Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction

Project Report

Zoltán Alexin[1], János Csirik[1], András Kocsor[2], Márton Miháltz[3], György Szarvas[1]

[1] University of Szeged, Department of Informatics,
6720 Szeged, Árpád tér 2., Hungary,
[2] MTA-SZTE, Research Group on Artificial Intelligence,
6720 Szeged, Aradi Vértanúk tere 1., Hungary,

{alexin,csirik,kocsor,szarvas}@inf.u-szeged.hu

[3] MorphoLogic Kft.
1126 Budapest, Orbánhegyi út 5., Hungary

mihaltz@morphologic.hu

Abstract

This report present a recent Hungarian project started in Spring of 2005. The goals of the project are to produce a Hungarian version of the EuroWordNet ontology database, to extend it with concepts specific to business domain, and to develop a demonstration version of an ontology-based Information Extraction (IE) system. The system will extract summarized data from short business articles concerning company fusions, acquisitions, profits, new products, new plants etc. A consortium of three leading Hungarian human language technology institutions won substantial governmental support lasting till 2007.

1 Project data

The work began in 2005 and finishes in 2007. The consortium partners are: University of Szeged, Department of Informatics, Human Language Technology Group (coordinator); Institute of Linguistics at the Hungarian Academy of Sciences, Department of Corpus Linguistics, and the MorphoLogic Ltd. Budapest. This consortium had successful cooperation during previous R&D projects. They produced the versions of Szeged Corpus[1] [1], the Szeged Treebank [3], and some applications like a preliminary information extraction software.

2 Aims and scope of the project

The Hungarian ontology database bases on the multilingual architecture of EuroWordNet (EWN) [8] and its South-Eastern European successor, BalkaNet (BN) [6]. It will contain Hungarian nouns, verbs and adjectives, which are connected to the English synsets of the Princeton WordNet 2.0 (PWN), the Inter-Lingual Index (ILI) defined in the BalkaNet project. The Base Concept Set (BCS) of the BalkaNet has a wider coverage (BN: 8,516 synsets; EWN: 1,310 synsets), therefore the consortium chose BalkaNet as a starting point of the project. In addition, the latter provides openly available tools and resources.

The first phase of the project will follow the expand approach. First, the 8,516 English synsets comprising the BCS are translated into Hungarian synsets. Aiding the translation by automatic methods [5] developed earlier will also be considered. The obtained results should undergo thorough manual checking and editing. Adding new concepts upon corpus-statistics measures is investigated. In the second phase of the work, this core database is extended in a top-down fashion with additional Hungarian synsets. Monolingual resources such as explanatory dictionary and thesauri can ease the manual work.

Extending the ontology database with pointers, either to the Hungarian explanatory dictionary, by the help of which one will be able to obtain sense definitions; or to the verb-frame lexicon by which one can reach semantic and syntactic descriptions of verb argument structures for more than 17,000 entries is a reasonable issue. The quality of the database is continuously monitored by automatic validation and consistency-checking methods developed in the BN project.

The core technology of IE implemented by the consortium bases on the application of semantic frames [4], and detailed linguistic (morphologic and syntactic) analysis. Semantic frames are formal abstractions of events in which actors in some decisive roles do something in circumstances described by the event arguments/parameters. IE means labeling of sentence constituents that identify the event described in the text, and its most relevant arguments. Identification is done by pattern matching, that is, given an incoming sentence tree

[1] http://www.tei-c.org/Applications/sz01.xml

decorated by linguistic attributes is checked against semantic frames. If there is a considerably good fit, then an algorithm collects the information appear in the slots defined by the semantic frame.

Domain specificity is a requirement for decent analysis of textual data to achieve significant accuracy at each step of natural language processing comprising IE (e.g. segmentation, morphological analysis, named entity recognition, shallow parsing, etc.). The IE demonstration project aims to focus on economy. A part of the Szeged Treebank [3] is used for training and testing, provided it contains 6453 short business news articles collected from MTIEco (MTI, Hungarian News Agency: http://www.mti.hu).

The Hungarian ontology is planned to aid the IE by integrating the role relations and semantic constraints defined in the frames into the ontology in the form of new relations and attributes of synsets (e.g. ROLE_SELLER, ROLE_GOODS, ANIMATE, etc.). This also means the extension of the Hungarian WordNet with a business-specific domain ontology to help IE. This way searching for relevant information and semantic analysis can be performed parallel to acquire the relevant sentence constituents in the following way:

- Target word identification (determines the economic event).
- Identification of those sentence constituents that satisfy all syntactic and semantic requirements set for one of the roles in the event.
- Heuristic algorithms for resolving ambiguities, selecting the best fit semantic frame by the help of ROLE and other relations (hypo/hypernimy, etc.).

3 Results

Previous results:

- Automatic methods have been developed to attach Hungarian nouns of an English-Hungarian bilingual dictionary to English synsets (PWN version 1.6) [5]. Nine different heuristics, based on the information extracted from the bilingual dictionary and the sense definitions of a monolingual explanatory dictionary were applied in order to find the correct candidate English synsets.
- By combining the best quality result sets, we produced different versions of the automatically generated nominal Hungarian WordNet database. The first version covered 12,500 synsets with an estimated precision of 65%, the second covered 7,300 synsets with 75% precision.
- Corpus of short business news articles of size 200.000 words, with full syntactic annotation.
- Prototypes of an IE system that performs extraction in business domain based on shallow parsing. Accuracy: 92.65% precision in identifying the correct economic event, and role accuracy of 70.3% compared to human annotation. In 50.49% of the cases the same semantic frame (same role set) is used by human annotator and the system, which means exact matching.

Ongoing research:

- Conversion of our resources (bilingual and monolingual dictionaries, a thesaurus, verb frame lexicon) to the VisDic XML format introduced by BalkaNet.
- Adaptation of the automatic methods described earlier to translate the 8,516 BCS synsets to Hungarian equivalents.
- Automatic topic classifier based on machine learning methods for business news articles to aid corpus expansion. Artificial Neural Network [2] shows 94.53%, Support Vector Machine classifier [6] shows 95.01% accuracy in selecting domain-relevant articles.
- Corpus expansion with further business news from MTI.
- Learning models for construction of business domain ontology to speed up domain specific extension of Hungrian EuroWordNet.

Bibliography

1. Alexin, Z., Csirik, J., Gyimóthy, T., Bibok K., Hatvani, Cs., Prószéky, G., Tihanyi, L.: Manually Annotated Hungarian Corpus, in Proc. of the Research Note Sessions of the EACL 2003, Budapest, Hungary 15-17 April, pp. 53–56 (2003).
2. Bishop, C. M.: Neural Networks for Pattern Recognition, Oxford University Press Inc., New York (1996).
3. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS Tagged and syntactically Annotated Hungarian Natural Language Corpus. In: proceedings of the TSD-2004 conference, LNAI 3206, Sojka et al., 41–47 (2004).
4. Fillmore, C. J.: Frame semantics and the nature of language, In Annals of the New York Academy of Sciences: Conference on the Origin of Language and Speech, Vol. 280, pp. 20–32, New York (1976).
5. 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).
6. 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).
7. Vapnik, V. N.: *Statistical Learning Theory*, John-Wiley & Sons Inc. (1998).
8. Vossen, P. (ed.): EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document Deliverable D032D033/2D014, (1999).