Noname manuscript No.
(will be inserted by the editor)

Web 2.0, Language Resources and standards to automatically build a multilingual Named Entity Lexicon

Antonio Toral · Sergio Ferrández · Monica Monachini · Rafael Muñoz

the date of receipt and acceptance should be inserted later

Abstract This paper proposes to advance in the current state-of-the-art of automatic Language Resource (LR) building by taking into consideration three elements: (i) the knowledge available in existing LRs, (ii) the vast amount of information available from the collaborative paradigm that has emerged from the Web 2.0 and (iii) the use of standards to improve interoperability.

We present a case study in which a set of LRs for different languages (WordNet for English and Spanish and Parole-Simple-Clips for Italian) are extended with Named Entities (NE) by exploiting Wikipedia and the aforementioned LRs. The practical result is a multilingual NE lexicon connected to these LRs and to two ontologies: SUMO and SIMPLE. Furthermore, the paper addresses an important problem which affects the Computational Linguistics area in the present, interoperability, by making use of the ISO LMF standard to encode this lexicon. The different steps of the procedure (mapping, disambiguation, extraction, NE identification and postprocessing) are comprehensively explained and evaluated. The resulting resource contains 974,567, 137,583 and 125,806 NEs for English, Spanish and Italian respectively. Finally, in order to check the usefulness of the constructed resource, we apply it into a state-of-the-art Question Answering system and evaluate its impact; the NE lexicon improves the system's accuracy by 28.1%. Compared to previous approaches to build NE repositories, the current proposal represents a step forward in terms of automation, language independence, amount of NEs acquired and richness of the information represented.

Keywords Language Resources · Named Entities · Web 2.0 · standards

Antonio Toral
NCLT, School of Computing, Dublin City University, Ireland

Monica Monachini
Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche, Pisa, Italy

Sergio Ferrández · Rafael Muñoz
Natural Language Processing and Information Systems Group, Department of Computing Languages and Systems, University of Alicante, PO BOX 99, E-03080 Alicante, Spain

1 Introduction

World knowledge is a requirement for dealing with the semantic level of natural languages. Conceptualisations of reality have occupied human beings since the Ancient Greeks, where the term Ontology (from the Greek ὄν, genitive ὄντος: *of being* (part. of εἶναι: *to be*) and -λογία: *science, study, theory*) was introduced by Aristotle (Aristotle, 1908). A long time later, at the end of the XX century, the first attempts to give *common sense* to computers by building Knowledge Bases (KBs) were initiated in the field of Artificial Intelligence. Examples of this are the CYC project (Lenat, 1998), MindNet (Richardson et al, 1998) and, more related to natural language, WordNet (Miller, 1995).

Computational Linguistics is an interdisciplinary field related to Artificial Intelligence that deals with human-level understanding and generation of natural languages. World knowledge is necessary for attaining truly intelligent computer systems. In the case of language, this knowledge is contained in Language Resources (LRs), and in fact, these play a central role in the field of Computational Linguistics as they are practically indispensable for carrying out any automatic understanding of language. The research community has therefore dedicated a lot of effort to the manual construction of LRs during the last two decades.

In spite of the amount of work devoted to LRs, which has led to the availability of robust and high coverage LRs, some types of linguistic information are not exhaustively covered in these resources. Two paradigmatic examples are those of Named Entities (NEs)[1] and domain-specific terms. It is clear that the manual population and maintenance of these two kinds of terms into LRs would be unfeasible, as the amount of terms involved is huge and their nature, especially that of NEs, is much more volatile than that of the terms that make up the core of traditional LRs (common nouns, adjectives, verbs and adverbs). This is related with the following assertion: "building a proper noun ontology is more difficult than building a common noun ontology as the set of proper nouns grows more rapidly" (Mann, 2002). The problem is then that a proper noun resource should be constantly updated. Keeping with this, (Philpot et al, 2005) states that "the need for machine-assisted ontology construction is stronger than ever" because "humans cannot manually structure the available knowledge at the same pace as it becomes available". Hence, in order to fill this gap, automatic procedures are needed. The so called *knowledge acquisition bottleneck* is a recognised issue within the Natural Language Processing (NLP) community.

