Development of the Hungarian WordNet Ontology and its Application to Information Extraction

Márton Miháltz
mihaltz@morphologic.hu
MorphoLogic
Budapest, Orbánhegyi út 5, H-1126, Hungary

This paper presents an outline of the construction process of the Hungarian WordNet Ontology, and the description of an information extraction application utilizing the ontology. The work presented was produced by three Hungarian natural language technology research institutions (The Research Institute for Linguistics of the Hungarian Academy of Sciences, Department of Informatics, University of Szeged, and MorphoLogic) in a 3-year project funded by the European Union ECOP program (GVOP-AKF-2004-3.1.1.)

The Princeton WordNet (WN) linguistic ontology ([1]) has become a standard and an invaluable semantic resource within the natural language technology community. The WN lexical semantic network consist of nodes called synsets (sets of synonymous words that are interchangeable in a context) corresponding to linguistic concepts and connecting edges corresponding to semantic relationships like hypernymy (is-a relationship), meronymy (part-of relationship), antonymy etc. The EuroWordNet (EWN) project ([5]) extended the WN architecture to a multilingual level, with the synsets of the English WN serving as interlingua (ILI) among the concepts of the various other languages. A common starting set (Common Base Concepts) was implemented in each participating language and then was expanded individually in a top-down manner by each partner. In addition to the 11 European languages covered by EWN, the BalkaNet project ([4]) several years later introduced connected Wordnets for 5 more Southeast-European languages.

The Hungarian WordNet (HuWN) project follows the BalkaNet project's resources: Princeton WordNet 2.0 as ILI, 8500 base concept synsets as a starting point and the VisDic XML-based ontology/dictionary editor. We also decided to integrate existing semantic resources into HuWN: on the one hand, we tried to map each Hungarian synset to a sense in the EKSz Hungarian explanatory dictionary, in order to obtain definitions, and on the other hand, for each verbal synset we registered corresponding entries in our existing verb frame description lexicon.

For nouns, adjectives and adverbs, the work followed the so-called expansion approach, which means we took English WordNet as a starting point for our concept networks. We used several machine-translation heuristics ([2]) to obtain a rough translation of the English synsets, which were then all manually examined, corrected and extended as necessary, with adaptation to the semantic conditions of the Hungarian language. For verbs, this approach proved to be unsustainable because of the major differences between the English and Hungarian verb typing systems. In this case, we used a mixed approach, by translating only a subset of the common concepts and creating the rest from scratch from frequent items in corpora. We also added new semantic relations and the so-called nucleus structure ([3]) in order to represent aspects of verb meanings unique to Hungarian.

After the Hungarian representation of the BCS synsets for nouns and adjectives was complete, we added the so-called local base concepts, frequent and basic concepts not yet covered and which were formed from frequent genus words in the EKSZ definitions. After this, several iterations of concentric extension followed, by adding and translating 1st-level hyponyms selected by corpus frequencies. In addition to this, for a few important domains such as geography, languages, money, public administration etc. we decided to implement the entire vocabulary of PWN by importing complete hyponym trees in these domains. For the

sake of the information extraction tool described below, we also added business domain concepts from the most frequent terms in a short business news corpus. We also added named entity lists covering Hungarian geographical regions, places, person names etc. For adverbs, we translated the 1000 most frequent senses in English sense-tagged corpora, then restructured the concepts to reflect Hungarian specifics. This way, the HuWN ontology contains 40,000 synsets (31,600 noun, 3,300 verb, 4,100 adjective and 1000 adverb synsets).

Our information extraction engine was developed to identify the event type (such as selling, provatisation, litigation etc.) and the participating entities (eg. the seller, buyer and the price in a sale) expressed in short business news texts.

We created so-called event frame descriptions manually after analyzing the business news corpus. Each frame description defines an event, and contains participants that correspond to the main verb and its typical arguments. In the implementation of the IE engine, a parser first identifies the main syntactic constituents in the input text, then it tries to match these to the elements of the candidate event frames. There are several kinds of constraints that need to be satisfied for a match. Lexical constraints can either be specified by strings, or by synset ids corresponding to hyponym subtrees of the HuWN ontology. Semantic constraints are expressed by so-called semantic meta-features, or basic semantic categories, such as "human", "company", "currency" etc. that are mapped to HuWN synsets and all their hyponyms. There are also syntactic and morphologic constraints, which are checked against the output of the parser and the underlying morphologic analyzer. Finally, the IE engine ranks the candidate event frame matches for the output according to the ratio of event participants matched.

In this approach, the use of ontological categories allows for a simpler and better understandable layout for the event frames. The main advantage of the use of synset ids and semantic types lies in the fact that the vocabulary of the IE engine can be easily customized and extended by adding new concepts to the ontology, without the need to modify the original event frame descriptions.

References

[1] Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998)

[2] Miháltz, M., Prószéky, G.: Results and Evaluation of Hungarian Nominal WordNet v1.0. In Proceedings of the Second International WordNet Conference (GWC 2004), Brno, Czech Republic, pp. 175-180 (2004).

[3] Moens, M., Steedman, M.: Temporal ontology and temporal reference. Computational Linguistics Volume 14 Issue 2 (1988) 15-28

[4] Tufiş, D., D. Cristea, S. Stamou (2004): BalkaNet: Aims, Methods, Results and Perspectives, A General Overview. In: Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1–2).

[5] Vossen, P. (ed.): EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document Deliverable D032D033/2D014, (1999).