

Building High Quality Topic Signatures

Montse Cuadros Oller

cuadros@lsi.upc.edu

Directors

Lluís Padró Cirera

German Rigau Claramunt

Memòria del DEA i Projecte de Tesi

Programa de Doctorat en Intel·ligència Artificial

Departament de Llenguatges i Sistemes Informàtics

Universitat Politècnica de Catalunya

Maig 2006

Contents

1	Introduction	5
2	State of the Art	9
2.1	Introduction	9
2.2	WordNet	10
2.3	EuroWordNet	12
2.4	Enriching Existing WordNets	13
2.4.1	eXtended WordNet	14
2.4.2	WordNet Domains	14
2.5	Multilingual Central Repository	15
2.6	Large-Scale Knowledge Acquisition	19
2.6.1	Topic Signatures	20
2.7	Evaluation Frameworks	21
3	Our approach	23
3.1	Main Goal	23
3.1.1	Illustrative Example	24
4	Evaluating Large-scale Knowledge Resources	29
4.1	Introduction	29
4.2	Senseval Evaluation Framework	30
4.3	Knowledge Resources	31
4.4	Indirect Evaluation on Word Sense Disambiguation	31

4.5	Evaluating the quality of knowledge resources	32
4.5.1	Baselines	33
4.5.2	Performance of the knowledge resources	34
4.5.3	Senseval-3 system performances	34
4.6	Conclusions and future work	36
5	Large-scale Knowledge Acquisition	39
5.1	Retrieving automatically Topic Signatures	39
5.1.1	Examples	41
5.2	Experiments	43
5.3	Combination of Knowledge Resources	45
5.4	Conclusions and Future Work	47
6	A proposal for disambiguating large-scale knowledge resources	49
6.1	Aim	49
6.2	Disambiguating process	49
6.3	Preliminaries Results	50
7	Thesis project	55
7.1	Research roadmap	55
7.2	Working Plan	56
7.3	Related Publications	56

Chapter 1

Introduction

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) that attempts to automatically process the human language. Early systems, working in restricted “closed worlds” with restricted vocabularies, worked extremely well, leading researchers to excessive optimism which was soon lost when the systems were extended to more realistic situations with real-world ambiguity and complexity.

Thus, the need for large and sophisticated semantic lexicons is recognized as one of the major problems in NLP applications both because of the need for substantial vocabulary in current NLP systems and because of its increasing complexity. Furthermore, the need of large-scale semantic resources is becoming more pressing since NLP systems are making the transition from laboratories to industry. However, the current requirements of NLP applications are widely exceeding the capabilities of the existing large-scale semantic resources.

The task of constructing realistic semantic resources for NLP deals with enormous amounts of knowledge, and a large set of languages to cover. There are many concepts, words and many distinct types of information about them potentially relevant to different kinds of NLP tasks and applications.

Using large scale lexico-semantic knowledge bases (such as WordNet, ONTOS, Mikrokosmos, Cyc, etc.) has become a usual, often necessary, practice for most current NLP systems. Building large-scale resources of this nature for open domain semantic processing is a hard and expensive task, involving large research groups during long periods of

development. For example, dozens of person-years are been invested world-wide into the development of wordnets for various languages (Fellbaum, 1998a), (Vossen, 1998b).

For instance, in more than eight years of manual construction (from version 1.5 to 2.1), WordNet passed from 103,445 semantic relations to 204,074 semantic relations¹. That is, only around twelve thousand semantic relations per year.

However, the diffusion and success of WordNet have also determined the emergence of several projects that aim either to build wordnets for languages other than English² (Hamp & Feldweg, 1997; Artale et al., 1997), or to develop multilingual wordnets, such as EuroWordNet (Vossen, 1998a) or Balkanet (Stamou et al., 2002b), to develop wordnets for particular domains, such as EuroTerm (Stamou et al., 2002a) or MuchMore (Buitelaar & Sacaleanu, 2002), or to automatically enrich existing wordnets, as for instance eXtended WordNet (Mihalcea & Moldovan, 2001), Omega (A.Philpot et al., 2005) or MCR (Rigau et al., 2002).

Unfortunately, the outcomes of these projects have been, usually, large and complex semantic structures, hardly compatible with resources developed in other projects and efforts. Obviously, this fact has severely hampered NLP progress. Thus, one of the main issues in the last years concerning NLP activities has been focused on the fast development, tuning and reuse of general semantic resources.

Obviously, the unique way to fast built Knowledge Bases, is using automatic methods. Obviously, manual methods seem to be more reliable but also time-consuming and high-costly.

Automatic lexical acquisition is also an old open issue in NLP. A large battery of methods have been used to obtain implicit information from structured and unstructured lexical resources. Obtaining large, explicit lexicons rich enough for NLP has proved difficult. Methods for automatic lexical acquisition have been developed for many topics and include several techniques in order to obtain the desired data.

The acquisition of knowledge from large-scale document collections is one of the major challenges for the next generation of text processing applications.

¹Symmetric relations are counted only once

²Find a list of wordnets currently under development at <http://www.globalwordnet.org>

Obviously, using annotated text is most accurate, but also in this case it is very expensive and its coverage is usually small. In fact, many types of useful lexical relationships have never been annotated. In the 5th Framework MEANING project (Rigau et al., 2002), EHU, UPC, ITC-irst and UoS carried out extensive experiments into acquiring several different types of lexical/conceptual information purely automatically, from large text collections or from the internet. This is probably the largest focused effort of lexical acquisition ever taken. As a result, the project provided with two of the largest semantic resources available: the Multilingual Central Repository (MCR) (Atserias et al., 2004) with more than 1,5 million of semantic relations and the Topic Signatures (TS) acquired from the web (Agirre & de la Calle, 2004).

In fact, the main goal of our research is to increase the existing Knowledge Resources with knowledge acquired automatically using large sets of word-weight pairs called Topic Signatures which are related to a specific topic (word sense). Once acquired, the TS associated to a particular word sense will be integrated into the MCR by assigning a sense for each word of the TS.

For instance, the noun car has 5 senses, and the first sense has 22 relations in WordNet1.6. However, the MCR partially derived by automatic means has 305. But the Topic Signature associated to this sense of car has more than one thousand words. Our plan is to accurately integrate all this new knowledge into the existing semantic resources.

Structure of the Document

The rest of this document is organized as follows. Chapter 2 present the state-of-the-art regarding large-scale knowledge acquisition for Natural Language applications.

Chapter 3 presents our approach which is further detailed in:

- Chapter 4: Evaluating Large-scale Knowledge Resources.
- Chapter 5: Large-scale Knowledge Acquisition.
- Chapter 6: Enrichment of Knowledge Resources using a Word Sense Disambiguation method.

Finally, chapter 7 and section 7.3 state the thesis project and presents the related publications.

Chapter 2

State of the Art

2.1 Introduction

The task of constructing realistic semantic resources for NLP deals with enormous amounts of knowledge and large set of languages to cover. There are many concepts, words and many distinct types of information potentially relevant to different kinds of NLP tasks. Using large scale lexico–semantic knowledge bases (such as WordNet (WN), ONTOS, Mikrocosmos, Cyc, etc.) has become a usual, often necessary, practice for most current NLP systems.

Semantic Resources are often used to develop Natural Language Processing systems and also they are applied with semantic distance measures, WSD, lexical chains detections, etc. In application level, these resources are been used for tasks such as Information Retrieval, Conceptual text Mining, Automatic Summarization, Question Answering, Information Integration, etc.

However, the diffusion and success of WordNet¹ have also determined the emergence of several projects that aim either to build wordnets for languages other than English² (Hamp & Feldweg, 1997; Artale et al., 1997), or to develop multilingual wordnets, such as EuroWordNet (Vossen, 1998b) or Balkanet (Stamou et al., 2002b), to develop wordnets for particular domains, such as EuroTerm (Stamou et al., 2002a) or MuchMore (Buite-

¹<http://mira.csci.unt.edu/~wordnet>

²Find a list of wordnets currently under development at <http://www.globalwordnet.org>

laar & Sacaleanu, 2002), or to automatically enrich existing wordnets, as for instance extended WordNet (XWN) (Mihalcea & Moldovan, 2001) and Multilingual Central Repository (Rigau et al., 2002).

Unfortunately, the outcomes of these projects have been, usually, large and complex semantic structures, hardly compatible with resources developed in other projects and efforts. Obviously, this fact has severely hampered NLP progress. Thus, one of the main issues in the last years concerning NLP activities has been focused on the fast development, tuning and reuse of general semantic resources.

2.2 WordNet

The Princeton WordNet (Miller et al., 1991),(Fellbaum, 1998b) is a lexical database which contains information about nouns, verbs, adjectives and adverbs in English and is organized around the notion of a *synset*. For example, $\langle party, political_party \rangle$ form a synset because they can be used to refer to the same concept. A synset is often further described by a gloss: "an organization to gain political power".

A synset is a set of words with the same part-of-speech (PoS) that can be interchanged in a certain context and they are related to each other by semantic relations.

The actual version of WordNet is 2.1. Table 2.1 show the different versions of WordNet and their statistics taking account the number of synsets. Table 2.2 show the number of words, synsets and word-senses pairs of WordNet version 2.1.

In WordNet each word could have one or more senses. A word would be monosemous if it has only one sense while it would be polysemous in case of having several senses. Table 2.3 presents some figures of the monosemous and polysemous words in numbers.

WordNet encodes 26 different types of semantic relations such as Hypernym, Hyponym, Antonym, Meronym, Holonym, etc. between synsets of the same part-of-speech.

PoS	WordNet					
	1.5	1.6	1.7	1.7.1	2.0	2.1
Noun	60557	66025	74488	75804	79689	81426
Verb	11363	12127	12754	13214	13508	13650
Adjective	13231	17915	18523	21460	18563	22141
Adverb	3243	3575	3612	3629	3664	3644
Total <i>synset</i>	91591	99642	109377	111223	115424	117597

Table 2.1: Number of *synsets* in the different versions of WordNet

PoS	Unique Strings	<i>Synsets</i>	Total Word-Sense Pairs
Noun	11709	81426	145104
Verb	11488	13650	24890
Adjective	22141	18877	31302
Adverb	4601	3644	5720
Total	155327	117597	207016

Table 2.2: Number of words, *synsets* and senses in WN2.1

PoS per <i>synset</i>	Monosemous Words and Senses	Polysemous Words	Polysemous Senses
Noun	101321	15776	43783
Verb	6261	5227	18629
Adjective	16889	5252	14413
Adverb	3850	751	1870
Totals	128321	27006	78695

Table 2.3: Polysemy in WN2.1

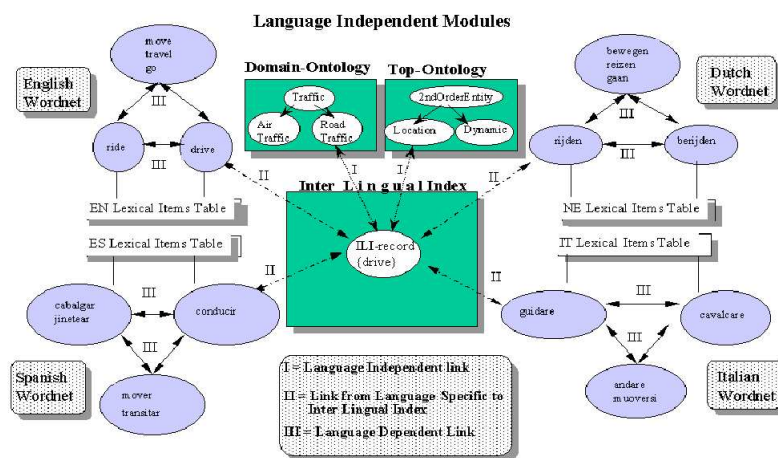


Figure 2.1: EuroWordNet architecture

2.3 EuroWordNet

EuroWordNet provided a natural laboratory to experiment different approaches for building large-scale semantic networks.