In order to clarify this issue, let us take a look at the state of NEs in WordNet -the most widely used English LR nowadays-. From version 2.1., this LR explicitly distinguishes between common nouns (called classes) and proper nouns (called instances) (Miller and Hristea, 2006). While WordNet's

[1] By Named Entities we refer in this paper to entities belonging to several semantic types (e.g. person, location, organisation) which take the form of proper nouns.

coverage of open domain common nouns is quite high, it contains very few proper nouns (only 7,669 synsets are tagged as instances in WordNet 2.1).

Following with NEs, most of the research done up to now relates directly to their recognition and classification in text according to small predefined sets of categories, such as the four category set (person, organisation, location, miscellaneous) of CoNLL (Tjong Kim Sang, 2002). With regards to NE resources, even if mature repositories of geographical NEs (also called gazetteers) do exist (e.g. geonames[2]), there is a lack of more general resources. However, the availability of general LRs with NEs could be very useful for NLP tasks; (Mann, 2002) shows how the use of a proper noun ontology, even if the ontology used has a low coverage, improves the precision of a Question Answering (QA) system. Moreover, this kind of resources could play a crucial role in NE Recognition systems that consider an extended hierarchy of entity types like that proposed in (Sekine et al, 2002).

Let us clarify the role that a NE rich LR could play in NLP by presenting a QA example. Consider the question 161 from the QA track at the 2006 edition of CLEF[3]: "Who is Fernando Henrique Cardoso?". This question would be easily answered if this person NE was present in a LR with semantic links to other entries, such as being an instance of "Brazilian", "politician", "president" or "minister".

1.1 Motivation and roadmap

Our present work aims at devising a generic methodology to extend existing LRs with NEs. The approach should be general enough so that it could be applied to different kinds of LRs and furthermore it should be language independent. NEs should not be only introduced in the LR but also linked to relevant existing entries by means of semantic relations. Moreover, the procedure should be fully automatic and produce a high quality final resource.

Because of the requirements posed to the task (high quality automatic extension of LRs with up-to-date NEs) we come up with two main ideas that will characterise our approach. The first is to exploit the information already present in LRs; these resources have been manually built by expert lexicographers and hence, the information encoded has high quality and can be used to support and guide their own extension. The second regards taking advantage of the so called *New Text* sources.

Up to now, research devoted to the automatic population of LRs has mostly focused on extracting the required information from two kinds of sources: Machine Readable Dictionaries (MRDs) and raw corpora. However, both present disadvantages. While MRDs are small in size and thus limit the quantity of information that can be extracted, corpora consist of unstructured text and therefore make it harder to extract valuable information.

[2] www.geonames.org

[3] www.clef-campaign.org

According to (Hearst, 1998), relations found in unrestricted text tend to be subjective judgements compared to the more established statements present in dictionaries and encyclopaedias. This is in line with the study conducted by (Wiebe et al, 2004). They analysed the Wall Street Journal Treebank Corpus and divided it into opinion and non opinion pieces . They discover that 70% of the sentences in opinion pieces are subjective and 30% are objective whereas in non opinion pieces, 44% of the sentences are subjective and only 56% are objective. Therefore, unless some post-process is carried out, these kind of textual sources are not appropriate for an automatic acquisition process. Wiebe and Riloff (Wiebe and Riloff, 2005) tackle this problem by creating subjective and objective sentence classifiers. Nevertheless, the results are far from being perfect; the best classifier, which is supervised, obtains 76% accuracy while the best unsupervised one achieves 73.8%.

Following with corpora based methods, they might, if no special treatment is applied, acquire the same instance with different lexical forms (Fleischman et al, 2003) (e.g. Bill Clinton and William Clinton) and therefore include them as different instances in the created resource.

