MEANING: a Roadmap to Knowledge Technologies

German Rigau. TALP Research Center. UPC. Barcelona. rigau@lsi.upc.es
Bernardo Magnini. ITC-IRST. Povo-Trento. magnini@itc.it
Eneko Agirre. IXA group. EHU. Donostia. eneko@si.ehu.es
Piek Vossen. Irion Technologies. Delft. Piek.Vossen@irion.nl
John Carroll. CCS. SUS. Brighton. johnca@cogs.susx.ac.uk

Abstract

Progress is being made in Natural Language Processing (NLP) but there is still a long way towards Natural Language Understanding (NLU). An important step towards this goal is the development of technologies and resources that deal with concepts rather than words. 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 large-scale enrichment of Lexical Knowledge Bases.

1. Introduction

The field of Information Society Technologies (IST) is one of the main thematic priorities of the European Commission for the 6th Framework programme. In that field, *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 for various languages, 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:

- In order to achieve accurate WSD, we need far more linguistic and semantic knowledge than is available in current lexical knowledge bases (e.g. current wordnets[1]).

[1] A wordnet is a conceptually structured knowledge base of word senses. The English WordNet (Miller 90, Fellbaum 98) has been developed at Princeton University over the past 14 years. EuroWordNet (Vossen 1998) is a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). Balkanet is building wordnets for the Balkan languages following EuroWordNet design.

- 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*.

2 Major research goals in NLP

In order to extend the state-of-the-art in human language technologies (HLT) future research must devise (1) innovative processes and tools for automatic acquisition of lexical knowledge from large-scale document collections; (2) novel techniques for accurately selecting the sense of open-class words in a large number of languages; (3) ways to enrich existing multilingual linguistic knowledge resources with new kinds of lexical information by automatically mapping information across languages.

2.1 Dealing with knowledge acquisition

The acquisition of linguistic knowledge from corpora has been a very successful line of research. Research in the acquisition of subcategorization information, selectional preferences, in thematic role assignments and diathesis alternations (Agirre and Martínez 2001, 2002, McCarthy and Korhonen, 1998; Korhonen et al., 2000; McCarthy 2001), domain information (Magnini and Cavaglià 2000), topic signatures (Agirre et al. 2001b), etc. has obtained remarkable results. The acquisition process usually involves large bodies of text which have been previously processed with shallow language processors.

Much of the use of the acquired knowledge has been hampered by the fact that the texts are not sense-disambiguated, and therefore, only knowledge for words can be acquired, that is, subcategorization for words, selectional preferences for words, etc. It is a well established fact that much of the linguistic behavior of words can be better explained if put into reference to word senses.

Having texts automatically sense-tagged with high accuracy will produce significantly better acquired knowledge at a sense level, including subcategorization frequencies, domain information, topic signatures, selectional preferences, specific lexico-semantic relations, thematic role assignments and diathesis alternations. Furthermore, it will allow to investigate automatic methods for dealing with new senses not present in current wordnets and clustering of word senses.

2.2 Dealing with WSD

Word Sense Disambiguation (WSD) is the task of assigning the appropriate meaning (sense) to a given word in a text or discourse. Ide and Veronis (1998) argue that word sense ambiguity is a central problem for many established HLT applications (for example Machine Translation, Information Extraction and Information Retrieval). This is also the case for associated sub-tasks (i.e. reference resolution and parsing). For this reason many international research groups are working on WSD, using a wide range of approaches. However, no large-scale broad-coverage accurate WSD system has been built up to date[2]. With current state-of-the-art accuracy in the range 60-70%, WSD is one of the most important open problems in NLP.

