A computational approach to resolve the polysemy of postpositions in Korean



Seongmin Mun UMR 7114 MoDyCo - CNRS, University Paris Nanterre

seongmin.mun@parisnanterre.fr





Introduction

This research features a project on the resolution of polysemy involving Korean postpositions. An adverbial postposition -(u)lo, for instance, is either directional or instrumental (Choo, 2008), as exemplified in (1) and (2).

(1) -(u)lo as directional ('(I) went to the road.')

도로 <i>-(으)</i> 로	갔다.
tolo <i>-(u)lo</i>	ka-ass-ta.
road-DIR	go-PST-SE

Visualization

Our visualization selects postposition, function, distributional semantic model, and context window size. Visualization is divided into three parts. The first part provides a distributional semantic map using TSNE to represent the distribution of co-occurring words on a reduced two-dimension depending on the selected options. The second part shows the sentences in concordance with the selected postposition and its function. The third part calculates the similarity between postposition and co-occurring words using cosine formula and provides the results with a force-directed graph and table. For a demo, see https://seongmin-mun.github.io/PostNetwork.ko/index.html

(2)	- <i>(u)lo</i> as instru	umental ('(I) went by bicycle')
	자전거 <i>-(으)</i> 로	갔다.
	cacenke <i>-(u)lo</i>	ka-ass-ta.
	bicycle-INS	use-PST-SE

Previous research computational linguistics has attempted to resolve the polysemy of postpositions in Korean (Shin et al., 2005; Kim et al., 2006). However, due to their focus on computational power to the detriment of linguistic expertise, the models have done a poor job at resolving polysemy. To tell the distinct meanings apart, our method consists in (a) limiting the scope to three of the most frequent postpositions (*-ey*, *-eyse*, and *-(u)lo*) as found in the Sejong Corpus (Shin, 2008), and (b) implementing three kinds of distributional semantic models:

- SVD (Eckart & Young, 1936)
- a combination of PPMI & SVD (Turney & Pantel, 2010)
- SGNS (Tomas et al., 2013)

The annotated corpus designed to represent the functions was used as trianing data set, and the optimal model was calculated by comparing the recognition accuracy of the learning models obtained by the combination of the distributional semantic models and context window sizes.

Data Processing

PostNetwork.ko	6					ົ	()	? ()	
Postposition -에 (-ey) 🗘	Distributional semantic map with T-SNE			Force directed graph					
Function LOC (장소, Location) \$ Method SVD \$ Context window size window 1 \$ Node size	· 특/kul/NNG · 특석부/pwuksepwu/NNG · 중양부/cwungangpwu/NNG · 경경·동성·ድ/nskife@ik/Mc/NNP · 여리/tasi/MAG · 호르_01/hulu_01/VV · 대왕(201/hulowan/NNG · 주/cwu/VV · 주/cwu/VV · · · · · · · · · · · · · · · · · · ·						····································		
frequency \$							Nearest words		
POS Text switch On/Off				● 書を ● この ● こ		id 0 1	name 에/ey/JKB 말01/mal01/NNG	simil 1 0.99	
Select POS ■ NNG (일반명사, Common				3	읍01/up01/NNG 안/an/MAG	0.99			
 NNP (고유명사, Proper Nc NNB (의존명사, Bound No 	I	d name	function	sentences	lexeme with POS	5	혼자01/honca01/NNG	0.99	
 ○ NP (대명사, Pronoun) ○ NR (수사, Numeral) 	2	-에 (-ey) 2 -에 (-ey)	LOC	보주성 거란이 압록강 동쪽 연안에 쌓은 성. 마침 총무님이 사무실에 남아 있었다.	보주성/NNP 거란/NNP 이/JKS 압록강/NNP 동쪽/NNG 연안02/NNG 에/JKB 쌓/VV 은/ETM 성 마침/MAG 총무/NNG 님/XSN 이/JKS 사무실/NNG 에/JKB 남/VV 아/EC 있/VX 었/EP 다/EF ./SF	6 7	똣/mos/MAG 누구/nwukwu/NP	0.99	
 ○ VV (동사, Verb) ○ VA (형용사, Adjective) ○ MAQ (양)바닐 나 Queneral 4 	\$	3 -에 (-ey)	LOC	거실에 불이 켜져 있었다.	거실02/NNG 에/JKB 불01/NNG 이/JKS 켜01/VV 어/EC 지/VX 어/EC 있/VX 었/EP 다/EF	8	어머니/emeni/NNG	0.99	
 MAG (일만무사, General A MAJ (접속부사, Conjuncti JKB (부사격조사, Adverbia 	Ę	-에 (-ey) 5 -에 (-ey)	LOC	개울 옆 언덕 밑에 샘물이 졸졸졸 흘렀습니다. 옥점의 코밑에 땀방울이 방울방울 맺혔다.	개울/NNG 옆/NNG 언덕/NNG 밑01/NNG 에/JKB 샘물/NNG 이/JKS 졸졸졸/MAG 흐르01/\ 옥점/NNP 의/JKG 코밑/NNG 에/JKB 땀방울/NNG 이/JKS 방울방울/NNG 맺히/VV 었/EP 다/EF .	9	형01/hyeng01/NNG) 문제06/mwuncey06/NNG	0.99	

Figure 2: The interface of our visualization represents the relation and network to the co-occurring words obtained from trained models

Evaluation

We conducted case study limited to adverbial postposition -(u)lo to assess the performance of the models. The learning curves shows the accuracy of how accurately function of adverbial postposition -(u)lo is classified (figure 3). The performance of the model for SGNS outranked that the other models and It is not significantly underperforming in every context window size, which aligns with findings of previous research (Levy et al., 2015). PPMI&SVD yields high performance in context window size 1, accuracy decreased as context window size became larger. It appears that the size of the context window influences the model performance for PPMI & SVD. This, therefore, means that PPMI&SVD tends to induce more syntactic representations since it has the best performance in context window size 1 that the information comes from immediately nearby words (e.g., Jurafsky, 2019; Lison & Kutuzov, 2017).

The meaning of a word in a sentence can be approximated by its relation to the co-occurring words (the Distributional Hypothesis). It is thus assumed that we can identify the polysemy of a word based on information obtained from surrounding words and the network of mutual associations between polysemous word and the surrounding words with which they occur. In this study, we focus on three postpositions (*-ey, -eyse,* and *-(u)lo*) that frequently appear in the Sejong Corpus. The adverbial postposition *-ey* has 8 functions, *-eyse* had 2 and *-(u)lo* 6. The models were created by a combination of three distributional semantic models and context window sizes and the dimensions of word embedding were reduced to two dimensions using TSNE.





Figure 1: Data processing structure. Framework for training models using unsupervised learning

Figure 3: Learning curves for each distributional semantic models (*-(u)lo*) X-axis: context window size; Y-axis: accuracy (%)