However, new types of text -the so called *New Text*- have emerged as a consequence of the appearance of new forms of communication. By *New Text* we refer to "new types of text - dynamic, reactive, multilingual, with numerous cooperating or even adversarial authors and little or no editorial control" which have arisen due to "recent advances in publication and dissemination systems" (Karlgren, 2006). We are interested in using these kinds of sources because (i) they tend to have some degree of structure which facilitates the extraction of valuable information and (ii) they are dynamic and thus a sensible source to guarantee up-to-date information. Making use of these new kinds of information could present important advantages for Information Extraction compared to the aforementioned kinds of sources. New types of sources such as folksonomies (aka social tagging) and wikis contain semi-structured semantic information (categorisation tags, interlingual and multilingual links, attribute-value tables, etc) that is not only useful to recognise the elements to be extracted but also to disambiguate and normalise them. Besides, these sources are dynamic, thus change with time, and because they are collaboratively built, reflect language variety. The challenge consists of adapting state-of-the-art extraction techniques in order to derive the maximum benefit from these new kinds of sources.

One of these new kinds of text is known as wiki. Wikis can be defined as on-line texts that allow users to easily edit and change the contents. These characteristics make them an effective tool for collaborative authoring. The most widely known example of a wiki resource is Wikipedia, a multilingual encyclopaedia that follows the wiki philosophy. Wikipedia is an interesting textual source for the automatic creation of LRs because, being an encyclopaedia, it contains facts dealing with the entire range of human knowledge and, as it is developed by a large amount of people[4], therefore reflects the variations of

[4] On 2008/03/11 the English version has 9,141,485 registered users.

language and human thought. The quality of Wikipedia's content is comparable to traditional encyclopaedias, according to (Giles, 2005), which compares its English version to the Encyclopaedia Britannica, and to a study carried out by the WIND research institute for the Stern magazine[5][6], which confronts the German version to the Brockhaus On-line encyclopaedia.

Several aspects make this research different from previous work within lexical and semantic knowledge acquisition. Compared to research that relies on corpora, our research avoids problems due to subjective judgements[7] and inconsistencies due to calling instances in different manners whereas compared to research that uses MRDs, our method is not limited by the small size of the input resource.

Table 1 compares the relevant characteristics of corpora, MRDs and Wikipedia for their application to knowledge acquisition. Taking into account all the four features considered (structure, subjectivity, size and nature), Wikipedia emerges as the resource offering the best trade-off.

Table 1 Comparison of corpora, MRDs and Wikipedia

	corpora	**MRDs**	**Wikipedia**
structure	none	high	medium
subjectivity	high	low	low
size	big	small	big
nature	static	static	dynamic

Apart from the knowledge bottleneck, another important problem of the field has to do with interoperability. The lack of long-term planning has led to LRs in different formats (often incompatible), aimed to specific subfields. It is only in the last years that the community has realised about this problem. Several actions are being taking nowadays to address it though, including to mention but a few:

- The establishment in 2002 of a technical subcommittee in ISO, TC37/SC4[8], devoted to the creation of standards for LRs in order to maximise their applicability.
- Research efforts to create linked resources, examples are the Global WordNet Association[9], constituted in 2000, and the Meaning project[10].

[5] `http://www.stern.de/media/pdf/wiki_test_750.jpg`

[6] `http://www.stern.de/computer-technik/internet/:stern-Test-Wikipedia-Brockhaus/604423.html?q=Brockhaus\%20wikipedia`

[7] Specifically, Wikipedia, being an encyclopaedia and having strong policies regarding neutrality, does not suffer from such problems

[8] http://www.tc37sc4.org

[9] `http://www.globalwordnet.org/`

[10] `http://www.lsi.upc.es/~nlp/meaning/` (2002–2004)

– The creation of an international conference devoted to LR interoperability (The International Conference on Global Interoperability for LRs (ICGL)[11]) whose first edition was celebrated in 2008.

An added value of our proposal is the use of standards in order to make both the procedures more generic and independent from the specific resource(s) used and to improve the interoperability and future sharing of different LRs. Concerning this matter, we will study the use the Lexical Markup Framework (LMF) -an ISO standard for LRs- as the representation format of the resulting NE resource. The aims of this format are to provide a common model for the creation and use of lexicons, to manage the exchange of data between these resources and to enable the merging of resources.

The rest of the paper is organised as follows. The following section discusses the start-of-the-art. Next, we describe the LRs used in the present research. After that we present our methodology. This is followed by a discussion of the experiments that have been carried out. Finally, we introduce an application to QA and present the conclusions.