EuroWordNet, an interesting example of semi-automatic building process is the EuroWordNet(EWN) project (Vossen, 1998b), has been one of the largest effort to build multilingual semantic resources, coordinating the parallel development of independent wordnets for several European languages.

Figure 2.1 shows an schema of EuroWordNet architecture.

The resulting wordnets are structured in the same way as the WordNet from Princeton in terms of synsets with basic semantic relations between them. Each wordnet represents a unique language-internal system of lexicalizations and was developed independently. In addition, the individual wordnets are linked to an Inter-Lingual-Index, based on WordNet. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. The index also gives access to a shared Top Concept Ontology of 63 semantic units and a Domain Ontology.

This Top Concept Ontology provides a common semantic framework for all the languages, while language specific properties are maintained in the local wordnets.

2.4 Enriching Existing WordNets

There is a lot of work focused to automatically or manually enrich WordNets in order to create new resources or improve the existing ones.

For instance, Alfonseca and Manandhar (Alfonseca & Manandhar, 2002b) propose a simple unsupervised approach for placing in the right place a Named Entity not present in WordNet. They assume that the synsets of WordNet have been enriched with Topic Signatures (Agirre et al., 2001a).

Moldovan and Girju (Moldovan & Girju, 2000) start with a set of manually selected seed concepts for acquiring new concepts and relations between these concepts and other WordNet synsets. (Hearst & Schutze, 1993), propose a simple algorithm to transform the hierarchic structure of WordNet with a set of plain categories. (Montoyo et al., 2001b), propose a system to reduce the polysemy of WordNet selecting similar Word-senses in content, and eliminating rare-used synsets. Tuning wordnet to a new domain can involve several activities: enriching already available wordnets with new concepts (usually consisting of terminological information) and relationships, integrating generic and domain-specific ontologies and ranking and selecting nodes by relevance, including dropping those nodes irrelevant in the new domain.

Turcato (Turcato et al., 1998) adapts WordNet to a specific domain by removing those synonymy relations that do not hold in the domain.

Hearst and Schutze (Hearst & Schutze, 1993) propose a very simple algorithm for rearranging WordNet converting its hierarchical structure into a reduced set of flat categories. The main goal is to create categories based on clusters.

Mihalcea and Moldovan (Montoyo et al., 2001a) propose a more conservative system for dealing with the same problem. Tomuro (Tomuro, 1998) tries to relate WordNet synsets by the semi-automatic induction of underspecified semantic classes,

Vossen (Vossen, 2001) presents a system for customising a multilingual WordNet database for technical document collections in specific domains.

Buitelaar and Sacaleanu (Buitelaar & Sacaleanu, 2001) present a system for determining the domain-specific relevance of GermaNet synsets.

2.4.1 eXtended WordNet

In the eXtended WordNet project³ (Mihalcea & Moldovan, 2001) WordNet glosses are syntactically parsed, transformed into logic forms and the content words are semantically disambiguated. The key idea of the Extended WordNet project is to exploit the rich information contained in the definitional glosses that is now used primarily by humans to identify correctly the meaning of words. In the first version of the eXtended WordNet released, XWN 0.1, the glosses of WordNet 1.7 are parsed, transformed into the logic forms and the senses of the words are disambiguated. There is now extended WordNet for WordNet 2.0 version.

The Extended WordNet may be used as a Core Knowledge Base for applications such as Question Answering, Information Retrieval, Information Extraction, Summarization, Natural Language Generation, Inferences, and other knowledge intensive applications. The glosses contain a part of the world knowledge since they define the most common concepts of the English language.

2.4.2 WordNet Domains

WordNet Domains is a lexical resource developed by Magnini and Cavaglià (Magnini & Cavaglià, 2000), where synsets are semi-automatically annotated with one or several domain labels from a 165 label-set organize hierarchically derived from "Dewey Decimal Classification"⁴. The original version was performed over WN1.6. Actually and due to the technology appropriate to have the mappings from one WordNet version to other WordNets (Rigau, 2000; Daudé et al., 2003), there is available this resource for all the WordNet versions. (Castillo et al., 2003), automatically assigns WordNet Domain labels to wordNet synsets.

Table 2.4 shows the distribution of number of labels per *synset* in WN2.0. The vast majority of *synset* have only one domain label.

The domain labels are complemented with the information contained in WordNet. One domain can include *synsets* of different syntactic categories: for example MEDICINE can

³<http://xwn.hlt.utdallas.edu>

⁴<http://www.oclc.org/dewey>

No. labels	Noun	Verb	Adj	Adv	Total	Percentage
1	66827	12506	17343	3548	100224	86,83%
2	10461	893	1097	109	12560	10,88%
3	1849	97	114	7	2067	1,79%
4	477	11	8	0	496	0,43%
5	75	1	0	0	76	0,07%
6	0	0	1	0	1	0,00%
Total	79689	13508	18563	3664	115424	100,00%

Table 2.4: Label Domain Distribution number per *synset* in WN2.0

content senses of nouns like *doctor#n#1* and *hospital#n#1*, and from verbs such as *operate#v#7* and a domain can include senses from different hierarchies of WN: for example SPORT contains different senses derived from *lifeform#n#1*, *physicalobject#n#1*, *act#n#2*, *location#n#1*.

2.5 Multilingual Central Repository

The Multilingual Central Repository (MCR) has been developed by the MEANING(IST-2001-34460)⁵ project. It is the result of the fusion of many different resources (different wordnet versions, Ontologies, lexicons) using the de facto standard sense repository, WordNet.

MCR constitutes a natural multilingual large-scale linguistic resource for a number of semantic processes that need large amounts of linguistic knowledge to be effective tools (e.g. semantic web ontologies). The fact that word senses are linked to concepts in MCR allow the appropriate representation and storage of the acquired knowledge.

The MCR integrates into the same EuroWordNet framework, an upgraded release of Base Concepts and Top Ontology, the MultiWordNet Domains, the Suggested Upper Merged Ontology, five local wordnets (with four English WordNet versions) and hundreds of thousand of new semantic relations, instances and properties fully expanded. In fact, the resulting MCR is the largest and richest multilingual lexical knowledge base ever build.

This multilingual structure allows to port the knowledge from one WordNet to the

⁵<http://www.lsi.upc.es/~ilp/meaning>

rest of languages via the EuroWordNet Ili, maintaining the compatibility among them. In that way, the Ili structure (including the Top Concept ontology and the Domain ontology) acts as a natural backbone to transfer the different knowledge acquired from each local wordnet to the rest of wordnets, balancing resources and technological advances across languages. In the same way, all the different resources (e.g. different ontologies) could be related through the ILI, and thus cross-checked.

MCR includes only conceptual knowledge. This means that only semantic relations between synsets will be acquired, uploaded and ported across local wordnets. However, when necessary, the relations acquired can be underspecified. In that way, they will be uploaded and ported to be ready useful for other acquisition processes and languages. For instance, consider the following relation <gain> involved <money> captured as typical object. Although, this relation may be further specified into <gain> involved-patient <money> posteriorly, other processes can take profit from a ported relation <ganar> involved <dinero>.

Currently, the MCR integrates:

- Ili aligned to WordNet 1.6 (Fellbaum, 1998b)
 - EuroWordNet Base Concepts (Vossen, 1998b)
 - BalKaNet Base Concepts (Stamou et al., 2002b)
 - EuroWordNet Top Concept Ontology (Vossen et al., 1997)
 - MultiWordNet Domains (Magnini & Cavaglià, 2000)
 - Suggested Upper Merged Ontology (sumo) (Niles & Pease, 2001)
- Local wordnets
 - Princeton English WordNet 1.5, 1.6, 1.7, 1.7.1, 2.0 (Fellbaum, 1998b)
 - eXtended WordNet (XWN) (Mihalcea & Moldovan, 2001)
 - Basque wordnet (Agirre et al., 2002)
 - Italian wordnet (Bentivogli et al., 2002)
 - Catalan wordnet (Benítez et al.,)

- Spanish wordnet (Atserias et al., 1997)
- Large collections of semantic preferences
 - Direct dependencies from Parsed SemCor (Agirre & Martinez, 2001)
 - Acquired from SemCor (Agirre & Martinez, 2001; Agirre & Martinez, 2002)
 - Acquired from BNC (McCarthy, 2001)
 - Large collections of Sense Examples
 - SemCor (Miller et al., 1993)
- Instances
 - Named Entities from IRST (Bentivogli et al., 2002)
 - Instances from sumo (Niles & Pease, 2001)
 - Named Entities from Alfonseca (Alfonseca & Manandhar, 2002a)
 - Verb Lexicon
 - VerbNet (Kipper et al., 2000)

A full description of the MCR and its contents can be found in (Atserias et al., 2004).

However, although these resources have been derived using different WordNet versions (from WN1.5 to WN2.0), the research community have the technology for the automatic alignment of wordnets (Daudé, 2004). This technology provides a mapping among synsets of different WordNet versions, maintaining the compatibility to all the knowledge resources which use a particular WordNet version as a sense repository. Furthermore, this technology allows to port the knowledge associated to a particular WordNet version to the rest of WordNet versions already connected.

In Figure 2.2, we present the main scheme of theMCR, with the directions which the relations are uploaded and ported.

Table 2.5 shows the number of relations of each source integrated in MCR.

Table 2.6 shows the number of synset-pairs with overlapping semantic relations. Most of the relations in the MCR between synsets-pairs are unique.

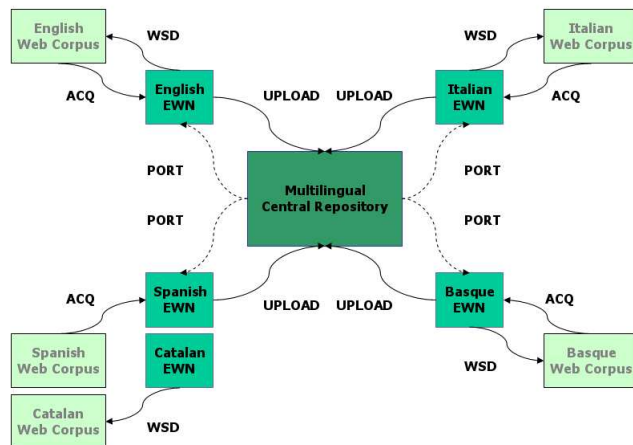


Figure 2.2: MCR architecture

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 2.5: Main sources of semantic relations

Type of Relations	#relations
Total Relations	1,642,389
Different Relations	1,531,380
Unique Relations	1,390,181
Non-unique relations (>1)	70,425
Non-unique relations (>2)	341
Non-unique relations (>3)	8

Table 2.6: Overlapping relations in the MCR

2.6 Large-Scale Knowledge Acquisition

There is a constant increasing interest of researchers in large-scale Knowledge Acquisition, in particular large knowledge resources. In order to be used in several Natural Language Processing Tasks that require it, such as Word Sense Disambiguation, Question Answering, Information Retrieval, etc.

