Context window and polysemy interpretation: A case of Korean adverbial postposition –(u)lo

Seongmin Mun¹ & Gyu–Ho Shin²
Université Paris Ouest Nanterre La Défense¹, University of Hawaii at Manoa²

Introduction

• Korean postposition –(u)lo has three major functions - INS (instrumental), FNS (final state), DIR (directional)

(1b) –(u)lo as INS (instrument)

na-nun kamca-ulo kulah-lo ssel-ess-ta.

I-TOP potato-ACC knife-INS cut-PST-DECL

‘I cut a potato with a knife.’

(1a) –(u)lo as FNS (final state)

Yongho-ulo pancang-ulo senchwul-ha-ass-ta.

Yongho-ACC class.leader-FNS election.do-PST-DECL

‘(We) elected Yongho as the leader of the class.’

(1c) –(u)lo as DIR (direction)

Mia-ka mikiwul-ulo tteta-sst-ta.

Mia-NOM America-DIR leave-PST-DECL

‘Milan left to America.’

• Assumption: construal of a polysemous word occurs in conjunction with a series of words, delivering various frame-semantic meanings (Goldberg, 2006) and yet purporting similar interpretations (Harris, 1954)

• Context window: a range of words surrounding a target word, affecting the determination of its characteristics

• Question: how does context window address polysemy interpretation in Korean?

Methods [1]

• Input: A portion of Sejong corpus (Shin, 2008), with semantic annotations of –(u)lo cross-verified by three native speakers of Korean (κ = 0.95)

  • Training set: 1,890 sentences
  • Test set: 210 sentences

• 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

• Classification: similarity-based estimate (Dagan et al., 1993) by calculating cosine similarity scores between -(u)lo and its co-occurring content words

Abbreviation: 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.

REFERENCES


• Shin, hyo-Pi. 2008. The 2nd sejong project : with a focus on sejong electronic lexicon of korean and the knc(korean national corpus), in In The 3rd International Joint Conference on Natural Language Processing.

Corresponding Author: Seongmin Mun(simon.seongmin.mun@gmail.com)

Results

• 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

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