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

Polysemy, one type of ambiguity, occurs when one form delivers multiple meanings/functions (Glynn and Robinson, 2014).
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 kolkomk-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)
<table>
<thead>
<tr>
<th>Polysemy in Korean</th>
</tr>
</thead>
</table>

**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).
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

<table>
<thead>
<tr>
<th>Study</th>
<th>Corpus type</th>
<th>Data size</th>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bae et al. (2015)</td>
<td>Korean PropBank</td>
<td>4,882 sentences</td>
<td>One-hot encoding + Structural SVM &amp; FFNN (Feed-Forward Neural Network)</td>
<td>0.75</td>
</tr>
<tr>
<td>Kim &amp; Ock (2016)</td>
<td>Sejong corpus</td>
<td>59,220 sentences</td>
<td>One-hot encoding + CRF (Conditional Random Fields Model)</td>
<td>0.83</td>
</tr>
<tr>
<td>Lee et al. (2015)</td>
<td>Korean PropBank</td>
<td>4,882 sentences</td>
<td>Word2vec (SGNS) + Structural SVM (Support Vector Machine)</td>
<td>0.77</td>
</tr>
<tr>
<td>Mun &amp; Shin (2020)</td>
<td>Sejong corpus</td>
<td>2,100 sentences</td>
<td>PPMI &amp; SVD + Similarity-based estimate</td>
<td>0.74</td>
</tr>
<tr>
<td>Park &amp; Cha (2017)</td>
<td>Sejong corpus</td>
<td>14,335 sentences</td>
<td>Word2vec (SGNS) + CRF</td>
<td>0.77</td>
</tr>
<tr>
<td>Shin et al. (2005)</td>
<td>Sejong corpus</td>
<td>4,355 sentences</td>
<td>Word token-based embedding + SVM</td>
<td>0.71</td>
</tr>
<tr>
<td>Yoon et al. (2016)</td>
<td>Korean PropBank</td>
<td>4,714 sentences</td>
<td>One-hot encoding + Bidirectional LSTM-CRFs</td>
<td>0.66</td>
</tr>
</tbody>
</table>
Context window: a range of words surrounding a target word, affecting the determination of its characteristics (Lison and Kutuzov, 2017).
**Question:** How does context window address polysemy interpretation in Korean?
Corpus

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

이/JKS 넣두리/NNG (으)로/JKB FNS 나타나/VV 냝다/EF ./SF
dal_05/NNG 이/JKS 어느새/MAG 서쪽/NNG (으)로/JKB DIR 기울/VV 고/EC 있/VX 있/EP 슬니다./EF ./SF

Test set

해숙/NNP 이/JKS 북도_04/NNG (으)로/JKB 나가/VV 있/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 넣두리/NNG (으)로/JKB_FNS 나타나/VV 냈다/EF ./SF
d알_05/NNG 이/JKS 어느새/MAG 서쪽/NNG (으)로/JKB_DIR 가을/VV 고/EC 있/VX 였/EP 습니다./EF ./SF

Test set

해숙/NNP 이/JKS 북도_04/NNG (으)로/JKB나가/VV 았/EP 다/EF ./SF

Figure: Example sentences used in the model training and testing (-(u)lo)
How does context window size address polysemy of adverbial postposition -(u)lo in Korean?

Chosun University
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)

Q: How to calculate the similarity score between ‘describes’ and ‘chapter’?

Input as a test item
[introduction, chapter (unknown), book, section]

Network from the training set
introduction
\[\text{describes} \quad 6.12 \quad 6.27 \quad \text{section} \quad \text{book}\]

Q: How to calculate the similarity score between ‘describes’ and ‘chapter’?

\[
\begin{matrix}
\text{Input as a test item} & \text{[introduction, chapter (unknown), book, section]} \\
\end{matrix}
\]

\[
\begin{array}{c|c|c|c|c}
(w_1, w_2) & \hat{I}(w_1, w_2) & f(w_1, w_2) & f(w_1) & f(w_2) \\
\hline
(\text{introduction, describes}) & 6.85 & 5 & 464 & 277 \\
(\text{book, describes}) & 6.27 & 13 & 1800 & 277 \\
(\text{section, describes}) & 6.12 & 6 & 923 & 277 \\
\hline
\text{Average:} & 6.41 & & & \\
\end{array}
\]

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

\[
\begin{array}{c|c|c|c|c|c|c|c}
(w_1, w_2) & \hat{I}(w_1, w_2) & f(w_1, w_2) & f(w_1) & f(w_2) \\
\hline
(\text{introduction, describes}) & 6.85 & 5 & 464 & 277 \\
(\text{book, describes}) & 6.27 & 13 & 1800 & 277 \\
(\text{section, describes}) & 6.12 & 6 & 923 & 277 \\
\hline
\text{Average:} & 6.41 & & & \\
\end{array}
\]

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

Chosun University
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)

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

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

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

Input as a test item

[-가/VV (go), -(u)lo (unknown), -자전거/NNG (bicycle), -대장/NNG (captain)]
How does context window size address polysemy of adverbial postposition -(u)lo in Korean?
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 in Korean?

Similarity Based Estimation: -(u)lo

Context window size

window 1

Input Sentence

Input your sentence ...

Analyze
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.