Lexical semantics (the semantics of words) is becoming an increasingly important part of Natural Language Processing, as researchers and systems are beginning to address semantics at large scale.

Several projects are working to acquire large-scale Knowledge Resources such:

- Open Mind Common Sense⁶: is a knowledge acquisition system designed to acquire commonsense knowledge from the general public using a web⁷.
- MindNet (L.Vanderwende et al., 2005): is a knowledge representation project that uses our broad-coverage parser to build semantic networks from dictionaries, encyclopedias, and free text. MindNets are produced by a fully automatic process that takes the input text, sentence-breaks it, parses each sentence to build a semantic dependency graph (Logical Form), aggregates these individual graphs into a single large graph, and then assigns probabilistic weights to subgraphs based on their frequency in the corpus as a whole.
- Omega⁸ (A.Philpot et al., 2005) a large terminological ontology obtained by re-emerging WordNet and Mikrokosmos, adding information from various sources, and subordinating the result to newly designed feature-oriented upper model.

There also work related to Lin and Pantel (Lin & Pantel, 1994), related to Knowledge Acquisition. They automatically discover concepts from text using a clustering algorithm called CBC (Clustering by Committee).

⁶<http://www.kurzweilai.net/meme/frame.html?main=/articles/art0371.html>

⁷<http://openmind.media.mit.edu>

⁸<http://omega.isi.edu>

2.6.1 Topic Signatures

Topic Signatures (TS) are word vectors related to a particular topic. Topic Signatures are built by retrieving context words of a target topic from large text collections. In most cases, the word senses of WN have been considered as topics. In particular, the task consists on:

- A) Acquire the best possible corpus examples for a particular word sense.
- B) Build the TS by deriving from the selected corpora the context words that best represents the word sense.

TS have been used in a variety of ways, such as in Summarization Tasks (Lin & Hovy, 2000), ontology population (Alfonseca et al., 2004) where they compare different weighting measures to create TS and approximate the link distance between synsets in WordNet (Fellbaum, 1998a), or word sense disambiguation (WSD) (Agirre et al., 2000) and (Agirre et al., 2001b). (Agirre & de Lacalle, 2003) shows that the best method to cluster wordnet nominal senses in comparison to other methods proposed in the same work is by using TS. (Agirre & de la Calle, 2004) provide the Topic Signatures for all nominal senses of WordNet using the web as a corpus, while (Cuadros et al., 2005) and (Cuadros et al., 2006) explore the acquisition of Topic Signatures using BNC as a corpus.

Obviously, part of the success for building high quality TS resides on acquiring high quality sense examples. The acquisition of sense examples could be done using different techniques. There are works related to the acquisition of Sense Examples using WordNet as a knowledge base to query the Web (Leacock et al., 1998), (Mihalcea & Moldovan, 1999a), (Agirre & Martinez, 2000), (Agirre & de la Calle, 2004) and querying a controlled Corpora (Cuadros et al., 2005), (Cuadros et al., 2006). In (Cuadros et al., 2005), (Cuadros et al., 2006), the tool used to query the corpora to retrieve sense examples is ExRetriever (Cuadros et al., 2004).

Table 2.7 shows an example of Topic Signature related to the word party from the TS from the WEB⁹.

⁹<http://ixa.si.ehu.es/Ixa/resources/sensecorpus>

1. sense: party, political_party "an organization to gain political power;"

democratic 12.66; tammany 12.42; alinement 12.24; federalist 11.59; missionary 10.33; whig 09.99; greenback 08.93; anti-masonic 08.34; nazi 08.11; republican 07.41; alcoholics 07.39; bull 07.07; socialist 06.26; organization 06.06; conservative 05.92; populist 05.38; dixiecrats 05.18; know-nothing 04.96; constitutional 04.51; pecking 04.37; democratic-republican 04.07; republicans 03.96; labor 03.92; salvation 03.83; moose 03.62; farmer-labor 03.43; labour 03.28;

Table 2.7: First words of the noun party sense 1 from the TS from the WEB

2.7 Evaluation Frameworks

There has been an increasing interest in the evaluation of NLP applications in recent years, particularly in the field of natural language understanding (NLU). Evaluations of NLU systems go back to the US-organised evaluation programmes of the 1980s such as TREC¹⁰ or DUC¹¹. However these evaluations programs were designed to evaluate whole systems, usually on large tasks.

This type of evaluation has become quite widespread within NLP, with the more recent introduction of evaluations of systems performing a wider range of more specialised tasks. For example Senseval¹² (de-facto standard benchmark for WSD systems).

In other tasks such as Word Sense Disambiguation the only way to test the systems is using a semantically-labelled corpus with a specific Knowledge Resource such as WordNet. (Klapaftis & Manandhar, 2005) test their systems using Semcor, while the vast majority use a specific Word Sense Disambiguation Framework obtained from an International Word Sense Disambiguation Competition, Senseval. (Mihalcea & Moldovan, 1999b) test their system with human evaluators.

Senseval competition consist of several Word Sense Disambiguation tasks. The only drawback of this evaluation framework, is the size of the set of evaluation data and the poor coverage.

There are two main differences in the evaluation methodologies, the direct and the

¹⁰<http://trec.nist.gov/>

¹¹<http://duc.nist.gov/>

¹²<http://www.senseval.org>

indirect. The main issue in the direct evaluation is using a Framework which was developed for evaluating this specific task while in the indirect evaluation framework, the task is evaluated with an Evaluation Framework not conceived for.

Knowledge Resource Evaluation is difficult. There is not a common approach neither a direct Evaluation Framework to evaluate Knowledge Bases, neither a concrete framework to see which are better. (Cuadros & Rigau, 2006) evaluate Knowledge Resources Indirectly using them as Topic Signatures in Senseval. In the Meaning project, the IRION Company used MCR in a search engine called Twenty One, as an Evaluation Framework.

Chapter 3

Our approach

3.1 Main Goal

This chapter presents an overview of our current research on building high quality knowledge resources, further detailed in the next chapters.

After studying the different approaches appearing in the literature for acquiring open domain knowledge resources and studying its main characteristics (see chapter 2), we present a general overview of the complete process we plan to perform for building automatically highly connected wordnets of several orders of magnitude larger from the current WordNet version. This process will exploit the acquired high quality Topic Signatures by disambiguating the words appearing in their vectors. Having the large collections of words (with Part-of-Speech and weight) captured for each synset in WordNet, our plan consist on disambiguating accurately those words (assigning the correct sense for each word of the Topic Signature). In that way, a new and very large resource with millions of new relations between synsets will be derived automatically and uploaded into the Multilingual Central Repository (MCR).

Our current research is presented in the next three chapters; Chapter 4, 5 and 6.

- Chapter 4 presents an empirical evaluation of the quality of the knowledge resources currently available.
- Chapter 5 studies the relative performance of different ways for automatically de-

iving Topic Signatures, in order to find out which are the best parameters to acquire them. In order to compare the knowledge already present in these resources, this Chapter also provides an evaluation of the combination of these knowledge resources.

- Chapter 6 presents a proposal for disambiguating the words appearing in the Topic Signatures. As a side effect, this process will produce large volumes of new relations between synsets which we plan to integrate into the MCR.

3.1.1 Illustrative Example

In order to illustrate the whole process, this section provides a small example using the word *party* of the Senseval-3 English Lexical Sample Task. *party* has 5 nominal senses in WordNet1.6.

Table 3.1 shows which are the semantics of the word *party* and the information related to its synsets, the gloss and the synonyms.

Table 3.2 and 3.3 show the first twenty words of the Topic Signatures associated to the first two senses of *party*. Obviously, both sets of words represent very different contexts and different underlying topic.

Table 3.4 show how many relations have these synsets in the existing MCR nowadays. The source of the relations are the following:

- Manual: all manually derived relations, mainly from WordNet1.6 and 2.0 (Fellbaum, 1998b).
- SemCor: They are the selectional preferences acquired from Semcor (Agirre & Martinez, 2001; Agirre & Martinez, 2002).
- BNC: They are the selectional preferences acquired from BNC (McCarthy, 2001).
- XWN: They are the selectional preferences acquired from XWN (Mihalcea & Moldovan, 2001).
- MCR: They are all the relations contained into the MCR (Rigau et al., 2002).

<i>Sense 1</i> <party, political party>	(an organization to gain political power; "in 1992 Perot tried to organize a third party at the national level") organization, organisation – (a group of people who work together)
<i>Sense 2</i> <party>	(an occasion on which people can assemble for social interaction and entertainment; "he planned a party to celebrate Bastille Day") affair, occasion, social occasion, function, social function – (a vaguely specified socialevent; "the party was quite an affair"; "an occasion arranged to honor the president"; "a seemingly endless round of social functions")
<i>Sense 3</i> <party, company>	(a band of people associated temporarily in some activity; "they organized a party to search for food"; "the company of cooks walked into the kitchen") set, circle, band, lot – (an unofficial association of people or groups; "the smart set goesthere"; "they were an angry lot")
<i>Sense 4</i> <party>	(a group of people gathered together for pleasure; "she joined the party after dinner") social gathering, social affair – (a gathering for the purpose of promoting fellowship)
<i>Sense 5</i> <party>	(a person involved in legal proceedings; "the party of the first part") person, individual, someone, somebody, mortal, human, soul – (a human being; "there was too much for one person to do")

Table 3.1: WordNet 1.6 senses for noun party, the gloss and the syns associated

party 0.5115; union 0.5041; trade 0.4939; political 0.3220; government 0.2216; support 0.2047; movement 0.1923; people 0.1700; power 0.1683; local 0.1661; years 0.1649; election 0.1625; policy 0.1615; nt 0.1606; leader 0.1469; year 0.1400; time 0.1400; system 0.1371; conference 0.1326; unions 0.1309
--

Table 3.2: First twenty words related to the noun party sense 1

birthday 1.8875; party 1.8677; tea 0.9478; cocktail 0.9162; house 0.4577; dance 0.4030; ceilidh 0.3996; barn 0.3797; photo 0.3067; teaparty 0.2759; night 0.2732; nt 0.2604; guests 0.2458; opportunity 0.2399; st 0.2249; photoopportunity 0.2170; people 0.2010; houseparty 0.2005; music 0.1813; acid 0.1811

Table 3.3: First twenty words rated to the noun party sense 2

Obviously this table presents a very different number of relations depending on the method used. It is worth highlighting that the number of relations obtained automatically is bigger than those obtained manually. However, *Are those automatically acquired relations of a comparable quality to the ones obtained manually?* Chapter 4 tries to clarify this point providing some clues regarding the acquisition method (manual vs. automatic). In fact, we will empirically demonstrate that merging the resources acquired by manual and automatic means, the resulting resource surpass both of them individually.

