Constructing a Nominal Hungarian WordNet Ontology with Automatic Methods

Márton Miháltz

MorphoLogic Kft.
1118 Budapest, Késmárki utca 8.
mihaltz@morphologic.hu

Keywords: WordNet construction, Hungarian Nominal WordNet, automatic extraction of semantic information

Abstract

This paper presents recent results of the ongoing project aimed at creating the nominal database of the Hungarian WordNet. We present 9 different automatic methods, developed for linking Hungarian nouns to WN 1.6 synsets. Nominal entries are obtained from two different machine-readable dictionaries, a bilingual English-Hungarian and an explanatory monolingual (Hungarian). The results are evaluated against a manually disambiguated test set. The final version of the nominal database is produced by combining the verified result sets and their intersections when confidence scores exceeded certain threshold values.

Our basic strategy was to attach Hungarian entries of a bilingual English-Hungarian dictionary to the nominal synsets of Princeton WordNet (WN), following the so-called extension approach. This way, the synsets formed by the Hungarian nouns can inherit the English WN semantic relations. In order to achieve this, we used heuristic methods, developed partly by previous similar projects and partly by us, which rely on information extracted from two machine-readable dictionaries (MRDs). This approach relies on the assumption that nominal conceptual hierarchies, which describe the world, would be similar across English and Hungarian languages to a degree which is sufficient for producing a preliminary version of our WordNet.

The first MRD we used was a bilingual English-Hungarian (17,700 Hungarian nominal entries, 12,400 English equivalents), serving as the basis of the attachment procedure. The *Magyar Értelmező Kéziszótár* (EKSz) monolingual explanatory dictionary (42,000 nominal headwords, 64,000 different definitions) was used to gain semantic information in order to assist the disambiguation heuristics

A number of these methods relied on structural information extracted from connections in the bilingual dictionary and WN (3 heuristics), and morpho-semantic information gained from the Hungarian side of the bilingual (1 heuristic). For the further support of the task we used the morphologically analyzed nominal definitions of the explanatory dictionary. The synonyms and hypernyms extracted this way were used by 4 additional heuristics. Further help was provided by the Latin translation equivalents available for a number of EKSz entries (1 heuristic).

In order to evaluate the precision of the results from the different information sources, we randomly selected 400 nouns from the Hungarian side of the bilingual and accomplished manual disambiguation against WN for these. By combining result sets whose precision scores were estimated with the help of this gold standard, we produced two preliminary versions of the Hungarian nominal database. The first version covers 7,900 nominal entries and 6,500 synsets (with 75% estimated precision), and the second, larger version covers 13,600 entries and 15,100 synsets (with 63% estimated precision).