2 Background

This section reports on the state-of-the-art and it is divided in three subsections. First, we present a survey on general lexical acquisition and automatic construction of Language Resources. This is followed by a more specific section on the acquisition of NEs and the construction of onomastica. Finally, the section is closed with a summary of the use of Web 2.0 sources, and more specifically Wikipedia, in NLP during the last years.

2.1 General lexical acquisition and enrichment of Language Resources

Research on automatic lexical automatic acquisition began in the 1980s and initially focused on acquiring lexical information from MRDs. During the next decade, due both to the availability of large corpora and NLP tools needed for their accurate processing (PoS taggers, chunkers, etc.) and to the drawbacks of MRDs, the emphasis shifted to corpus-based approaches. Recent years have seen what could be called "a quantitative evolution"; the increasing processing power of computers together with the availability of robust statistical NLP tools have led to research proposals where the reference corpus is the World Wide Web.

The ACQUILEX project (Acquisition of Lexical Knowledge for NLP Systems, 1989-1992) pioneered on the derivation of lexica from very incipient samples of MRDs. Relevant publications from this period include (Calzolari, 1992), (Nakamura and Nagao, 1988) and (Alshawi, 1987). A later work (Rigau,

[11] http://icgl.ctl.cityu.edu.hk

1998) presents a detailed proposal regarding the massive acquisition of lexical knowledge from monolingual and bilingual MRDs. Apart from designing a productive methodology to build and validate a multilingual KB, a software system (called SEISD) was implemented.

(Hearst, 1992) criticises the utilisation of MRDs in knowledge acquisition because of their fixed size and proposes the extraction of semantic knowledge from corpora by using lexical patterns. Six patterns are proposed together with a methodology to find new ones. The follow-up of ACQUILEX, ACQUILEX-II (1993-1995), made considerable use of corpora as a further source of data for the semi-automatic construction of lexical resources. SPARKLE (1995-1996) demonstrated the important role of shallow parsing for acquiring several types of linguistic information such as subcategorisation, argument structure or selectional preferences. MEANING (2002-2005)(Atserias et al, 2004) acquired EuroWordNet-based information from corpora to support Word Sense Disambiguation. (Snow et al, 2006) [12] extends WordNet with up to 400,000 new synsets by applying a semantic taxonomy induction algorithm that exploits heterogeneous evidence.

(Agichtein and Gravano, 2000) addresses the scalability problem and proposes an efficient method when dealing with large corpora. (Etzioni et al, 2008) introduces *Open Information Extraction*, an extraction paradigm designed for large corpora in which the system makes a single pass and extracts tuples without any human input. The authors also present TextRunner, an implementation of this paradigm.

2.2 Onomastica acquisition and creation

This section presents an overview of research work regarding the creation and acquisition of onomastica, i.e. dictionaries of proper nouns. The most relevant approaches found in the literature follow.

(Sheremetyeva et al, 1998) presents the structure of a multilingual onomasticon made up of a set of monolingual onomastica cross–referenced by translation links. The entries are organised in a hierarchy made up of 45 semantic categories. A semi-automatic population procedure is proposed, which is supported by an acquisition and administration interface.

Prolexbase, a multilingual database of proper nouns, was created within the Prolex project (Tran et al, 2004) (Krstev et al, 2005) (Maurel, 2008). It is based on an ontology which has four layers (instances, linguistic, conceptual and meta-conceptual) and several relations (synonymy, meronymy, antonomasia, etc). Entries are linked to EuroWordNet's Inter-Lingual Index. The population of Prolex is done manually. It contains mainly French proper nouns, 75,368 lemmas. There are also translations for Serbian and German (13,000 entries).

(Mann, 2002) creates a proper noun ontology from newswire text. The proposal consists of extracting phrases from a 1 gigabyte corpus by applying

[12] `http://ai.stanford.edu/~rion/swn/`

a Part-of-Speech pattern (a common noun followed by a proper noun). This allows the author to gather 113,000 different proper nouns and to reach a precision of 60% (84% for proper nouns referring to people and 47% for the rest). The author also points out that the employed methodology is problematic with polysemous words and that it is not straight-forward to integrate the proper noun ontology created with the WordNet taxonomy of nouns.