Then, Chapter 5 explores different ways for automatically deriving Topic Signatures in order to find out which are the best parameters to acquire them. In order to compare the knowledge already present in these resources, this Chapter also provides an evaluation of the combination of these knowledge resources. Again, we will empirically demonstrate that merging the resources acquired by manual and automatic means, the resulting combination surpass both of them individually.

	sense 1	sense 2
Manually	29	7
SemCor	20	7
BNC	226	181
XWN	111	15
MCR	386	210

Table 3.4: The number of relations contained in MCR related to sense 1 and 2

party#n#2	cocktail	N
	terrace	N
	baby	N
	flagstaff	N
	potter	N
	birth	N

Table 3.5: Some words related to the word party sense 2 appearing in the TS from the WEB

Finally, Chapter 6 presents a proposal for disambiguating the words appearing in the Topic Signatures. As a side effect, this process will produce large volumes of new relations between synsets which we plan to integrate into the MCR.

Table 3.5 show some words of the Topic Signature related to the second sense of noun *party*. Observing these words, we can easily imagine a semantic relation between the noun *party* sense number 2 and the related words. Furthermore, there are these relations do not appear currently into the MCR (there have not been derived by manual or automatical means).

Moreover, once disambiguated, these relations will be completely underspecified. We plan to further continue identifying the relation by means of semantic properties already present into the MCR (Climent et al., 2005). For instance, after the disambiguation process we could derive an anonymous relation between party#n#2 and terrace#n#1. However, we know from the ontological properties appearing into the MCR that terrace#n#1 is a place. Obviously, this knowledge could help to infer that the relation between party#n#2 and terrace#n#1 is LOCATION.

In that way, we plan to derive by complete automatic means a new resource highly connected that we expect to provide further semantic capabilities to the exiting NLP applications.

Chapter 4

Evaluating Large-scale Knowledge Resources

4.1 Introduction

This chapter presents an empirical evaluation of the quality of publicly available large-scale knowledge resources.

The study considers a wide range of existing manually and automatically derived large-scale knowledge resources.

Obviously, all these semantic resources have been derived using a very different set of methods, tools and corpus, resulting on a different set of new semantic relations between synsets. In fact, each resource has a different volume and accuracy figures. Although isolated evaluations have been performed by their developers in different experimental settings, to date no comparable evaluation has been carried out in a common and controlled framework.

This work tries to establish the relative quality of these semantic resources in a neutral environment.

The quality of each large-scale knowledge resource is indirectly evaluated on a Word Sense Disambiguation task. In particular, we used a well defined Word Sense Disambiguation evaluation benchmark (Senseval-3 English Lexical Sample task) to evaluate the

quality of each resource.

4.2 Senseval Evaluation Framework

Senseval ¹ is an organized competition with several Word Sense Disambiguation (WSD) tasks.

Word Sense Disambiguation is a well-know task consisting on giving a word-sense to a target word taking into account several issues such as context words, knowledge bases, lexical resources with supervised methods or unsupervised methods.

The concrete Word Sense Disambiguation tasks Senseval Evaluation Framework cover are:

- All Words Task (English and Italian).
- Lexical Sample Task (Basque, Catalan, Chinese, English, Italian, Romanian, Spanish and Swedish)
- Automatic subcategorization acquisition
- Multilingual Lexical Sample Task
- WSD of WordNet glosses
- Semantic Roles (English and Swedish)
- Logic Forms

The Word Sense Disambiguation task used in the evaluations and tests of our work is the English Lexical Task.

The English Lexical Sample Task is a Word Sense Disambiguation task based on the determining the sense of a tagged word in a sentence, having the words resting in the sentence as context for the disambiguation algorithms.

¹<http://www.senseval.org>

4.3 Knowledge Resources

The semantic resources compared in this study are the following:

- **WN**: This knowledge resource uses the direct relations encoded in WordNet 1.6 or 2.0. We also tested WN-2 (using relations at distance 1 and 2) and WN-3 (using relations at distance 1, 2 and 3) (Fellbaum, 1998a).
- **XWN**: This knowledge resource uses the direct relations encoded in eXtended WordNet (Mihalcea & Moldovan, 2001).
- **XWN+WN**: This knowledge resource uses the direct relations included in WN and XWN.
- **spBNC**: This knowledge resource contains the selectional preferences acquired from BNC (McCarthy, 2001).
- **spSemCor**: This knowledge resource contains the selectional preferences acquired from SemCor (Agirre & Martinez, 2001; Agirre & Martinez, 2002).
- **spBNC+spSemCor**: This knowledge resource uses the selectional preferences acquired from BNC and SemCor.
- **MCR-spBNC**: This knowledge resource uses the direct relations in MCR except those appearing in spBNC.
- **MCR**: This knowledge resource uses the direct relations included in MCR (Atserias et al., 2004).
- **TSWEB**: This knowledge resources uses the topic Signatures related from the web (Agirre & de la Calle, 2004).

4.4 Indirect Evaluation on Word Sense Disambiguation

Knowledge Resources Evaluation is difficult. There is not a common approach to evaluate them, neither a concrete framework to see which is better. (Cuadros & Rigau, 2006) evaluate Knowledge Resources using them as Topic Signatures.

In order to measure the quality of the knowledge resources considered in the previous section, we performed an indirect evaluation by using all these resources as Topic Signatures (TS). That is, word vectors with weights associated to a particular synset from all senses of a set of words. Word Vectors are obtained by collecting those specific relations related to the synset and from a concrete Knowledge Resource. This simple representation tries to be as neutral as possible with respect the evaluation framework.

All knowledge resources are directly evaluated on a Word Sense Disambiguation (WSD) task; In particular on the noun-set of Senseval-3 English Lexical Sample task which consist of 20 nouns. All performances are evaluated on the test data using the fine-grained scoring system provided by the organizers.

Furthermore, trying to be as neutral as possible with respect to the semantic resources studied, we applied systematically the same disambiguation method to all of them. Recall that our main goal is to establish a fair comparison of the knowledge resources instead of providing the best disambiguation technique for a particular semantic knowledge base.

The Evaluation procedure used, follows these steps:

- We perform this procedure for each noun-test example, between the test and all the Topic Signatures related to the possible senses of the noun:
 - A simple word overlapping (by occurrence or weight) is defined as:
 - * **occurrence** evaluation measure: counts the amount of words from the TS occurring in the sense examples.
 - * **weight** evaluation measure: adds up the weights of the words occurring from the TS in the sense example.

The synset having higher overlapping for a chosen method is selected for a particular test example.

4.5 Evaluating the quality of knowledge resources

In order to establish a clear picture of the current state-of-the-art of publicly available wide coverage knowledge resources we will consider a number of basic baselines.

4.5.1 Baselines

We have designed several baselines in order to establish a relative comparison of the performance of each semantic resource:

- **RANDOM**: This method selects a random sense for each target word. This baseline can be considered as a lower-bound.
- **WordNet MFS (WN-MFS)**: This method selects the most frequent sense (the first sense in WordNet) of the target word.
- **TRAIN-MFS**: This method selects the most frequent sense in the training corpus of the target word.
- **Train Topic Signatures (TRAIN)**: This baseline uses the training corpus to directly build a Topic Signature using TFIDF measure for each word sense. Note that in this case, this baseline can be considered as an upper-bound of our evaluation framework.²

Formula 4.1 presents the F1 measure (harmonic mean of recall and precision) we have considered in our evaluations.

Table 4.1 shows the precision, recall and f1 measures for these baselines.

In the table, TRAIN has been calculated with a fixed vector size of 450 words. As it was expected, the RANDOM baseline obtains the poorest results while the most frequent sense of WordNet (WN-MFS) is very close to the most frequent sense of the the training corpus (TRAIN-MFS). Nevertheless both are far below from the Topic Signatures acquired using the training corpus (TRAIN).

$$F1 = \frac{2 \times P \times R}{P + R} \quad (4.1)$$

²This TS is built following part of the process EXRET TS from chapter 5 follow

Baselines	P	F1	R
TRAIN	65.2	65.1	65.1
TRAIN-MFS	54.5	54.5	54.5
WN-MFS	53.0	53.0	53.0
RANDOM	19.1	19.1	19.1

Table 4.1: P, R and F1 measure of the Baselines

4.5.2 Performance of the knowledge resources

Table 4.2 presents the performance of each knowledge resource uploaded into the MCR. The best results for precision, recall and F1 measures appear in bold. The lowest result is obtained by the knowledge gathered directly from WN mainly due to its poor coverage (Recall of 17.6 and F1 of 25.6). Its performance is improved using words at distance 1 and 2 (F1 of 33.3), but it decreases using words at distance 1, 2 and 3 (F1 of 30.4). The best precision is obtained by WN (46.7). But the best performance is achieved by the combined knowledge of MCR-spBNC (Recall of 42.9 and F1 of 44.1). This represents a recall 18.5 points higher than WN. That is, the knowledge integrated into the MCR (WordNet, extended WordNet and the selectional preferences acquired from SemCor) although partly derived by automatic means performs much better in terms of precision, recall and F1 measures than using the knowledge currently present in WordNet alone. It also seems that the knowledge from spBNC always degrades the performance of their combinations³.

Figure 4.1 shows the performance of all knowledge resources with Baselines in terms of F1 and vector-size.

Regarding the baselines, all knowledge resources integrated into the MCR surpass RANDOM, but none achieves neither WN-MFS, TRAIN-MFS nor TRAIN.

