MEANING Cycle 2: Assessment

Document Number	D7.2
Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Project URL	http://www.lsi.upc.es/~nlp/meaning/meaning.html
Availability	Public
Authors: John Carroll (Sussex), Bernardo Magnini (ITC-irst), German Rigau, Eneko Agirre (UPV/EHU), Lluis Padro (UPC)	

meaning

INFORMATION SOCIETY TECHNOLOGIES

Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Security (Distribution level)	Public
Contractual date of delivery	February 29 2004
Actual date of delivery	March 30 2004
Document Number	D7.2
Type	Report
Status & version	v Final
Number of pages	13
WP contributing to the deliberable	Work Package 7
WPTask responsible	John Carroll (Sussex)
Authors	John Carroll (Sussex), Bernardo Magnini (ITC-irst), German Rigau, Eneko Agirre (UPV/EHU), Lluis Padro (UPC)
Other contributors	
Reviewer	
EC Project Officer	Evangelia Markidou
Authors: John Carroll (Sussex), Bernardo Magnini (ITC-irst), German Rigau, Eneko Agirre (UPV/EHU), Lluis Padro (UPC)	
Keywords: NLP, Lexical Knowledge Representation, Acquisition, WSD	
Abstract: In this deliverable we make a qualitative assessment of the progress made in the second cycle of the MEANING project.	

Contents

1 Introduction

The main objectives of work package 7 are:

- to evaluate the quality and accuracy of the developed software and the acquired data; and
- to assess the progress of the project, and if necessary, provide the necessary information to devise corrective actions.

Evaluations of quality and accuracy are given in the relevant work packages (for example Deliverable 5.2 presents the work on acquisition carried out in the second MEANING cycle, and presents quantitative evaluation results).

In this document we give a *qualitative* assessment of progress made in the second MEANING cycle, and make proposals fo work that could be carried out in the next cycle.

2 Linguistic Processors

Deliverable D3.2 presents an inventory of tools, lexical resources, and corpora which have been developed or enhanced within the MEANING project. These tools and resources have benn used to suport the work carried out in WP5 (Acquisition) and WP6 (Word Sense Disambiguation).

3 Knowledge Integration

Deliverable D4.2 describes the MCR1, the second version of the MEANING Multilingual Central Repository, and the second Porting process (PORT1). In this section we summarise the information uploaded and integrated into MCR1.

The final content of MCR1 includes:

- ILI
  - Aligned to WordNet 1.6 [Fellbaum, 1998]
  - EuroWordNet Base Concepts [Vossen, 1998]
  - EuroWordNet Top Concept Ontology [Vossen, 1998]
  - WordNet Domains version 070501 [Magnini and Cavagli, 2000]
  - Suggested Upper Merged Ontology (SUMO) [Niles and Pease, 2001]
- Local wordnets
  - English WordNet 1.5, 1.6, 1.7, 1.7.1, 2.0 [Fellbaum, 1998]
  - eXtended WordNet 1.7 [Mihalcea and Moldovan, 2001]
  - Basque wordnet [Agirre *et al.*, 2002]
  - Italian wordnet [Pianta *et al.*, 2002]
  - Catalan wordnet [Benítez *et al.*, 1998]
  - Spanish wordnet [Atserias *et al.*, 1997]
- Large collections of semantic preferences
  - Direct dependencies from Parsed SemCor [Agirre and Martinez, 2001]
  - Acquired from SemCor [Agirre and Martinez, 2001; Agirre and Martinez, 2002]
  - Acquired from BNC (2nd release) [McCarthy, 2001]
- Large collections of Sense Examples
  - SemCor
- Instances
  - Named Entities [Alfonseca and Manandhar, 2002]
  - Named Entities [Niles and Pease, 2001]
  - Named Entities [Pianta *et al.*, 2002]

source	#relations
Acquired from Princeton WN1.6	138.091
Selectional Preferences acquired from SemCor	203.546
Selectional Preferences acquired from BNC	707.618
New relations acquired from Princeton WN2.0	42.212
Gold relations from eXtended WN	17.185
Silver relations from eXtended WN	239.249
Normal relations from eXtended WN	294.488
Total	1.642.389

Table 1: Main sources of semantic relations

The current version of the MCR integrates wordnets from five different languages. The final version of the MCR contains 1,642,389 unique semantic relations between concepts (ILI-records). This represents one order of magnitude larger than the Princeton wordnet (138,091 unique semantic relations in WN1.6). Table 1 summarizes the volumes of the main sources for semantic relations integrated into MCR1.

