The MEANING Project*

German Rigau,
IXA Group
Euskal Herriko Unibertsitatea
Donostia.
{rigau}@si.ehu.es

Eneko Agirre,
IXA Group
Euskal Herriko Unibertsitatea
Donostia.
{eneko}@si.ehu.es

Jordi Atserias
TALP Research Center
Jordi Girona Salgado, 1-3.
08034 Barcelona
{batalla}@talp.upc.es

Resumen: A pesar del progreso que se realiza en el Procesamiento del Lenguaje Natural (PLN) aún estamos lejos de la Comprensión del Lenguaje Natural. Un paso importante hacia este objetivo es el desarrollo de técnicas y recursos que traten conceptos en lugar de palabras. Sin embargo, si queremos construir la próxima generación de sistemas inteligentes que traten Tecnología de Lenguaje Humano en dominios abiertos necesitamos resolver dos tareas intermedias y complementarias: Resolución de la ambiguedad léxica de las palabras y enriquecimeinto automático y a gran escala de bases de conocimiento léxico
Palabras clave: Wordnet, EuroWordnet, MultiWordnet, Adquisición Multilingüe

Abstract: Progress is being made in Natural Language Processing (NLP) but there is still a long way towards Natural Language Understanding. An important step towards this goal is the development of technologies and resources that deal with concepts rather than words. However, to be able to build the next generation of intelligent open domain Human Language Technology (HLT) application systems we need to solve two complementary and intermediate tasks: Word Sense Disambiguation (WSD) and automatic large-scale enrichment of Lexical Knowledge Bases.
Keywords: Wordnet, EuroWordnet, MultiWordnet, Multilingual Acquisition

1 *Project Participants*

- **Universitat Politècnica de Catalunya** Barcelona, Spain
- **ITC-IRST** Trento, Italy
- **University of the Basque Country** Donostia, Spain
- **University of Sussex**, Brighton, UK
- **Irion Technologies B.V.** The Netherlands

The MEANING project is funded by the EU 5^{th} Framework IST Programme. **Coordinator:** German Rigau. Euskal Herriko Unibertsitatea. Donostia. rigau@si.ehu.es

2 *Introduction*

Knowledge Technologies aim to provide meaning to petabytes of information content our societies will generate in a near future. Information and knowledge management systems need to evolve accordingly, to help release next generation of intelligent open domain HLT to deal with the growing potential of the knowledge-rich and multilingual society.

To develop a trustable semantic web infrastructure and a multilingual ontology framework to support knowledge management it is required a wide range of techniques to progressively automate the knowledge lifecycle. In particular this include extracting high level meaning from the large collections of content data and its representation and management in a common knowledge bases.

Even now, building large and rich knowledge bases takes a great deal of expensive manual effort; this has severely hampered HLT application development. For example, dozens of person-years have been invest into the development of wordnets, but the data in these resources is still not sufficiently rich to support advanced concept-based HLT applications directly. Furthermore, resources produced by introspection usually fail to register what really occurs in texts. Applications will not scale up to working in the open domain without more detailed and rich general-purpose and also domain-specific linguistic knowledge. To be able to build the next generation of intelligent open domain HLT application systems we need to solve two complementary intermediate tasks: Word Sense Disambiguation (WSD) and large-scale enrichment of Lexical Knowledge Bases. However, progress is difficult due to the following paradox:

* This research has been partially funded by the European Comission (MEANING IST-2001-34460)

1. In order to achieve accurate WSD, we need far more linguistic and semantic knowledge than is available in current lexical knowledge bases (e.g. wordnets).
2. In order to enrich Lexical Knowledge Bases we need to acquire information from corpora, which have been accurately tagged with word senses.

Providing innovative technology to solve this problem will be one of the main challenges to access Knowledge Technologies.

3 Meaning

Meaning (Rigau et al., 2002) [1] will treat the web as a (huge) corpus to learn information from, since even the largest conventional corpora available (e.g. the Reuters corpus, the British National Corpus) are not large enough to be able to acquire reliable information in sufficient detail about language behaviour. Moreover, most languages do not have large or diverse enough corpora available.

Meaning proposes an innovative bootstrapping process to deal with the interdependency between WSD and knowledge acquisition:

1. Train accurate WSD systems and apply them to huge corpora by coupling knowledge-based techniques on EuroWordNet with ML techniques that combine huge amounts of labeled and unlabeled data. When ready, use also the knowledge acquired in 2.
2. Use the obtained accurate WSD data, shallow parsing techniques and domain tagging to extract new linguistic knowledge to incorporate to EuroWordNet.

This method will be able to break this interdependency in a series of cycles thanks to the fact that the WSD system will be based on all domain information, sophisticated linguistic knowledge, large numbers of automatically tagged examples from the web, and a combination of annotated and unannotated data. The first WSD system will have weaker linguistic knowledge, but the sole combination of the rest of the factors will produce significant performance gains. Besides, some of the required linguistic knowledge can be acquired from unnanotated data, and can therefore be acquired without using any WSD system. Once acceptable WSD is available, the acquired knowledge will be of a higher quality, and will allow for better WSD.

Multilingualism will be also helpful for Meaning. The idiosyncratic way the meaning is realised in a particular language will be captured and ported to the rest of languages involved in the project using EuroWordNet as a Multilingual Central Repository in three consecutive phases.

[1] http://www.lsi.upc.es/~nlp/meaning/meaning.html

4 Conclusions

Where the acquisition of knowledge from large-scale document collections will be one of the major challenge for the next generation of text processing applications, Meaning emphasises multilingual content-based access to web content. Moreover, it can provide a keystone enabling technologies for the semantic web. In particular, the Multilingual Central Repository produced by Meaning is going to constitute the natural knowledge resource for a number of semantic processes that need large amounts of linguistic data to be effective tools (e.g. web ontologies).

Current web access applications are based on words; Meaning will open the way for access to the multilingual web based on concepts, providing applications with capabilities that significantly exceed those currently available. Meaning will facilitate development of concept-based open domain Internet applications (such as Question/Answering, Cross Lingual Information Retrieval, Summarisation, Text Categorisation, Event Tracking, Information Extraction, Machine Translation, etc.). Furthermore, Meaning will supply a common conceptual structure to Internet documents, facilitating knowledge management of web content. This common conceptual structure is a decisive enabling technology for allowing the semantic web.

References

Rigau, G., B. Magnini, E. Agirre, P. Vossen, and J. Carroll. 2002. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of COLLING Workshop*, Taipei, Taiwan.