - *-(u)lo*: Final state(700), Instrument(700), Direction(700)

Creating training and test sets



Training set

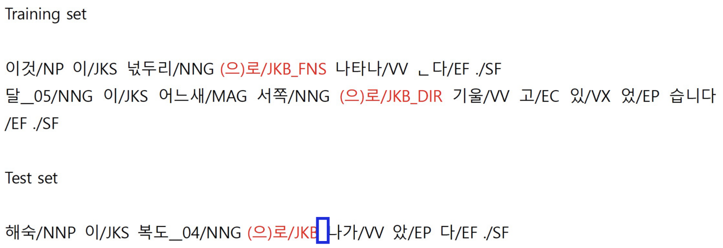
이것/NP 이/JKS 넋두리/NNG (으)로/JKB_FNS 나타나/VV ㄴ다/EF ./SF
달_05/NNG 이/JKS 어느새/MAG 서쪽/NNG (으)로/JKB_DIR 기울/VV 고/EC 있/VX 었/EP 습니다
/EF ./SF

Test set

해숙/NNP 이/JKS 복도_04/NNG (으)로/JKB 나가/VV 았/EP 다/EF ./SF

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

Creating training and test sets



Training set

이것/NP 이/JKS 넋두리/NNG (으)로/JKB_FNS 나타나/VV ㄴ다/EF ./SF

달_05/NNG 이/JKS 어느새/MAG 서쪽/NNG (으)로/JKB_DIR 기울/VV 고/EC 있/VX 었/EP 습니다 /EF ./SF

Test set

해숙/NNP 이/JKS 복도_04/NNG (으)로/JKB 나가/VV 았/EP 다/EF ./SF

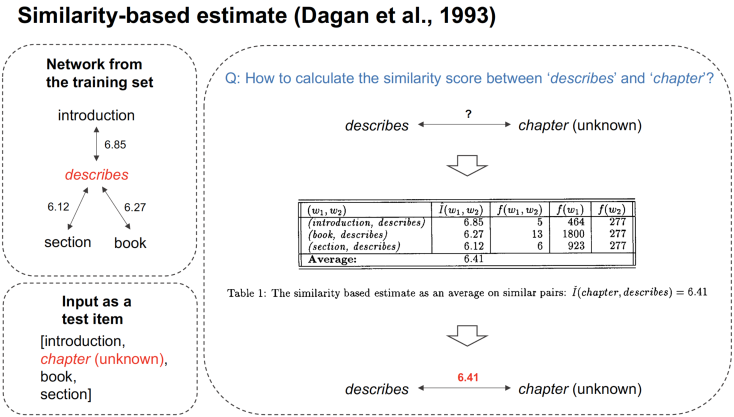
Figure: Example sentences used in the model training and testing (-*(u)lo*)

Methods

Word embedding model: PPMI-SVD

- **Model training**: Adapting a distributional semantic model (Harris,1954), an unsupervised learning algorithm was devised by combining Singular Value Decomposition with Positive Pointwise Mutual Information (i.e., PPMI-SVD).
- **Classification**: similarity-based estimate (Dagan et al., 1993) by calculating cosine similarity scores between -*(u)lo* and its co-occurring content words.

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



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

Network from the training set

Input as a test item

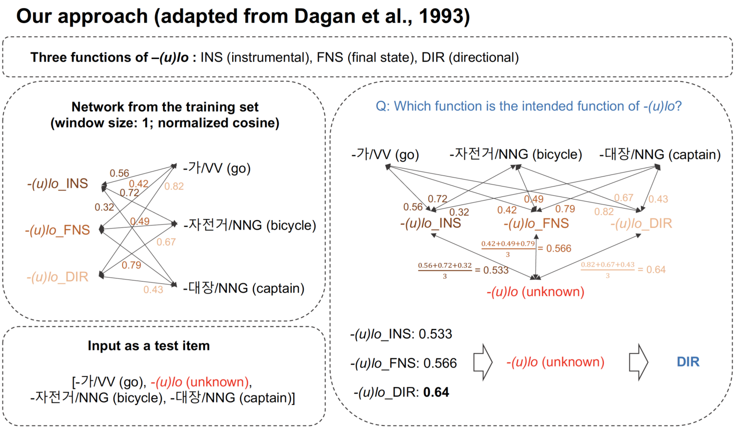
[introduction,
chapter (unknown),
book,
section]

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: $\check{I}(chapter, describes) = 6.41$

Our approach (adapted from Dagan et al., 1993)



Our 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*?

-*(u)lo*_INS: 0.533

-*(u)lo*_FNS: 0.566

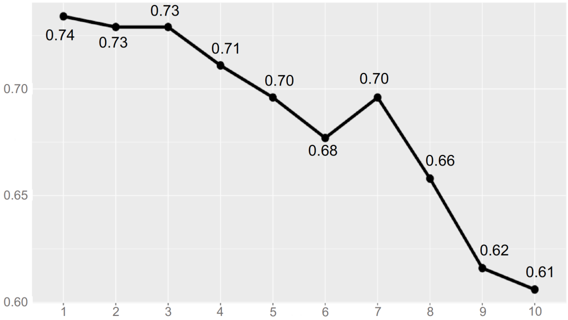
-*(u)lo*_DIR: **0.64**

-*(u)lo* (unknown)

DIR

Results

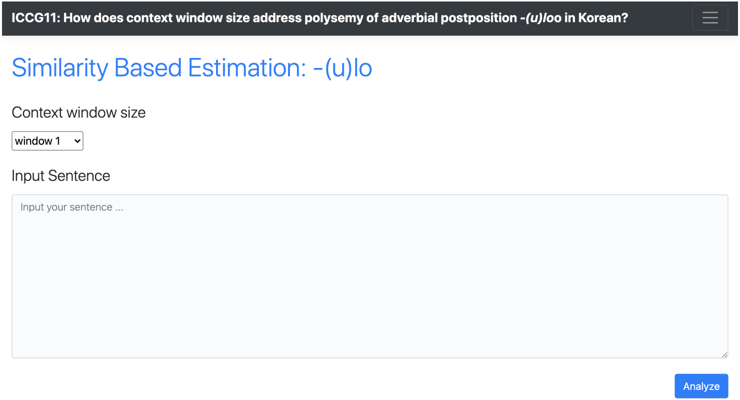
Classification: PPMI-SVD



X-axis: window size
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.

Evaluation



ICCG11: How does context window size address polysemy of adverbial postposition -*(u)lo*o in Korean?

Similarity Based Estimation: -(u)lo

Context window size

window 1

Input Sentence

Conclusion

- Classification
  - The result aligns with the small-window-size advantage (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.
- Evaluation
  - The size of the window affects the accuracy of polysemy interpretation.

Thank you for listening.