4.5.3 Senseval-3 system performances

For sake of comparison, tables 4.3 and 4.4 present the F1 measure of the fine-grained results for nouns of the the Senseval-3 lexical sample task for the best and worst unsupervised and

³All selectional preferences acquired from SemCor or BNC have been considered including those with very low confidence score

KB	P	R	F1
MCR-spBNC	45.4	42.9	44.1
MCR	41.8	40.4	41.1
spSemCor	43.1	38.7	40.8
spBNC+spSemCor	41.4	30.1	40.7
WN+XWN	45.5	28.1	34.7
WN-2	38.0	29.7	33.3
XWN	45.0	25.6	32.6
WN-3	31.6	29.3	30.4
spBNC	36.3	25.4	29.9
WN	46.7	17.6	25.6

Table 4.2: Fine-grained results for the resources integrated into the MCR

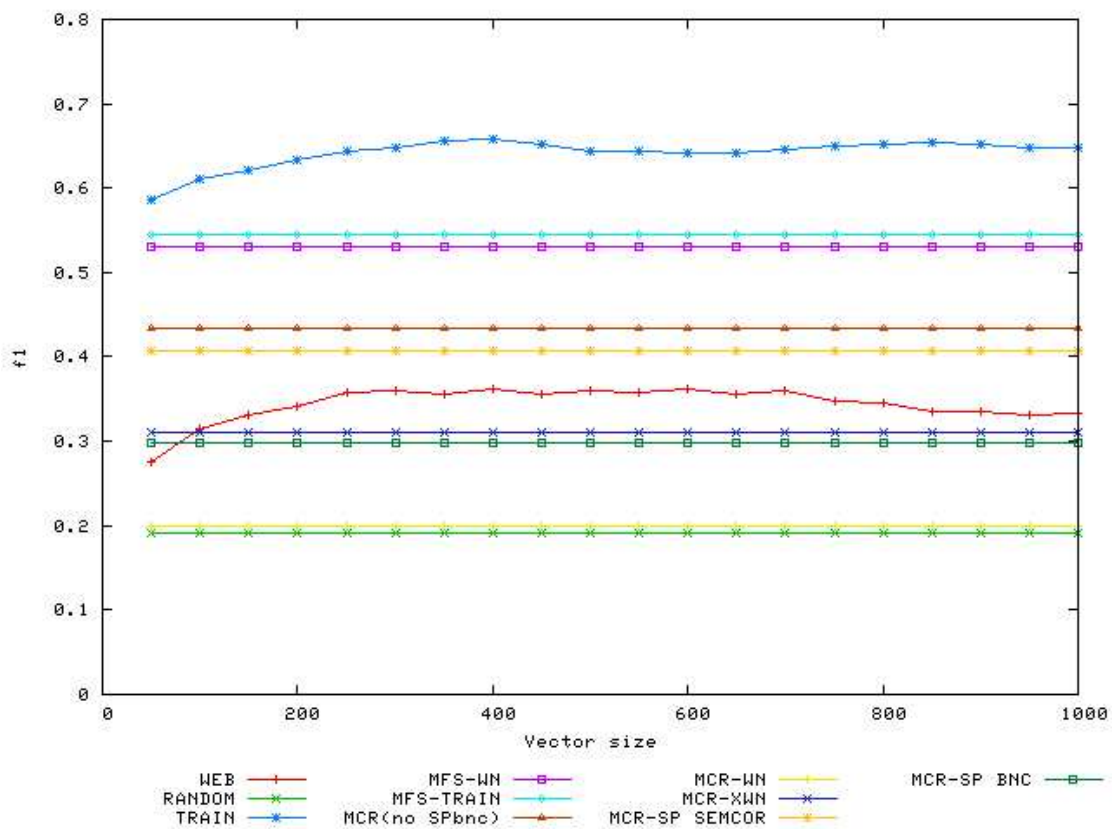


Figure 4.1: Fine-grained evaluation results for the knowledge resources with the baselines related. The figure presents F1 (Y-axis) in terms of the size of the word vectors (X-axis).

s3 systems	F1
s3_wsdiit	68.0
s3_Cymfony	57.9
WN-MFS	53.0
MCR-spBNC	44.1
s3_DLSI	17.8

Table 4.3: Senseval-3 Unsupervised Systems

s3 systems	F1
htsa3 U.Bucharest (Grozea)	74.2
TRAIN	65.1
TRAIN-MFS	54.5
DLSI-UA-LS-SU U.Alicante (Vazquez)	41.0

Table 4.4: Senseval-3 Supervised Systems

supervised systems, respectively.

We must recall that the main goal of this research is to establish a clear and neutral view of the relative quality of available knowledge resources, rather than to provide the best Word Sense Disambiguation algorithm using these resources. Obviously, much more sophisticated WSD systems using these resources could be devised.

Although we maintain the classification of the organizers, s3_wsdiit used the train data so we in our opinion used supervised knowledge.

4.6 Conclusions and future work

Summarizing, this study provides empirical evidence for the relative quality of publicly available large-scale knowledge resources. The relative quality has been measured indirectly in terms of precision, recall and f1 on a WSD task.

The study empirically demonstrates that automatically acquired knowledge bases (extended WordNet, Selectional Preferences acquired automatically from SemCor and BNC) clearly surpass both in terms of precision and recall to the knowledge manually encoded from WordNet (using relations expanded to one, two and three levels).

The better results have been obtained using occurrences (the weights are only used to order the words of the vector). Finally, we should remark that the results are not skewed

(for instance, for resolving ties) by the most frequent sense in WordNet or any other statistically predicted knowledge.

Chapter 5

Large-scale Knowledge Acquisition

In this section we present the different techniques and methods we have developed automatically to acquire large-scale open-domain Knowledge Resources. In our case, the acquired Knowledge Resources are Topic Signatures (Cuadros et al., 2005; Cuadros et al., 2006).

5.1 Retrieving automatically Topic Signatures

Topic Signatures (TS) are word vectors related to a particular topic (Lin & Hovy, 2000). Topic Signatures are built by retrieving context words of a target topic from large volumes of text. For this study we would use the large-scale Topic Signatures acquired from the web (TSWEB) (Agirre & de la Calle, 2004) and those acquired from the BNC (EXRET) (Cuadros et al., 2005) using ExRetriever (Cuadros et al., 2004).

The acquisition process followed by both approaches is the following:

- **TSWEB:** Topic Signatures acquired from the web ¹. Inspired by the work of (Leacock et al., 1998), these Topic Signatures were constructed using monosemous relatives from WordNet (synonyms, hypernyms, direct and indirect hyponyms, and siblings) querying Google and retrieving up to one thousand snippets per query (that is, a word sense). In particular, the method was as follows:

¹<http://ixa.si.ehu.es/Ixa/resources/sensecorpus>

- Organizing the retrieved examples from the web in collections, one collection per word sense.
- Extracting the words and their frequencies for each collection.
- Comparing these frequencies with those pertaining to other word senses using *tfidf* (formula 5.1).
- The words with distinctive frequency for one of the collections are gathered in an ordered list, which constitutes the Topic Signature for the respective word sense.

TSWEB Constitutes the largest available semantic resource with around 100 million relations (between synsets and words).

- **EXRET**: ExRetriever (Cuadros et al., 2004)² is a flexible tool to perform sense queries on large corpora.
 - This tool characterizes each sense of a word as an specific query using a declarative language.
 - This is automatically done by using a particular query construction strategy, defined *a priori*, and possibly using information from a knowledge base, such as WN.

In this study, ExRetriever has been tested using the BNC, WordNet as a knowledge base and three different measures (TFIDE, as shown in formula 5.1 (Agirre & de la Calle, 2004), Mutual Information (Church & Gale, 1991) and Association Ratio (Rigau et al., 1997)).

$$tfidf(w, C) = \frac{wf_w}{\max_w wf_w} \times \log \frac{N}{Cf_w} \quad (5.1)$$

²<http://www.lsi.upc.es/nlp/meaning/downloads.html>

$$MI(w, s) = \log \frac{P(w \wedge s)}{P(w)P(s)} \quad (5.2)$$

$$AR(w, C) = Pr(w/C) \log_2 \left(\frac{Pr(w/C)}{Pr(w)} \right) \quad (5.3)$$

Where w stands for word context and C stands for Collection (all the corpus gathered for a particular word sense). For TFIDE, wf , and Cf stands for the frequency of words and Collections.

The strategy used for building the word sense queries is very important. In this study, considered two different query strategies:

- Monosemous A (**queryA**): (OR monosemous-words). That is, the union set of all synonym, hyponym and hyperonym words of a WordNet synset which are monosemous nouns (these words can have other senses as verbs, adjectives or adverbs).
- Monosemous W (**queryW**): (OR monosemous-words). That is, the union set of all words appearing as synonyms, direct hyponyms, hypernyms indirect hyponyms (distance 2 and 3) and siblings. In this case, the nouns collected are monosemous having no other senses as verbs, adjectives or adverbs.

TSWEB used the query construction **queryW**, and ExRetriever used both.

5.1.1 Examples

Figure 5.1 presents an example of Topic Signature from TSWEB using **queryW** and the web. Figure 5.2 and 5.3 show an example of Topic Signature from EXRET using **queryA** and the BNC for the first sense of the noun party.

Although both automatically acquired TS seems to be closely related to the first sense of the noun party having the gloss "*an organization to gain political power:*", they do not have words in common.

democratic	0.0126	socialist	0.0062
tammany	0.0124	organization	0.0060
alinement	0.0122	conservative	0.0059
federalist	0.0115	populist	0.0053
missionary	0.0103	dixiecrats	0.0051
whig	0.0099	know-nothing	0.0049
greenback	0.0089	constitutional	0.0045
anti-masonic	0.0083	pecking	0.0043
nazi	0.0081	democratic-republican	0.0040
republican	0.0074	republicans	0.0039
alcoholics	0.0073	labor	0.0039
bull	0.0070	salvation	0.0038

Table 5.1: Topic Signature for party#n#1 using TSWEB (24 out of 15881 total words)

party	4.9350	trade	1.5295
political	3.7722	parties	1.4083
government	2.4129	politics	1.2703
election	2.2265	campaign	1.2551
policy	2.0795	leadership	1.2277
support	1.8537	movement	1.2156
leader	1.8280	general	1.2034
years	1.7128	public	1.1874
people	1.7044	member	1.1855
local	1.6899	opposition	1.1751
conference	1.6702	unions	1.1563
power	1.6105	national	1.1474

Table 5.2: Topic Signature for party#n#1 using ExRetriever (queryA) with TFIDF (24 out of 9069 total words)

party	0.5115	policy	0.1615
union	0.5041	nt	0.1606
trade	0.4939	leader	0.1469
political	0.3220	year	0.1400
government	0.2216	time	0.1400
support	0.2047	system	0.1371
movement	0.1923	conference	0.1326
people	0.1700	unions	0.1309
power	0.1683	work	0.1202
local	0.1661	leaders	0.1193
years	0.1649	public	0.1178
election	0.1625	working	0.1170

Table 5.3: Topic Signature for party#n#1 using ExRetriever(queryW) with TFIDF (24 out of 9069 total words)

```
<instance id="party.n.bnc.00008131" docsrc="BNC"> <context> Up to
the late 1960s , catholic nationalists were split between two main political
groupings . There was the Nationalist Party , a weak organization
for which local priests had to provide some kind of legitimation . As a
<head>party</head> , it really only exercised a modicum of power in
relation to the Stormont administration . Then there were the republican
parties who focused their attention on Westminster elections . The disor-
ganized nature of catholic nationalist politics was only turned round with
the emergence of the civil rights movement of 1968 and the subsequent
forming of the SDLP in 1970 . </context> </instance>
```

Table 5.4: Example of test num. 00008131 for party#n which its correct sense is 1

As an example, table 5.4 shows a test example of Senseval-3 corresponding to the first sense of the noun party. In bold there are the words that appear in EXRET-queryA. There are several important words that appear in the text that also appear in the TS.

5.2 Experiments

A wide set of Experiments have been developed on the nominal part of Senseval-3 English Lexical Sample task using TSWEB and EXRET in order to measure the effect of several parametres when building TS: The considered features were:

- Corpus: British National Corpus vs WEB.
- Measures: TFIDF (formula 5.1), MI (formula 5.2) and AR (formula 5.3).
- Knowledge Resources: WordNet.
- Query construction: queryW vs queryA

While TSWEB was developed using TFIDE, queryW and the information retrieved from the web, EXRET has been developed using TFIDF, MI and AR, the two possible query constructions strategies using BNC.