A promising current line of research uses semantically annotated corpora to train Machine Learning (ML) algorithms to decide which word sense to choose in which contexts. The words in these annotated corpora are tagged manually with semantic classes taken from a particular lexical semantic resource (most commonly

[2] See the conclusions of the SENSEVAL-2 competition: http://www.sle.sharp.co.uk/senseval2/

WordNet). Many standard ML techniques have been tried, such as Bayesian learning, Exemplar based learning, Decision Lists, and recently margin-based classifiers like Boosting and Support Vector Machines (Escudero et al., 2000a, 2000b, 2000c, 2000d, 2001; Martínez and Agirre, 2000). These approaches are termed "supervised" because they learn from previously sense annotated data and therefore they require a large amount of human intervention to annotate the training data.

Supervised WSD systems are data hungry. They suffer from the "knowledge acquisition bottleneck", it takes them mere seconds to digest all of the processed corpus contained in training materials that take months to annotate manually. So, although Machine Learning classifiers are undeniably effective, they are not feasible until we can obtain reliable unsupervised training data. Ng (1997) estimates that the manual annotation effort necessary to build a broad coverage word-sense annotated English corpus is about 16 person-years; and this effort would have to be replicated for each different language. Unfortunately, many people think that Ng’s estimate might fell short, as the annotated corpus thus produced is not guaranteed to enable high accuracy WSD.

Some recent work is focusing on reducing the acquisition cost and the need for supervision in corpus-based methods for WSD. Leacock et al. (1998) and Mihalcea and Moldovan (1999) automatically generate arbitrarily large corpora for unsupervised WSD training, using the synonyms or definitions of word senses provided in WordNet to formulate search engine queries over the Web. On another independent research area (Yarowsky, 1995) and (Blum and Mitchell, 1998) have shown that it is possible to reduce the need for supervision with the help of large amounts of unannotated data. Following this ideas, (Agirre and Martínez, 2000) has developed knowledge-based prototypes for obtaining accurate examples from the web for specific WordNet synsets, as well as, large quantities of unannotated examples.

But in order to make significant advances in WSD system accuracy, systems need to be able to use types of lexical knowledge that are not currently available in wide-coverage lexical knowledge bases: for example subcategorisation frequencies for predicates (particularly verbs) rely on word senses, selectional preferences of predicates for classes of arguments, amongst others (Carroll and McCarthy, 2000; McCarthy et al., 2001; Agirre and Martínez, 2002;).

2.3 Dealing with multilingualism

Language diversity is at the same time a valuable cultural heritage worth preserving, and an obstacle to achieving a more cohesive social and economic development. This situation has been further stressed as a major challenge in IST research lines. Improving language communication capabilities is a prerequisite for increasing industrial competitiveness, this way leading to a sound growth in key economic sectors.

However, this obstacle can be helpful because all languages realise the meaning in different ways. We can benefit from this fact using a novel multilingual mapping process that exploits the EuroWordNet architecture. In EuroWordNet local wordnets are linked via an Inter-Lingual-Index (ILI) allowing the connection from words in one language to translation equivalent words in any of the other languages. In that way, technological advances in one language can help the other. For instance, for Basque, being an agglutinative language with very rich morphological-syntactic information, it is possible to extract semantic relations that would be more difficult to capture in other languages. However, Basque is not largely present in the web as the others. Using this approach it is possible to balance both gaps.

Although the technology to provide compatibility across wordnets exits (Daudé et al, 1999, 2000, 2001), new research is needed for porting and uploading the various types of knowledge across languages, and new ways to test the validity of the ported knowledge in the target languages.

3. The MEANING Roadmap

The improvements mentioned above have been explored separately with relative success. In fact, no research group in isolation has tried to combine all this aforementioned factors. We designed the MEANING project[3] convinced that only a combination of all relevant knowledge and resources will be able to produce significant advances in this crucial research area.

MEANING 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 inter-dependency between WSD and knowledge acquisition:

1. Train accurate WSD systems and apply them to very large corpora by coupling knowledge-based techniques on the existing EuroWordNet (e.g. to populate it with domain labels, to induce automatically training examples) with ML techniques that combine very large amounts of labeled and unlabeled data. When ready, use also the knowledge acquired in 2.
2. Use the obtained accurate WSD data in conjunction with shallow parsing techniques and domain tagging to extract new linguistic knowledge to incorporate into 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 performance.