Furthermore, the current MCR have been also enriched with 466,972 semantic properties coming from different sources. Table 2 summarizes the main sources for semantic properties integrated into MCR1.

source	#properties
WordNet Domain	110.556
Top Concept Ontology	256.776
SUMO	99.640
Total	466.972

Table 2: Main sources of semantic properties

The resulting MCR1 is probably the largest and richest multilingual lexical–knowledge ever built. For MCR2 (the last version of the MCR) we are planning to provide also an Ontology for Named Entities, and an automatically constructed set of Base Concepts. The consortium will also ask for a new set of Base Concepts developed for WN1.7 for Balkanet and those developed at Princeton.

For the MCR2, we are also planning to upload VerbNet. VerbNet[1] is a verb lexicon with syntactic and semantic information for English verbs, using Levin verb classes to systematically construct lexical entries. For each syntactic frame in a verb class, there is a set of semantic predicates associated with it. Many of these semantic components are cross-linguistic. The lexical items in each language form natural groupings based on the presence or absence of semantic components and the ability to occur or not occur within

[1] `http://www.cis.upenn.edu/group/verbnet`

particular syntactic frames. The English entries are mapped directly onto English WordNet senses. We hope that this new resource will provide further structure and consistency to the selectional preferences acquired automatically.

From ACQ1 we obtained a large set of sense examples acquired automatically from the web (see Working Paper WP5.5 *Experiment 5.H a): Publicly available topic signatures for all WordNet nominal senses*). These examples have been obtained querying Google. For each word sense in WordNet, a program builds a complex query including sets of monosemous synonymous relatives. Using this approach, large collections of text can be obtained. This will represent hundreds of examples per word sense. Using this large-scale resource we plan to generate topic signatures for every word sense in WordNet that will be uploaded into MCR2.

The consortium also plans to upload a new parsed version of SemCor. The corpus will be parsed with a new version of RASP able to process XML files.

We will also study how to port the last version of the MCR to other formats including DAML+OIL, RDF, etc. that will be of utility of the Semantic Web communities.

4 Acquisition

In the second cycle of MEANING, work on acquisition focused on leveraging new types of lexical information acquired from large text collections and the web, for example distributional similarity data, topical vectors, and information mapped from one language to another.

We carried out a number of acquisition experiments in this cycle, as summarised below. (For more details, see D5.2 and working papers WP5.5–11 which describe fully the experiments and results).

- Experiment A consisted of two pilot studies into how lexical knowledge obtained in one language can help in the acquisition process performed in other languages, specifically semantic patterns for predicates, and attachment preferences for prepositional phrases. We obtained promising results, considering the high degree of polysemy involved.

- Experiment D investigated two different techniques for acquisition of *topic signatures* from corpora and the web. (Topic signatures associate a topical vector to each word sense.) The acquired topic signatures were successfully used to compute the similarity between nominal word senses, to cluster word senses, and for WSD (performing a little above the 'supervised' baseline, and beating random choice by a large margin).

- Experiment E was also concerned with unsupervised bootstrapping large amounts of data for supervised WSD training. It focused on acquiring examples of words used in particular senses. For this purpose a new software tool was developed. EXRETRIEVER characterises automatically each synset of a word as a query (using mainly synonyms, hyponyms and the words from the definitions in a WordNet), and then uses these queries to obtain sense examples (sentences) automatically from large text collections. Various novel acquisition strategies were defined, executed and evaluated.

- Experiment G evaluated a new method of acquiring selectional preferences from unannotated text, which instead of covering all the noun senses in WordNet, just gives a probability distribution over a portion of "prototypical classes" rather than all senses. This method improves in both precision and recall over previous similar methods. Using automatically disambiguated input data further increased recall.

- Experiment H explored a set of methods to hierarchically cluster the (often very fine-grained) word senses in WordNet. The best results came from using word usage examples obtained automatically from the web, based on topic signature data produced in Experiment D.

- Experiment I investigated a method for unsupervised acquisition of English phrasal verbs (as distinct from verbs that take prepositional complements). The method used distributional similarity data, and demonstrated that various measures using

the distributional nearest neighbours of a phrasal verb and its simplex counterpart show a highly significant correlation with human compositionality judgements.

- Experiment K developed a method for automatically determing the most prevalent WordNet sense for a word in a given corpus. Many words have skewed frequency distributions for their senses, and supervised WSD systems usually back off to the more prevalent sense if they do not have enough information.

In the next round, we will further scale up these acquisition techniques in order to produce data for the final MEANING WSD systems.