The Knowledge Base used for the different query construction strategies has been WordNet.

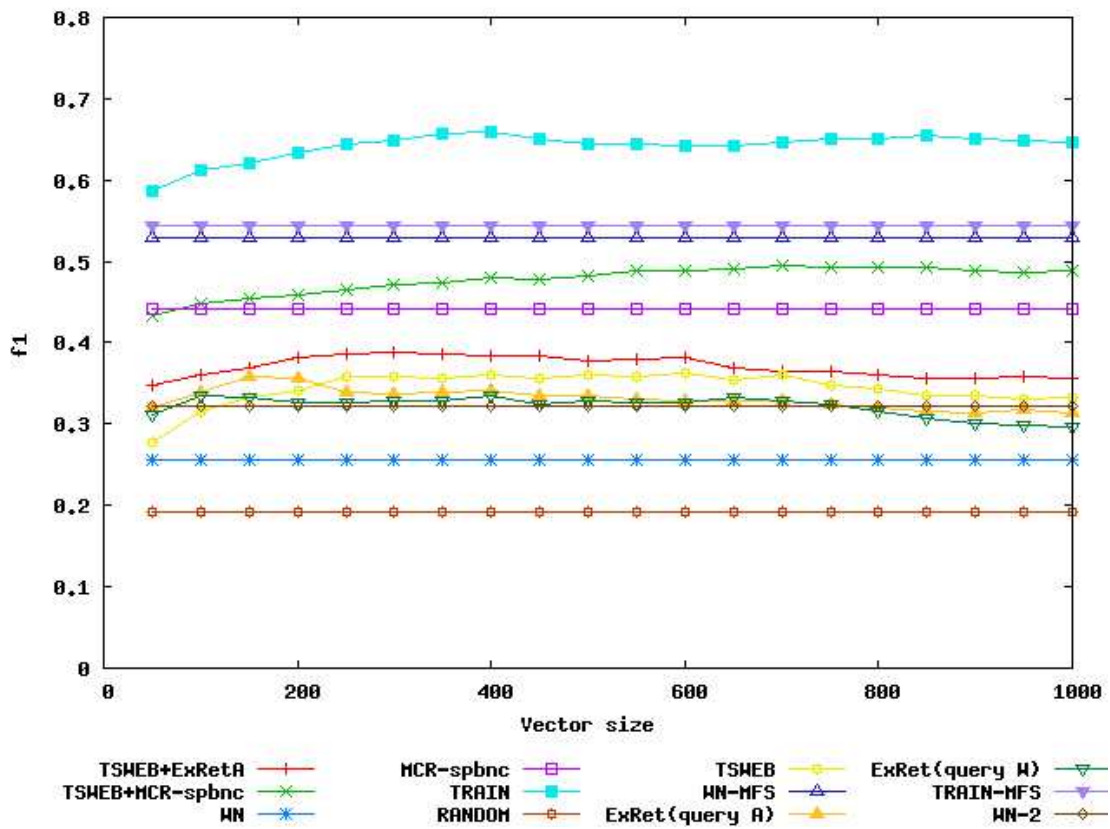


Figure 5.1: Fine-grained evaluation results for the knowledge resources and the combination. The figure presents F1 (Y-axis) in terms of the size of the word vectors (X-axis).

Included at figure 5.1, there are the best performing systems of methods TSWEB and EXRET with queryW and queryA. The best performing measure is TFIDF, while the higher performing query is queryW. The results obtained querying the WEB rate higher than the ones querying BNC using the same query construction queryW. In fact, the different query strategies perform quite similar using the same corpus BNC.

5.3 Combination of Knowledge Resources

In order to evaluate more deeply the quality of each knowledge resource, we also provide some evaluations of the combined outcomes of several knowledge resources. That is, to which extent each knowledge base provides new knowledge not provided by the others. The combinations designed are based on two or three Knowledge Resources, the ones performed are the following combinations:

- Baseline + MCR
- Baseline + TSWEB/EXRETA/EXRETW
- Baseline + MCR + TSWEB/EXRETA/EXRETW
- MCR + TSWEB/EXRETA/EXRETW
- TSWEB + EXRETA/EXRETW

Where MCR stands for any of the Knowledge Resources defined in the previous section, Baseline stands for any of the considered Baseline and TSWEB/EXRETA/EXRETW stands for the Automatically Acquired Topic Signatures.

The combinations are performed following a very simple voting method for each Knowledge Resource: For each word of all the combined KR:

- The scoring results of the word sense are normalized.
- The normalized scores of the word senses are added up.
- The word sense with higher score is selected.

Figure 5.1 plots F1 results of the fine-grained evaluation of the baselines (including upper and lower-bounds), the knowledge bases integrated into the MCR, the best performing Topic Signatures acquired from the web and the BNC evaluated individually and in combination with others. The figure presents F1 (Y-axis) in terms of the size of the word vectors (X-axis).

The main features of the different Topic Signatures are the corpus, the knowledge Resource used in the query Construction, the frequency measure used and in the Topic Signature.

Sumarizing:

- Corpus: Regarding Topic Signatures, as expected, in general the knowledge gathered from the web (TSWEB) is superior to the one acquired from the BNC either using queryA or queryW (EXRET-queryA and EXRET-queryW).
- Query Construction: The performance of EXRET-queryA and EXRET-queryW is pretty similar in general. Specifically, when using the first two hundred words of the TS queryA is slightly better than using queryW.
- Knowledge: Although EXRET-queryA and EXRET-queryW perform very similar, both knowledge resources contain different knowledge. This is shown when combining the outcomes of these two different knowledge resources with TSWEB.
- Combinations:
 - While no improvement is obtained when combining the knowledge acquired from the web (TSWEB) and the BNC (EXRETqueryW) when using the same acquisition method (queryW), the combination of TSWEB and EXRETqueryA (TSWEB+ExRetA) obtains better F1 results than TSWEB (EXRETqueryA have some knowledge not included into TSWEB).

- The knowledge already integrated into the MCR (MCR-spBNC) surpasses the knowledge from the Topic Signatures acquired from the web or the BNC, using queryA, queryW or their combinations.
- The combination of TSWEB and MCR-spBNC (TSWEB+MCR-spBNC) outperforms both resources individually indicating that both knowledge bases contain complementary information. The maximum is achieved with vectors of at most 750 words. In fact, the resulting combination is very close to the most frequent sense baselines. This fact indicates that the resulting large-scale knowledge resource almost encodes the knowledge necessary to be a most frequent sense tagger.

5.4 Conclusions and Future Work

The main contributions of this section are:

- **Conclusions related to the Topic Signatures Acquisition:** Regarding the automatic acquisition of large-scale Topic Signatures those acquired from the web are slightly better than those acquired from smaller corpora (for instance, the BNC). QueryW performs better than queryA but both methods (queryA and queryW) also produce complementary knowledge. Finally, TFIDF performs better than Association Ratio and Mutual Information. However, these weights were not useful for measuring the strength of a vote (they were only useful for ordering the words in the Topic Signature).
- **Conclusions related to the Combination of TS:** The knowledge contained into the MCR (WordNet, eXtended WordNet, Selectional Preferences acquired automatically from SemCor) is of a better quality than the automatically acquired Topic Signatures. In fact, the knowledge resulting from the combination of all these large-scale resources outperforms each resource individually indicating that these knowledge resources contain complementary information. Finally, we should remark that the

resulting combination is very close to the most frequent sense classifiers.

Once empirically demonstrated that the knowledge resulting from MCR and Topic Signatures acquired from the web is complementary and close to the most frequent sense classifier, we plan to integrate the Topic Signatures acquired from the web (of about 100 million relations) into the MCR. This process will be performed by disambiguating the Topic Signatures. That is, trying to obtain word sense vectors instead of word vectors. This will allow to enlarge the existing knowledge bases in several orders of magnitude by fully automatic methods.

Chapter 6

A proposal for disambiguating large-scale knowledge resources

This chapter describes the Word Sense Disambiguation process that we plan to perform in the automatically acquired Topic Signatures. That is, building word sense vectors instead of word vectors. In fact, we plan to integrate the resulting TS into the MCR.

6.1 Aim

In computational linguistics, word sense disambiguation (WSD) is the problem of determining in which sense a word having a number of distinct senses is used in a certain context. In our case, the context associated to a word sense would be a bag of words with a part-of-speech and a related weight (Topic Signatures).

The final goal of this disambiguation process is to enrich the existing Knowledge Resources by 40 times the number of the current relations.

6.2 Disambiguating process

The disambiguating process is focused on the point that we want to disambiguate the words contained in Topic Signatures related to a particular noun sense. The Topic Signa-

tures we would disambiguate are the Topic Signatures from the web (TSWEB)¹.

Each Topic Signatures contains a set of words with PoS related to an specific topic (word sense).

The main goal is to get out the sense of each of these words related to the noun sense according to WordNet.

We have devised several methods to disambiguate Topic Signatures, based on several techniques. The overviewed methods are:

- Method-1: This simple method uses the Topic Signatures to disambiguate the words associated to a particular word. This method selects the sense having associated a TS with the same word we are disambiguating (in figure 6.1 party).
- Method-2: This method counts the words between the topic Signature of the word to be disambiguated. The synset most overlapped, is selected (see figure 6.2).
- Method-3: (Figure 6.3) Instead of using the TS, other Knowledge resources and techniques could be used. For instance, WordNet, MCR (Multilingual Central Repository) together with the WordNet Similarity Package (Patwardhan & Pedersen, 2003).
- Method-4: (Figure 6.3) This method uses the same resources used in Method-3 but only comparing the words of the Topic Signature (not unique the word-sense of the TS).
- Method-5: An applicability of the Relaxation Labelling algorithm (Padró, 1996) will be considered.

The important issue in this disambiguation approach is the combining methodology used with the defined methods using a voting system to get the final WSD system.