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[4] using EuroWordNet as a Multilingual Central Repository in three consecutive phases (see figure 1).

[3] Started in March 2002, MEANING IST-2001-34460 "Developing Multilingual Web-scale Language Technologies" is a three years research project funded by the EC.

[4] MEANING will work with three major European languages (English, Spanish and Italian) and two minority languages (Catalan and Basque).

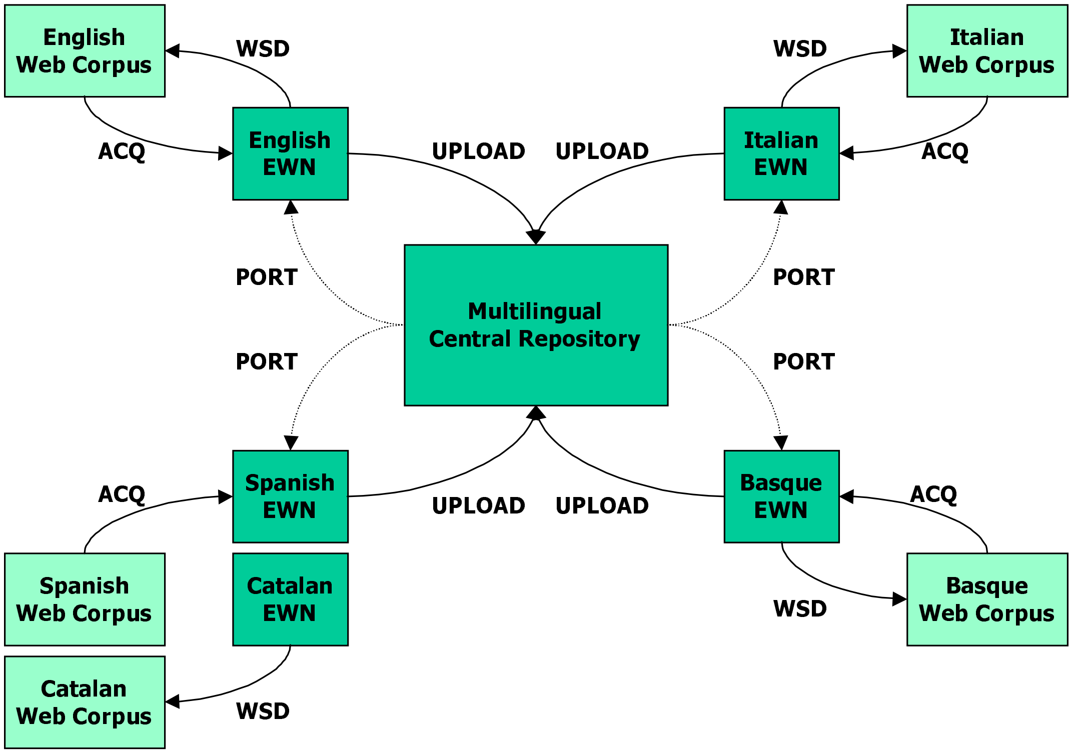

Figure 1: MEANING data flow.

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). NLP tools and software of the next generation will benefit from the MEANING outcomes.

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, thus facilitating knowledge management of web content. This common conceptual structure is a decisive enabling technology for allowing the semantic web.

References

Agirre E. and Martínez D. *Exploring automatic word sense disambiguation with decision lists and the Web*. Proceedings of the Workshop “Semantic Annotation And Intelligent Annotation” organized by COLING 2000. Luxembourg. 2000.

Agirre E. and Martinez D. *Learning class-to-class selectional preferences*. Proceedings of the Workshop "Computational Natural Language Learning" (CoNLL-2001). In conjunction with ACL'2001/EACL'2001. Toulouse. 2001.

