SzegedKoref: A Manually Annotated Coreference Corpus of Hungarian

Veronika Vincze[1,2], Klára Hegedűs[3], Richárd Farkas[1]

[1]University of Szeged, Department of Informatics
Szeged Árpád tér 2., e-mail: {vinczev,rfarkas}@inf.u-szeged.hu
[2]MTA-SZTE Research Group on Artificial Intelligence
[3]University of Szeged, Institute of Psychology
e-mail: klarahegedus92@gmail.com

Here we introduce the SzegedKoref corpus, in which coreference relations are manually annotated. For annotation, we selected the texts of Szeged Treebank, the biggest treebank of Hungarian with manual annotation at several linguistic layers.

We present the annotated texts, we describe the annotated categories of anaphoric relations, and we offer several examples of each annotated category. Currently, the corpus contains 309 sentences and 9,782 tokens from the newspaper domain and 3,712 sentences and 45,981 tokens from the student essay subcorpus. Altogether, there are 4021 sentences and 55,763 tokens in the current version of the corpus, however, the annotation process is still going on, and the amount of annotated texts is continuously growing.

In Hungarian, zero pronouns also mean a challenge to coreference resolution systems. We automatically inserted zero pronouns into the text before the manual annotation process, so they are also annotated in the data.

There are 2191 anaphoric chains in the student essay subcorpus and 265 in the newspaper domain, adding up to 2456 anaphoric chains altogether. The most frequent types of anaphor are pronominal anaphors and repetition, indicating that automatic coreference resolution systems should pay extra attention to these categories, together with zero pronouns.

Due to its size, the corpus can be exploited in training and testing machine learning based coreference resolution systems, which we would like to implement in the near future.