(Fleischman et al, 2003) extracts concept-instance relations from 15 gigabytes of newspaper text by using two Part-of-Speech patterns (common nouns followed by a proper noun and appositions). Machine Learning techniques are applied to increase the precision of the extracted info. 500,000 unique instances (Bill Clinton and William Clinton are considered as two different instances) are extracted. An evaluation over 100 concept-instance items is carried out, achieving a precision of 93%.

(Sundheim et al, 2006) studies the linkage of a gazetteer to WordNet. The paper proposes to incorporate the instances of a geographic nature from WordNet into the Integrated Gazetteer Database (IGDB). This is justified by the fact that both resources contain complementary information.

(De Loupy et al, 2004) proposes to use WordNet as a proper noun thesaurus for a QA system by enriching it with 130,675 proper nouns. These nouns are extracted from several knowledge bases (the authors do not specify which) and from the Internet. 55 types of entries are enriched with proper nouns. However, not all of them seem to contain proper nouns (e.g. "professions" contains "Academic teacher", "political titles" contains "1st secretary"). The methodology followed to build this thesaurus is not mentioned, which leads us to think that both the acquisition of proper nouns and their insertion in the correspondent synsets are carried out manually.

REPENTINO (*REPositório para reconhecimento de ENTidades com NOme*) (Sarmento et al, 2006) is a repository of monolingual (Portuguese) NEs. This resource contains 450,129 entities, which are organised according to a taxonomy made up of several top categories (abstract, art and media, nature, event, legal, localisation, organisation, product, being and substance) which in turn are subdivided into subcategories. The NEs are extracted from several corpora and web sources by using semi-automated methods. Details about the amount of NEs extracted from each source per category can be found at `http://poloclup.linguateca.pt/cgi-bin/repentino/fontes.pl`. There is a web interface [13] that allow users to both browse the repository and to suggest new NEs to be added.

2.3 Wikipedia and NLP

In the last few years there has been a growing academic interest for Web 2.0 collaborative resources and among them especially for Wikipedia[14]. This is

[13] `http://www.linguateca.pt/REPENTINO/`

[14] `http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies#Over_time`

particularly true for the area of Computational Linguistics, which perceives Wikipedia as a new LR of huge dimensions.

Wikipedia is an on-line encyclopedia which is constantly built in a collaborative way by a huge amount of volunteers. It has versions for more than 200 languages. Wikipedia contains several elements which make it an interesting potential source for Computational Linguistics. We briefly outline the main elements of its structure:

- Pages. The page is the main element of Wikipedia. It represents the concept of article or encyclopaedic entry.
- Redirects, can be associated to pages. They represent orthographic variants of entry titles.
- Categories, to which pages can be associated. The categories form a taxonomy; a category can have one or several subcategories and belongs to a supercategory.
- Intralingual links. They connect two pages that belong to the same language.
- Interlingual links. They connect equivalent pages that belong to different languages.

In this field several events in which Wikipedia has a central role have been organised lately including the evaluation tasks WiQA[15] and GikiCLEF[16] and the workshops NEW TEXT[17], WikiAI08[18], and The People's Web Meets NLP[19].

The community has also developed tools that allow researchers to access the information encoded in Wikipedia and other similar resources. Examples of these are JWPL[20] and JWKTL[21], APIs that allow to access the information contained in Wikipedia and Wiktionary respectively (Zesch et al, 2008).

Wikipedia has been exploited for a wide range of tasks such as Monolingual (Ahn et al, 2005; Jijkoun et al, 2005; Buscaldi and Rosso, 2006) and multilingual (Ferrández et al, 2007b) QA, Semantic relatedness (Ponzetto and Strube, 2007; Gabrilovich and Markovitch, 2007; Milne and Witten, 2008), Information Extraction (Wu et al, 2008), NE Disambiguation (Bunescu and Pasca, 2006) or NE Recognition (Nothman et al, 2009). Furthermore, several researchers have used it to build LRs. (Gregorowicz and Kramer, 2006) mine a term-concept network from Wikipedia. (Suchanek et al, 2007) introduces an ontology automatically derived from Wikipedia and WordNet. (Auer et al, 2008) extracts structured information from Wikipedia and makes in available on the Web. (Pedro et al, 2008) extracts a medical ontology. (Milne et al, 2006) mines a thesaurus for the agriculture domain. (Medelyan and Legg, 2008) integrates

