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-*(으)*로	갔다.
tolo-*(u)lo*	ka-ass-ta.
road-DIR	go-PST-SE

(2) *-(u)lo* as instrumental ('(I) went by bicycle')

자전거-*(으)*로	갔다.
cacenke-*(u)lo*	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

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.

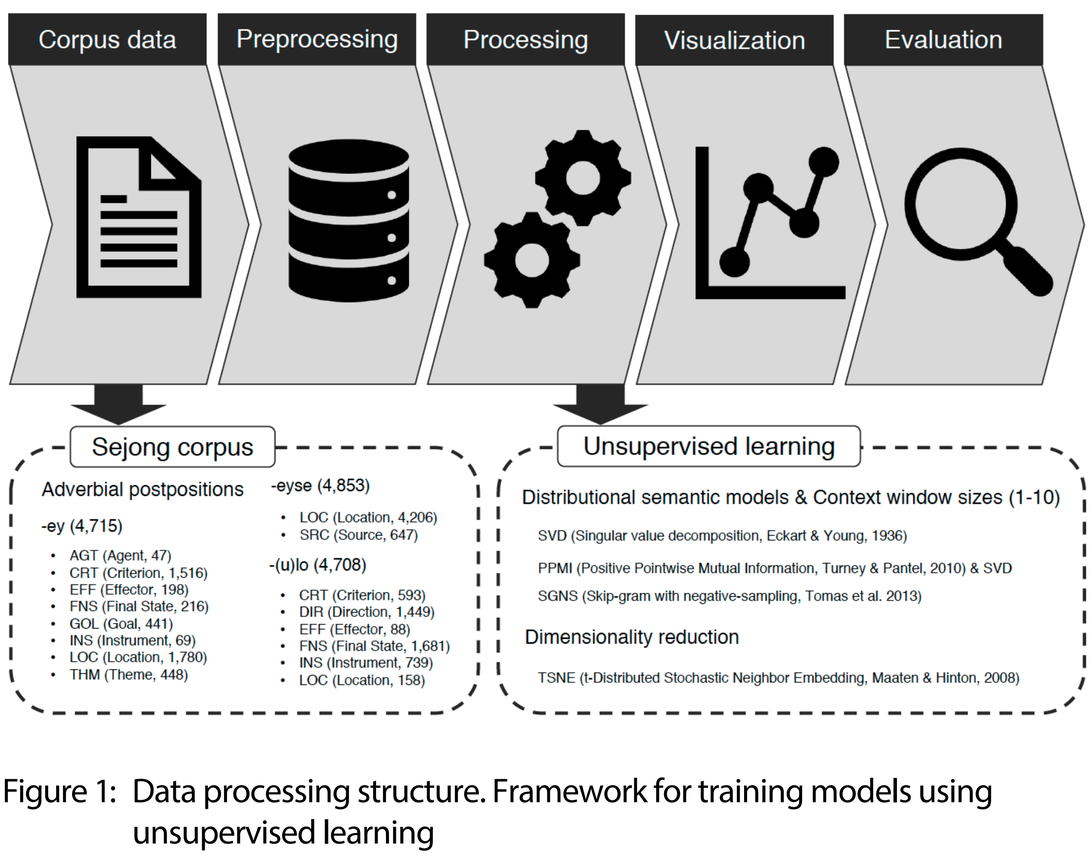

Figure 1: Data processing structure. Framework for training models using unsupervised learning

