Coreference resolution of Korean anaphoric zero objects: Towards an unsupervised learning approach

Korean is a language where grammatical role-denoting expressions such as the subject, the object, and the (time/place) adverbial, are phonologically suppressed, as follows:

(1) A: cheli-ka eceycenyek-ey phica-lul mek-ess-ni?

Cheli-Nom last night-in pizza-Acc ate-Past-Interrogative

'Did Cheli eat pizza last night?'

B: ye, [e] [e] [e] mek-ess-e-yo.

Yes, eat-Past-Informal-Hearer+Honor

'Yes, he ate it at that time.'

In the dialogue between speakers A and B, B's sentence contains the phonologically null/empty expressions (indicated by [e]'s) that substitute for the subject, the object, the adverbial underlined in A's sentence. Their syntactic/structural identity has been debated for a long time. Since the phonologically null/empty expressions are translated into the overt pronouns or pro-forms in English, they were analyzed as zero/null/empty pronouns/pro-forms (Huang (1984)). Put in another way, they were taken to derive from pro-drops or dropping of pronouns. More recently, an alternative analysis was provided (Takahashi (2008)): Since repeated sentence expressions can undergo ellipsis, the phonologically null/empty expressions such as those in (1B) were analyzed as deriving from elision of a subject/object/adverbial.

Aside from the issue of identifying their exact status, another issue bearing on them is how these apparently invisible expressions undergo interpretation or what people call coreferential resolution for them. Of course, this issue is intimately related to the first issue, because the identification of their exact status is sure to be a prerequisite for investigating the right way of finding their correct antecedents. Assuming with the traditional view of grammar that both pro-forms (including pronouns) and ellipsis belong to the same larger domain of anaphora (i.e., the use of an expression the interpretation of which depends upon another expression in context (its antecedent or postcedent)), in this paper we concentrate on the coreference resolution of anaphoric zero objects (AZO(s)) among zero expressions in sentences of Korean.

More specifically, we first examine what features in grammar or discourse theories are instrumental in best characterizing coreference resolution of AZOs in Korean. Acknowledging that the Centering theory or the features postulated in it are at present the most effective in determining the correct antecedents of AZOs, we adopt the corpus that Park et al. (2015) annotated with such features. Park et al. (2015) in fact used the corpus to take a supervised learning approach to coreference resolution of Korean AZOs. Departing from Park et al. (2015), in this paper we show

that a unsupervised probabilistic learning approach may outperform a supervised learning approach in coreference resolution of Korean AZOs.

We initially train our unsupervised resolver on overt object pronouns using the Expectation-Maximization (EM) algorithm. After training, we then apply the resulting model to resolve AZOs. More specifically, given (i) an AZO z and (ii) context features, we determine its antecedent from the set C of candidate antecedents of z as follows: (: hidden data). To fully specify our model, we reply on context features employed in Park et al. (2015). Experiments demonstrate that our unsupervised model outdoes its rivaling supervised counterparts in performance when resolving AZOs in the given corpus.