[15] http://ilps.science.uva.nl/WiQA/
[16] http://www.linguateca.pt/GikiCLEF/
[17] http://www.sics.se/jussi/newtext
[18] http://lit.csci.unt.edu/~wikiai08/index.php/Main_Page
[19] http://www.ukp.tu-darmstadt.de/acl-ijcnlp-2009-workshop
[20] http://www.ukp.tu-darmstadt.de/software/jwpl/
[21] http://www.ukp.tu-darmstadt.de/software/jwktl/

Cyc and Wikipedia. (Jones et al, 2008) builds a domain-specific multilingual dictionary by extracting the entries from a Wikipedia category, which is then used to customise a Machine Translation system. (Ruiz-Casado and Castells, 2006) extracts relations between entries of Wikipedia which are added to their corresponding WordNet entries.

3 Language Resources

This section introduces the LRs used in the present research for the different languages covered (English, Italian and Spanish). The LRs are WordNet (for English), EuroWordNet (for Italian and Spanish) and PAROLE-SIMPLE-CLIPS (for Italian). The following subsections briefly describe each of these LRs.

3.1 WordNet

WordNet is an on-line lexical database for English developed at the University of Princeton that contains nouns, verbs, adjectives and adverbs organised into sets of synonyms - called synsets- and contains several types of semantic relations among its nodes (Miller, 1995). It is manually developed by a team of linguists and its design is inspired by psycholinguistic theories of human lexical memory. This resource is widely used within the NLP community. In fact, it has become the *de facto* standard for several NLP tasks such as Word Sense Disambiguation.

The version of WordNet used in this research, 2.1., is made up of 117,597 synsets (81,426 nouns, 13,650 verbs, 18,877 adjectives and 3,644 adverbs) and 155,327 variants (117,097 nouns, 11,488 verbs, 22,141 adjectives and 4,601 adverbs).

3.2 EuroWordNet

EuroWordNet (EWN) (Vossen, 1998) is a project funded by the European Union with the aim of developing a multilingual database of inter-connected wordnets for several European languages. This project is inspired by WordNet but introduces important improvements. EWN contains new types of relationships, including some across parts of speech. Moreover, EWN is a multilingual resource; a module called inter-lingual-index (ILI) links "equivalent" synsets in the various wordnets by using as a pivot the synsets of WordNet 1.5. EWN introduces a set of 1,024 top concepts common to all the languages and a language independent Top Ontology built from 63 very abstract of these top concepts.

In this research we use two wordnets that belong to the EWN model: the Italian and Spanish wordnets.

3.2.1 Italian WordNet

The Italian WordNet (IWN) (Alonge et al, 1999) was built from different Italian lexical and corpora sources, such as the Italian Machine Dictionary, the Italian Reference Corpus and the PAROLE lexicon. IWN was originally created in the framework of the EWN project and then further extended in the Italian national project "Integrated System for the Automatic Language Processing" (SI-TAL).

In its current status, IWN provides the semantic description for around 67,000 Italian word senses (9,096 verbs, 32,099 common nouns, 3,450 proper nouns, 4,356 adjectives, and 513 adverbs, either single or multi-word units), which are clustered in approximately 50,000 synsets. IWN employs the same set of semantic relations used in EWN and there are currently 117,068 instances of language-internal relations.

3.2.2 Spanish WordNet

The Spanish WordNet (Verdejo, 1999) was built within the EWN project by a research team belonging to three universities: UNED, University of Barcelona and Technical University of Catalonia. It was afterwards extended, enriched and mapped to WordNet 1.6. The version used in this research[22] contains 30,485 synsets, 52,515 variants, 73,665 language internal relations and 28,283 equivalence relations to the ILI.