6.3 Preliminaries Results

We already performed Method-1 and Method-2 for all the 20 nouns of Senseval-3 Lexical Sample Task in order to obtain a preliminary coverage for these methods and a variety of

¹<http://ixa.si.ehu.es/Ixa/resources/sensecorpus>

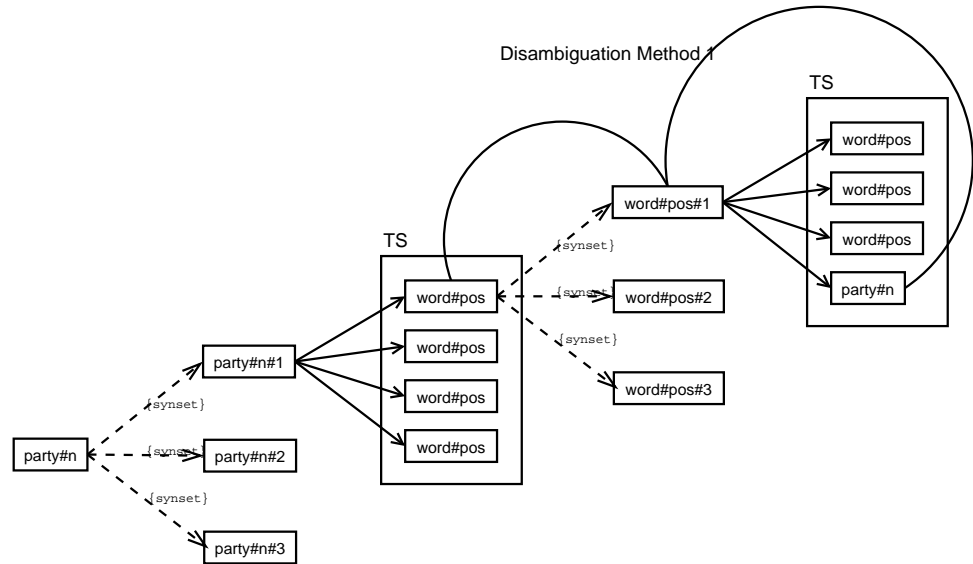


Figure 6.1: Method 1 for WSD

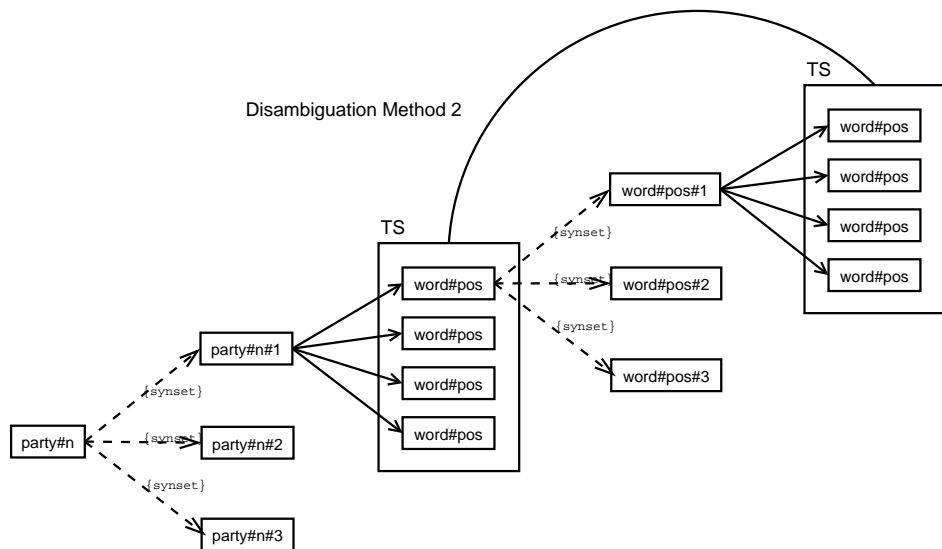


Figure 6.2: Method 2 for WSD

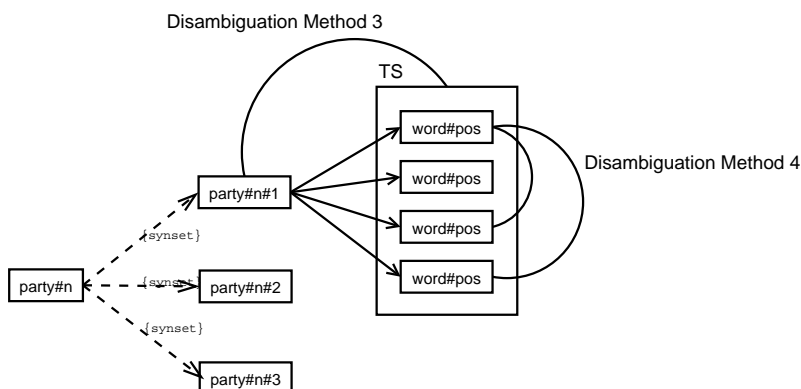


Figure 6.3: Method 3 and 4 for WSD

nouns.

Methods 1 and 2 require different Topic Signatures to the ones to be disambiguated (for instance, using the Topic Signatures associated to the words appearing in the Topic Signature to be disambiguated). However, methods 3 and 4 only require the Topic Signature of the noun to be disambiguated (for instance, the Topic Signature for the noun *party*). Fortunately, TSWEB is available for almost all nominal senses of WordNet allowing the possibility to test the performance of these methods for the four methods (for instance, ExRet-QueryA and ExRetQueryW where build only for the nouns of the Senseval-3 English Lexical Sample task).

In order to estimate the productivity of these approaches, we performed some initial calculations. In average, about 55% of the words in the Topic Signatures of TSWEB are nouns. From them, the vast majority (about 82%) is present in WordNet (and obviously have a Topic Signature for each synset). In average each synset belonging to the nominal part of the Senseval-3 English Lexical Sample task has Topic Signatures of 2,718 words. Thus, in average methods 1 and 2 could be applied to 1,256 words per synset.

Although we did not performed a formal evaluation, method-1 in average can be applied 73% of the times (around 916 words per synset) while method-2 can be applied 82% of the times (around 1,029 words per synset). And for the nouns appearing in the Topic Signature, methods 3 and 4 can be applied 82% of the times (around 1,029 words per synset).

These means that we can estimate the maximum productivity of these approaches to be of about thousand of new semantic relations per synset.

These means an increase of the size of about 3500% of the number of relations of Word-Net.

Obviously, these figures do not estimate the quality of the new relations obtained, which will be further studied. However, we also expect that these process could filter out not relevant words of the Topic Signature.

Chapter 7

Thesis project

In this chapter, we present an overview of the current situation and plans of the thesis project.

7.1 Research roadmap

First, a subset of publicly available large-scale knowledge resources have been evaluated in a common framework. We have also performed a large set of experiments on acquiring wide-coverage Topic Signatures from large text collections. Finally, a preliminary but complete proposal have been devised in order to fully integrate the already acquired Topic Signatures into the Multilingual Central Repository.

Firstly, a large set of already existing knowledge sources have been evaluated using a common framework (Senseval-3 English Lexical Sample task). Surprisingly, the combination of the manually and automatically derived knowledge resources already integrated into the MCR surpass both of them individually.

Secondly, a large set of experiments have been performed in order to build high quality Topic Signatures. These techniques and the derived resources have been evaluated also using the same framework (Senseval-3 English Lexical Sample task). Surprisingly, the Topic Signatures are of a lower quality of the knowledge already present into the MCR, which is much more smaller. However, the combination of both knowledge resources also surpass both of them individually, providing enough knowledge to build a most frequent

sense tagger.

Finally, we present a complete proposal for integrating the automatically acquired Topic Signatures into the MCR. The final resource could produce a knowledge base 40 times richer than the current MCR.

The research lines to be explored in a near future are the following:

- To complete some undergoing experiments on acquiring high quality Topic Signatures.
- To disambiguate the Topic Signatures, initially based on chapter 6.
- To characterize the acquired relations by using the ontological properties of the MCR.

7.2 Working Plan

It is expected to finish the thesis at summer of 2008, which means that the project presented here will be developed in two years time. During the next six months, we expect to finish the experiments for WSD. Then, the next months will be devoted to perform some experiments on inferring the semantic relations between synsets. This research, will long approximately one year, and will be focused in one or other direction depending on the results obtained in the first months. After this period, the experimentation and the results will be drawn, and the doctoral dissertation will be written.

7.3 Related Publications

Publications related to acquisition of Topic Signatures:

- Montse Cuadros, Lluís Padró, German Rigau
An Empirical Study for Automatic Acquisition of Topic Signatures
Proceedings of Third International WordNet Conference. pp 51-59. ISBN 80-210-3915-9. Jeju Island (Korea).
- Montse Cuadros, Lluís Padró, German Rigau
Comparing Methods for Automatic Acquisition of Topic Signatures

Recent Advances in Natural Language Processing (RANLP). (p. 181-186). Borovets, Bulgaria. 21-23 September, 2005.

Publications related to acquisition of Sense Examples:

- Montse Cuadros, Jordi Atserias, Mauro Castillo, German Rigau
The MEANING approach for automatic acquisition of sense examples
MEANING Workshop'05(p. 93). Trento. 3-4 february, 2005.
- Montse Cuadros, Jordi Atserias, Mauro Castillo, German Rigau
Automatic Acquisition of Sense Examples Using Exretriever
IBERAMIA Workshop on Lexical Resources and The Web for Word Sense Disambiguation.
(p. 97-104). Mexico. 22-23 november, 2004. ISBN 968-863-786-6

Publications related to the quality of Knowledge Resources:

- Montse Cuadros and German Rigau
Quality Assessment of Large Scale Knowledge Resources
To appear at EMNLP 2006. Sydney, Australia. 22-23 July,2006

Technical Reports from the MEANING project:

Reports related to Knowledge Resources (MCR), and Knowledge acquisition:

- WP4.4 "UPLOAD1" Jordi Atserias, Eva Naqui, Montse Cuadros, German Rigau, Samir Kanaan. Technical Report. MEANING PROJECT.
- D4.2 "PORT1" Jordi Atserias, Eva Naqui, Montse Cuadros, German Rigau. Technical Report. MEANING PROJECT.
- D4.3 "PORT2" Jordi Atserias (UPC), Montse Cuadros (UPC), Eva Naqui (UPC), German Rigau (UPV/EHU) WP5.14 Experiment 5.E: Sense Examples (2nd round)" Montse Cuadros, Jordi Atserias, Mauro Castillo, German Rigau. Technical Report. MEANING PROJECT.

- WP5.2 "Experiment 5.A a: Multilingual Knowledge Acquisition (Second Round)" Alicia Ageno (UPC), Eneko Agirre (UPV/EHU), Jordi Atserias (UPC), Aitziber Atutxa (UPV/EHU), John Carroll (Sussex), Montse Cuadros (UPC), Rob Koeling (Sussex), Bernardo Magnini (ITC-Irst), Diana McCarthy (Sussex), Octavian Popescu (ITC-Irst), Francis Real (UPC), German Rigau (UPV/EHU), Horacio Rodríguez (UPC). Technical Report. MEANING PROJECT.

Reports related to acquisition of Sense Examples:

- WP5.6 "Experiment 5.F: Sense Examples (2nd round)" Juan Francisco Fernandez, Mauro Castillo Valdés, German Rigau, Jordi Atserias, Jordi Turmo, Montse Cuadros. Technical Report. MEANING PROJECT.
- WP5.14 "Experiment 5.E: Sense Examples (3rd round)" Montse Cuadros, Mauro Castillo Valdés, Jordi Atserias, German Rigau. Technical Report. MEANING PROJECT.

Acknowledgments

This research has been funded by the European Commission (MEANING,IST-2001-34460) and is being funded by IXA NLP Group from the Basque Country University, CLASS (Basque Country University project) and ADIMEN (Basque Country Government project). Part of this research has been performed at the TALP Research Center, is recognized as a Quality Research Group (2001 SGR 00254) by DURSI.