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

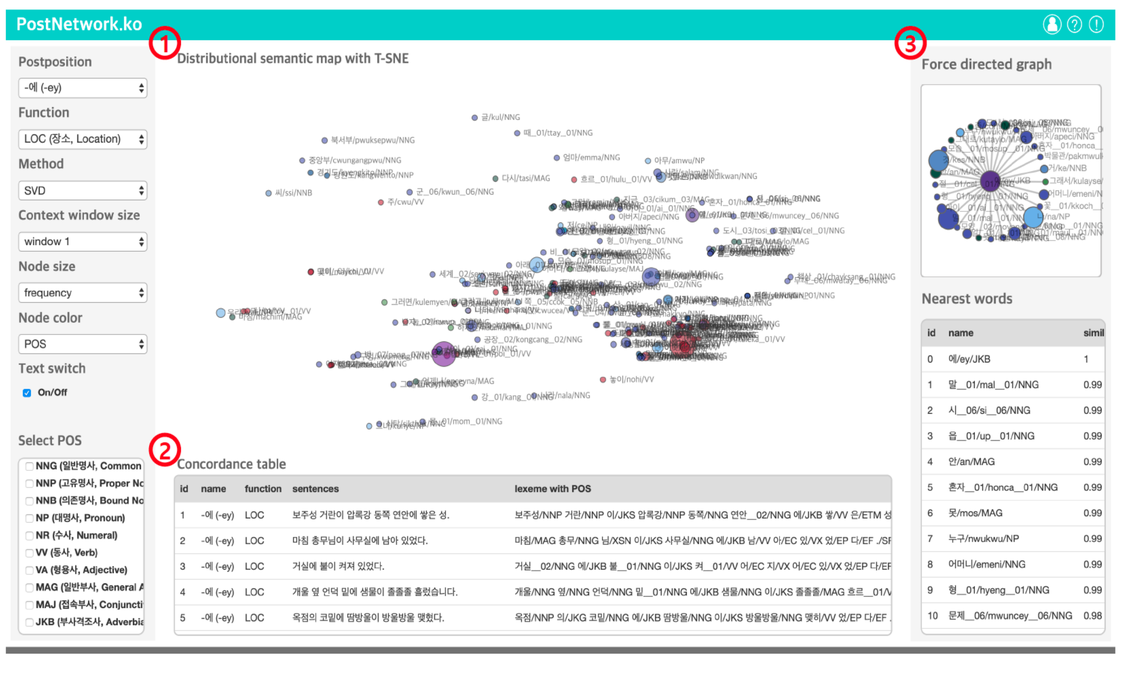

Figure 2: The interface of our visualization represents the relation and network to the co-occurring words obtained from trained models

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).

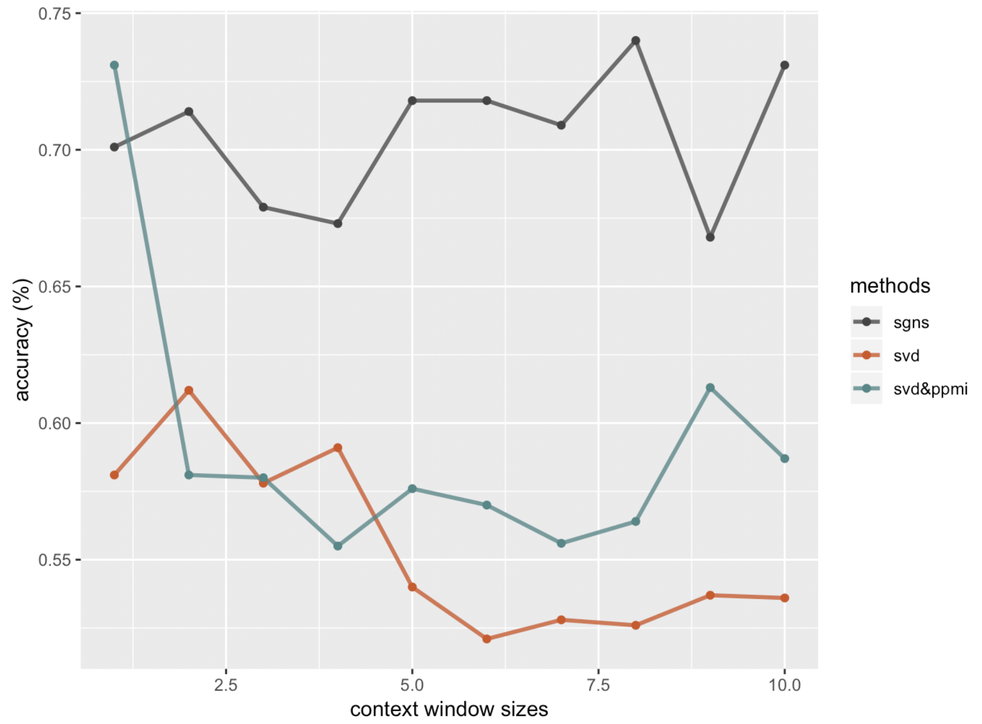

Figure 3: Learning curves for each distributional semantic models (*-(u)lo*) X-axis: context window size; Y-axis: accuracy (%)