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

Zoltán Alexin, János Csirik
György Szarvas
University of Szeged,
Institute of Informatics
Árpád tér 2.
H-6720, Szeged, Hungary,
{alexin,csirik,szarvas}@
inf.u-szeged.hu

András Kocsor
MTA-SZTE, Research
Group on Artificial Intelligence
Aradi Vértanúk tere 1.
H-6720 Szeged, Hungary,
kocsor@inf.u-szeged.hu

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

Abstract

This report describes a recent Hungarian project begun in the 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 the business domain, and to develop a demonstration version of an ontology-based Information Extraction (IE) system. The system will extract condensed data from short business articles concerning company mergers, acquisitions, balance reports, new products, new plants and so on. A consortium of three leading Hungarian human language technology institutions won substantial governmental support that will last until 2007.

1 Project Data

The project was started in 2005 and will be finish in 2007. The consortium partners are the University of Szeged, Institute of Informatics, Human Language Technology Group (coordinator); Institute of Linguistics at the Hungarian Academy of Sciences, the Department of Corpus Linguistics, and MorphoLogic Ltd. Budapest. This consortium worked well together in previous R&D projects. They produced different versions of the Szeged Corpus[1], the Szeged Treebank (Csendes et al., 2004), and some applications such as preliminary information extraction system which made use of the Szeged Corpora and Treebank.

2 Aims and scope of the project

The Hungarian ontology database is based on the multilingual architecture of EuroWordNet (EWN) (Vossen, 1999) and its South-Eastern European successor, BalkaNet (BN) (Tuffis, 2004). It will contain Hungarian nouns, verbs and adjectives, which are associated with 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), so the consortium chose BalkaNet as a starting point of the project which, in addition, provides openly available tools and resources.

The first phase of the project follows the expand approach (Vossen, 1999). First, the 8,516 English synsets comprising the BCS are translated into Hungarian synsets. Aiding the translation via automatic methods (Miháltz and Prószéky, 2004) developed earlier is also considered. The results obtained are manually checked and edited. In the second phase of the work this core database will be extended in a top-down fashion with additional Hungarian synsets.

To extend the ontology database with pointers we will apply the Hungarian explanatory dictionary that contains sense definitions and our verb-frame lexicon that contains semantic and syntactic descriptions of verb argument structures of more than 17,000 entries.

The core technology of IE was implemented by the consortium based on the application of semantic frames (Lowe et al., 1997), and detailed linguistic analysis. Semantic frames are formal abstractions of events in which actors in some decisive roles do something in certain circumstances. IE means the labelling of sentence constituents that identify the event described in the text, and its most relevant arguments. Identification is done by pattern matching, that is, an incoming sentence tree decorated by linguistic attributes is checked against semantic frames. If there is a fit, then an algorithm collects the information appearing in the slots defined by the semantic frame. The Hungarian ontology being developed is intended to aid IE by integrating the role relations and semantic constraints defined in the frames into the ontology database in the form of new relations and attributes of synsets (e.g. BUYER, SELLER, GOODS, ANIMATE). This way searching for relevant information and semantic analysis can be performed at the same time as acquiring the relevant sentence constituents. That is, the

- target word identification (determines the economic event).
- identification of those sentence constituents that satisfy all the syntactic and semantic requirements set for one of the roles in the event.

* The work is partially financed by the European Union and the Hungarian Government under the grant: GVOP-3.1.1.-2004-05-0191/3.0 (AKF) in the Economic Competitiveness Operating Program.

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

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 291–292. © Masaryk University, 2005

- heuristic algorithms for resolving ambiguities, selecting the best fit semantic frame with the help of ROLE and other relations like hyponymy or hyperonymy.

3 Conclusions

Automatic methods have been developed to associate Hungarian nouns of an English-Hungarian bilingual dictionary with English synsets. By combining nine different heuristics (Miháltz and Prószéky, 2004), we were able to produce 7,300 Hungarian nominal synsets automatically with an estimated precision of 75%. The participants constructed a corpus of short business news articles of 200,000 words in total with full syntactic annotation. A prototype of an IE system is developed that performs information extraction from short business articles. The accuracy for identifying the correct economic event was 92.65% and role accuracy was 70.3% compared to that for human annotation.

3.1 Ongoing research

We were able to provide Hungarian synonym suggestions for about 50% of the 8,516 BCS synsets with the automatic methods described earlier. This and the untranslated part is being edited using VisDic, which contains all our resources in a common XML format. Other directions of our research include:

- Enriching the corpus with additional business news from MTI.
- Automatic topic classifier based on machine learning methods for business news articles to aid corpus expansion.
- Developing machine learning models for the construction of a business domain ontology, which is an extension of the Hungarian EuroWordNet.

References

Csendes, D., Csirik, J., Gyimóthy, T. (2004) The Szeged Corpus: A POS Tagged and syntactically Annotated Hungarian Natural Language Corpus. In: proceedings of the TSD-2004 conference, LNAI 3206, Ed.: Sojka et al., 41–47.

Lowe J. B., Baker C. F., and Fillmore, C. J. (1997) A frame-semantic approach to semantic annotation. in *Proceedings of the SIGLEX workshop "Tagging Text with Lexical Semantics: Why, What, and How?"* Washington, D.C., USA.

Miháltz, M., Prószéky, G. (2004) 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.

Tufis, 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).

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