Agirre E., Ansa O., Martínez D. and Hovy E. *Enriching WordNet concepts with topic* *signatures*. Proceedings of the NAACL workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations. Pittsburg. 2001.

Agirre E. and Martinez D. *Integrating selectional preferences in WordNet*. Proceedings of the first International WordNet Conference. Mysore, India, 2002.

Blum A. and Mitchel T. *Combining labelled and unlabeled data with co-training*. In Proceedings of the 11th Annual Conference on Computational Learning Theory. 1998.

Carroll, J. and McCarthy, D. *Word sense disambiguation using automatically acquired verbal preferences*. Computers and the Humanities. Senseval Special Issue, Vol. 34, No 1-2. 2000.

Daudé J., Padró L. and Rigau G., *Mapping Multilingual Hierarchies using Relaxation Labelling*, Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99). Maryland, 1999.

Daudé J., Padró L. and Rigau G., *Mapping WordNets Using Structural Information* , 38th Anual Meeting of the ACL. Hong Kong, 2000.

Daudé J., Padró L. and Rigau G., *A Complete WN1.5 to WN1.6 Mapping*, Proceedings of NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations". Pittsburg, PA, 2001.

Escudero G., Màrquez L. and Rigau G., *Boosting Applied to Word Sense Disambiguation*. Proceedings of the 11th European Conference on Machine Learning. Barcelona. 2000.

Escudero G., Màrquez L. and Rigau G., *Naive Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited*. Proceedings of the 14th European Conference on Artificial Intelligence, Berlin. 2000.

Escudero G., Màrquez L. and Rigau G., A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation. Proceedings of Fourth Computational Natural Language Learning Workshop. Lisbon. 2000.

Escudero G., Màrquez L. and Rigau G., *An Empirical Study of the Domain Dependence of Supervised Word Sense Disambiguation Systems*. Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong. 2000.

Escudero G., Màrquez L. and Rigau G., *Using LazyBoosting for Word Sense Disambiguation*. Proceedings of 2nd International Workshop "Evaluating Word Sense Disambiguation Systems", SENSEVAL-2. Toulouse. 2001.

Fellbaum C. editor. *WordNet An Electronic Lexical Database*. The MIT Press. 1998.

Ide, N. and Vèronis, J. *Introduction to the special issue on word sense disambiguation: The state of the art*. Computational Linguistics, 24 (1), 1998.

Korhonen A., Gorrell, G. and McCarthy D. *Statistical Filtering and Subcategorization Frame Acquisition*. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong. 2000.

Leacock, C. Chodorow, M. and Miller, G.A. *Using Corpus Statistics and WordNet Relations for Sense Identication*, Computational Linguistics, 24(1), 1998.

Magnini B. and Cavaglià G., *Integrating subject field codes into WordNet*. In Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens. 2000.

Martínez D. and Agirre E. *One Sense per Collocation and Genre/Topic Variations*. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong, 2000.

McCarthy, D. and Korhonen, A. *Detecting verbal participation in diathesis alternations*. Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics COLING-ACL'98. Montreal. 1998.

McCarthy D., *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex. 2001.

McCarthy, D., Carroll J. and J. Preiss. J. *Disambiguating noun and verb senses using automatically acquired selectional preferences*. Proceedings of the SENSEVAL-2 Workshop at ACL/EACL'01, Toulouse. 2001.

Mihalcea R. and Moldovan D. *An automatic method for generating sense tagged corpora*. In Proceedings of American Association for Artificial Intelligence. 1999.

Miller G. *Five papers on WordNet,* Special Issue of International Journal of Lexicogrphy 3(4). 1990.

Ng. H. T. *Getting Serious about Word Sense Disambiguation*. In Proceedings of Workshop "Tagging Text with Lexical Semantics: Why, what and how?", Washington, 1997.

Vossen P. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht. 1998.

Yarowsky D., *Unsupervised word sense disambiguation rivaling supervised methods*. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. 1995.