5 Word Sense Disambiguation

The WSD working package in this second cycle has focused in the following:

- The development of state-of-the-art WSD systems for all languages: we have developed lexical-sample systems for all languages involved in MEANING, plus all-words systems for English and Italian.
- The use of automatically acquired examples. A large number of examples for each nominal word sense have been acquired from the web, and the first experiments have been carried out, showing that a difference can be made in the all-words English system.
- The use of the information acquired and ported in ACQ0. The WORDNETs ported to all MEANING languages have been used to develop all-words systems for languages other than English. Experiment E has shown concrete results for Italian.
- A number of experiments, new and old, have been active, trying to make progress both in the performance of lexical-sample systems, and in seeking ways to break the acquisition bottleneck.
- Participation in Senseval-3. The MEANING consortium is organizing lexical-sample tasks for all languages involved in the project. The consortium is participating in the lexical-sample tasks for all MEANING languages, plus all-words tasks for English and Italian, and the dictionary disambiguation task.

Below, we summarize the main conclusions for each experiment (for details check D6.2, or the working papers corresponding to each experiment):

- Experiment A: An English all-words system competitive with those presented at Senseval-2 has been entered for Senseval-3. This prototype includes a large number of supervised and unsupervised systems developed in MEANING.
- Experiment E: An all-words system (Domain Driven Disambiguation) has been evaluated for Italian. Other MEANING languages (Basque, Catalan, Spanish) do not have an all-words gold standard, so they have not been evaluated. Nevertheless, the Italian results are a good indication, given that the lexical resources and linguistic processing for these languages is on a par with Italian.
- Experiment F has shown that smoothing allows for improved results in WSD. The experiments are still ongoing.
- Experiment G has shown some improvement in the use of automatically acquired selectional preferences, compared to the first round. This experiment has used results from WSD0 to improve the acquisition (ACQ1) and then apply the acquired knowledge in this phase (WSD1).

- Experiments H have shown important advances in the bootstrapping process. Experiment H a) has used automatically acquired examples from the web in order to improve the English all-words system with positive results. Experiment H b) has shown that topic signatures acquired from a second language provide a competitive unsupervised WSD system. Experiment H c) showed that using untagged data we can improve supervised WSD results. Finally Experiment H d) tried (unsuccessfully) to use lexical relations to improve the English all-words systems.

- Experiment K has shown that it is possible to induce the priors (ranks) for the word senses of a word based solely on raw text. The results of the experiment are very encouraging.

- Experiment M has shown that the combination of knowledge- and corpus-based techniques help improve the results of each other.

For the next round, all experiments will try to produce further insight in the improvement of WSD systems. Our Senseval-3 participation will be also evaluated, and the final MEANING WSD prototypes will be released.

References

[Agirre and Martinez, 2001] E. Agirre and D. Martinez. Learning class-to-class selectional preferences. In *Proceedings of CoNLL01*, Toulouse, France, 2001.

[Agirre and Martinez, 2002] E. Agirre and D. Martinez. Integrating selectional preferences in wordnet. In *Proceedings of the first International WordNet Conference in Mysore*, India, 21-25 January 2002.

[Agirre *et al.*, 2002] E. Agirre, O. Ansa, X. Arregi, J.M. Arriola, A. Diaz de Ilarraza, E. Pociello, and L. Uria. Methodological issues in the building of the basque wordnet: quantitative and qualitative analysis. In *Proceedings of the first International WordNet Conference in Mysore*, India, 21-25 January 2002.

[Alfonseca and Manandhar, 2002] E. Alfonseca and S. Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the 1st International Conference on General WordNet*, Mysore, India, 2002.

[Atserias *et al.*, 1997] J. Atserias, S. Climent, X. Farreres, G. Rigau, and H. Rodríguez. Combining multiple methods for the automatic construction of multilingual wordnets. In *Proceedings of RANLP'97*, pages 143–149, Bulgaria, 1997.

[Benítez *et al.*, 1998] L. Benítez, S. Cervell, G. Escudero, M. López, G. Rigau, and M. Taulé. Methods and tools for building the catalan wordnet. In *Proceedings of the ELRA Workshop on Language Resources for European Minority Languages, First International Conference on Language Resources & Evaluation*, Granada, Spain, 1998.

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

[Magnini and Cavagli, 2000] B. Magnini and G. Cavagli. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*, Athens. Greece, 2000.

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

[Mihalcea and Moldovan, 2001] R. Mihalcea and D. Moldovan. Extended wordnet: Progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA, 2001.

[Niles and Pease, 2001] I. Niles and A. Pease. Towards a standard upper ontology. In *In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 17–19. Chris Welty and Barry Smith, eds, 2001.

[Pianta *et al.*, 2002] E. Pianta, L. Bentivogli, and C. Girardi. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India, 2002.

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