Bibliography

- Agirre, E., Ansa, O., Arregi, X., Arriola, J., Díaz de Ilarraza, A., Pociello, E., & Uria, L. (2002). Methodological issues in the building of the basque wordnet: quantitative and qualitative analysis. *Proceedings of the first International Conference of Global WordNet Association*. Mysore, India.
- Agirre, E., Ansa, O., Martinez, D., & Hovy, E. (2000). Enriching very large ontologies with topic signatures. *Proceedings of ECAI'00 workshop on Ontology Learning*. Berlin, Germany.
- Agirre, E., Ansa, O., Martinez, D., & Hovy, E. (2001a). Enriching wordnet concepts with topic signatures. *Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Pittsburg, USA.
- Agirre, E., Ansa, O., Martinez, D., & Hovy, E. (2001b). Enriching wordnet concepts with topic signatures. *Proceedings of the NAACL workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations*. Pittsburg.
- Agirre, E., & de la Calle, O. L. (2004). Publicity available topic signatures for all wordnet nominal senses. *LREC'04* (pp. 97–104).
- Agirre, E., & de Lacalle, O. L. (2003). Clustering wordnet word senses. *International Conference Recent Advances in Natural Language Processing, RANLP'03*. Borovets, Bulgaria.
- Agirre, E., & Martinez, D. (2000). Exploring automatic word sense disambiguation with decision lists and the web. *Proceedings of the COLING workshop on Semantic Annotation and Intelligent Annotation*. Luxembourg.

- Agirre, E., & Martinez, D. (2001). Learning class-to-class selectional preferences. *Proceedings of CoNLL01*. Toulouse, France.
- Agirre, E., & Martinez, D. (2002). Integrating selectional preferences in wordnet. *Proceedings of the 1st International Conference of Global WordNet Association*. Mysore, India.
- Alfonseca, E., Agirre, E., & de Lacalle, O. L. (2004). Approximating hierarchy-based similarity for wordnet nominal synsets using topic signatures. *Proceedings of the Second International Global WordNet Conference (GWC'04). Panel on figurative language*. Brno, Czech Republic. ISBN 80-210-3302-9.
- Alfonseca, E., & Manandhar, S. (2002a). Distinguishing concepts and instances in wordnet. *Proceedings of the first International Conference of Global WordNet Association*. Mysore, India.
- Alfonseca, E., & Manandhar, S. (2002b). An unsupervised method for general named entity recognition and automated concept discovery. *Proceedings of the first International Conference of Global WordNet Association*. Mysore, India.
- A.Philpot, E.Hovy, & P.Pantel (2005). The omega ontology. *In IJCNLP Workshop on Ontologies and Lexical Resources, OntoLex-05, 2005*.
- Artale, A., Magnini, B., & Strapparava, C. (1997). Lexical discrimination with the italian version of wordnet. *Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources*. Madrid. Spain.
- Atserias, J., Climent, S., Farreres, X., Rigau, G., & Rodríguez, H. (1997). Combining multiple methods for the automatic construction of multilingual wordnets. *proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP'97)*. Tzigov Chark, Bulgaria.
- Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., & Vossen, P. (2004). The meaning multilingual central repository. *Proceedings of the Second International Global WordNet Conference (GWC'04)*. Brno, Czech Republic. ISBN 80-210-3302-9.
- Benítez, L., Cervell, S., Escudero, G., López, M., Rigau, G., & M. Tauli, t. . .

- Bentivogli, L., Pianta, E., & Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. *First International Conference on Global WordNet*. Mysore, India.
- Buitelaar, P., & Sacaleanu, B. (2001). Ranking and selecting synsets by domain relevance. *NAACL Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations' (NAACL'2001)*. Pittsburg, PA, USA.
- Buitelaar, P., & Sacaleanu, B. (2002). Extending synsets with medical terms. *Proceedings of the 1st Global WordNet Association conference*. Mysore, India.
- Castillo, M., Real, F., & Rigau, G. (2003). Asignaci3n autom3tica de etiquetas de dominios en wordnet. *Proceedings of SEPLN'03*. Alcala de Henares, Spain. ISSN 1136-5948.
- Church, K. W., & Gale, W. A. (1991). A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of english bigrams. *Computer Speech and Language*, 5, 19–54.
- Climent, S., More, Q., Rigau, G., & Atserias, J. (2005). Towards the shallow ontologization of wordnet. a work in progress. *SEPLN 2005, Granada, Spain*.
- Cuadros, M., Castillo, M., Rigau, G., & Atserias, J. (2004). Automatic Acquisition of Sense Examples using ExRetriever. *Iberamia'04* (pp. 97–104).
- Cuadros, M., Padro, L., & Rigau, G. (2005). Comparing methods for automatic acquisition of topic signatures. *RANLP'05, Borovets, Bulgaria, september 2006* (pp. 181–186).
- Cuadros, M., Padro, L., & Rigau, G. (2006). An empirical study for automatic acquisition of topic signatures. *Global Wordnet Conference GWC'06, Jeju Island, Korea, 22-26 January 2006* (pp. 51–59).
- Cuadros, M., & Rigau, G. (2006). Quality assessment of large scale knowledge resources. *To appear in EMNLP'06, Sydney, Australia*.
- Daud3, J. (2004). *Mapping large-scale semantic networks*. Submitted to Phd. Thesis, Software Department (LSI). Technical University of Catalonia (UPC), Barcelona, Spain.

- Daudé, J., Padró, L., & Rigau, G. (2003). Making wordnet mappings robust. *Proceedings of the 19th Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN*. Universidad Universidad de Alcalá de Henares. Madrid, Spain.
- Fellbaum, C. (Ed.). (1998a). *WordNet. An Electronic Lexical Database*. The MIT Press.
- Fellbaum, C. (Ed.). (1998b). *Wordnet. an electronic lexical database*. Language, Speech, and Communication. The MIT Press.
- Hamp, B., & Feldweg, H. (1997). Germanet - a lexical-semantic net for german. *Proceedings of ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*. Madrid. Spain.
- Hearst, M., & Schutze, H. (1993). Customizing a lexicon to better suit a computational task. *Proceedings of the ACL SIGLEX Workshop on Lexical Acquisition*. Stuttgart, Germany.
- Kipper, K., Dang, H. T., & Palmer, M. (2000). Class-based construction of a verb lexicon. *In Proceedings of the Seventeenth National Conference on Artificial Intelligence, AAAI'00*.
- Klapaftis, I. P., & Manandhar, S. (2005). Google & wordnet based word sense disambiguation. *Proceedings of the 22nd International Conference on Machine Learning (ICML05) Workshop on Learning and Extending Ontologies by using Machine Learning Methods*. Bonn, Germany.
- Leacock, C., Chodorow, M., & Miller, G. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24, 147–166.
- Lin, C., & Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. *Proceedings of 18th International Conference of Computational Linguistics, COLING'00*. Strasbourg, France.
- Lin, D., & Pantel, P. (1994). Concept Discovery from Text. *15th International Conference on Computational Linguistics, COLING'02*. Taipei, Taiwan.
- L.Vanderwende, G.Kacmarcik, H.Suzuki, & A.Menezes (2005). An automatically-created lexical resource. *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations, Vancouver, British Columbia, Canada, October 2005*.

- Magnini, B., & Cavaglià, G. (2000). Integrating subject field codes into wordnet. In *Proceedings of the Second International Conference on Language Resources and Evaluation LREC'2000*. Athens, Greece.
- McCarthy, D. (2001). *Lexical acquisition at the syntax-semantics interface: Diathesis alternations, subcategorization frames and selectional preferences*. Doctoral dissertation, University of Sussex.
- Mihalcea, R., & Moldovan, D. (2001). extended wordnet: Progress report. *NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL'2001)*. (pp. 95–100). Pittsburg, PA, USA.
- Mihalcea, R., & Moldovan, I. (1999a). An Automatic Method for Generating Sense Tagged Corpora. *Proceedings of the 16th National Conference on Artificial Intelligence*. AAAI Press.
- Mihalcea, R., & Moldovan, I. (1999b). An automatic method for generating sense tagged corpora. *Proceedings of the 16th National Conference on Artificial Intelligence*. AAAI Press.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., & Teng, R. (1991). Five papers on wordnet. *Special Issue of the International Journal of Lexicography*, 3, 235–312.
- Miller, G. A., Leacock, C., Teng, R., & Bunker, R. T. (1993). A semantic concordance. *Proceedings of the ARPA Workshop on Human Language Technology*.
- Moldovan, D., & Girju, R. (2000). Domain-specific knowledge acquisition and classification using wordnet. *Proceedings of FLAIRS-2000 conference*. Orlando, FL.
- Montoyo, A., Palomar, M., & Rigau, G. (2001a). Automatic generation of a coarse grained wordnet. *Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Pittsburg, USA.
- Montoyo, A., Palomar, M., & Rigau, G. (2001b). Wordnet enrichment with classification systems. *Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Pittsburg, USA.

- Niles, I., & Pease, A. (2001). Towards a standard upper ontology. *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*. Ogunquit, Maine.
- Padró, L. (1996). Pos tagging using relaxation labelling.
- Patwardhan, S., & Pedersen, T. (2003). *The cpan wordnet::similarity package* (Technical Report). <http://search.cpan.org/author/SID/WordNet-Similarity-0.03/>.
- Rigau, G., Atserias, J., & Agirre, E. (1997). Combining unsupervised lexical knowledge methods for word sense disambiguation. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. Joint ACL/EACL* (pp. 48–55). Madrid, Spain.
- Rigau, G., Magnini, B., Agirre, E., Vossen, P., & Carroll, J. (2002). Meaning: A roadmap to knowledge technologies. *Proceedings of COLING Workshop A Roadmap for Computational Linguistics*. Taipei, Taiwan.
- Rigau, J. D. L. P. G. (2000). Mapping wordnets using structural information. *38th Annual Meeting of the Association for Computational Linguistics (ACL'2000)*. Hong Kong.
- Stamou, S., Ntoulas, A., Hoppenbrouwers, J., Saiz-Noeda, M., & Christodoulakis, D. (2002a). Euroterm: Extending the eurowordnet with domain-specific terminology using an expand model approach. *Proceedings of the 1st Global WordNet Association conference*. Mysore, India.
- Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., & Grigoriadou, M. (2002b). Balkanet: A multilingual semantic network for the balkan languages. *Proceedings of the 1st Global WordNet Association conference*.
- Tomuro, N. (1998). Semi-automatic induction of underspecified semantic classes. *Proceedings of the workshop on Lexical Semantics in Context: Corpus, Inference and Discourse at the 10th European Summer School in Logic, Language and Information (ESLLI-98)*. Saabruecken, Germany.

- Turcato, D., Popowich, F., Toole, J., Fass, D., Nicholson, D., & Tisher, G. (1998). Adapting a synonym database to specific domains. *Proceedings of ACL'2000 workshop on Recent Advances in NLP and IR*. Hong Kong, China.
- Vossen, P. (Ed.). (1998a). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.
- Vossen, P. (Ed.). (1998b). *Eurowordnet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers.
- Vossen, P. (2001). Extending, trimming and fusing wordnet for technical documents. *NAACL Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations' (NAACL'2001)*. Pittsburg, PA, USA.
- Vossen, P., Bloksma, L., Rodríguez, H., Climent, S., Roventini, A., Bertagna, F., & Alonge, A. (1997). *The eurowordnet base concepts and top-ontology* (Technical Report). Deliverable D017D034D036 EuroWordNet LE2-4003.