3.3 PAROLE-SIMPLE-CLIPS

PAROLE-SIMPLE-CLIPS (PSC) is an Italian computational lexicon which has been developed in the framework of three different projects. The first two, PAROLE (Ruimy et al, 1998) and SIMPLE (Lenci et al, 2000), were funded by the European Union and were devoted to the research and development of wide-coverage, multi-purpose and harmonised computational lexicons for twelve European languages. While PAROLE dealt with the morphological and syntactic layers, SIMPLE added a semantic layer to the PAROLE data. Finally, CLIPS (Ruimy et al, 2002) was an ulterior Italian national project where the Italian lexicon was enlarged and refined.

The semantic layer of PSC, the relevant one for the current research, contains about 55,000 semantic units (i.e. senses) organised in an ontology made up of 153 semantic types (i.e. ontology nodes).

From a theoretical point of view, the linguistic background of PSC is based on the Generative Lexicon theory (Pustejovsky, 1991). In this theory, the sense is viewed as a complex bundle of orthogonal dimensions that express the multidimensionality of word meaning. The most important component for representing the lexical semantics of a word sense is the qualia structure which

[22] available for research at urlhttp://www.lsi.upc.edu/ nlp

consists of four qualia roles (formal, constitutive, agentive and telic). Each qualia role can be considered as an independent element or dimension of the vocabulary for semantic description. The qualia structure enables us to express different or orthogonal aspects of word sense whereas a one-dimensional inheritance can only capture standard hyperonymic relations.

3.4 Mapping between PSC and IWN

Although PSC and IWN follow different lexical models, they also present compatible aspects (as a matter of fact, the ontologies of SIMPLE and EWN are compatible). Linking both resources offers the end-user more exhaustive lexical information combining features offered by the two lexical models. It provides not only reciprocal enhancements but also a validation of the two resources. Moreover the linking presents a multilingual vocation; on one hand IWN is linked to wordnets for other languages by using the ILI, on the other hand PSC shares the theoretical model, the representation language, the building methodology and a set of core entries with 11 other European lexicons. Regarding the current status of this linking, 72.37% of word senses about concrete entities and 69.59% of word senses about abstract entities and events have been mapped (Roventini et al, 2007) (Roventini and Ruimy, 2008).

4 Procedure

In this section we explain thoroughly the procedure followed to derive a lexicon of NEs from existing LRs and Wikipedia. It consists of several sequential phases which we will refer to as: mapping, disambiguation, extraction, NE identification and post-processing. A graphic depicting the overall process is presented in figure 1. As previously stated, the approach followed takes advantage of information already present in LRs and exploits the semi structured nature of *New Text*.

Our method maps the noun is-a hierarchy of LRs to Wikipedia categories, disambiguates eventual ambiguous mappings, extracts the articles present in the latter and identifies from them which are NEs. Several pieces of information from the NEs such as written variants, definitions, etc. are introduced into a NE lexicon. In a post-processing phase, (i) additional NEs are extracted exploiting the interlingual links of Wikipedia and (ii) the extracted NEs are linked to ontologies.

The following subsections deal with each phase. Afterwards we present the structure of the resulting NE lexicon.

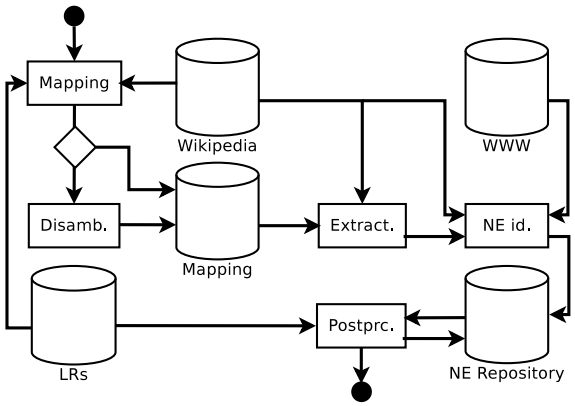

Fig. 1 Diagram of the procedure

4.1 Mapping

In this first step the instantiable nouns[23] present in the LRs are mapped to Wikipedia categories. These mappings are obtained by comparing the lemmas of the nouns to those of the categories. In order to do this, the categories of Wikipedia are lemmatised with Freeling 2.0 (Atserias et al, 2004), as this tool provides PoS-tagging machinery for the different languages considered (English, Spanish and Italian).

Once we have the subset of nouns that are instantiable and the categories have been lemmatised, we map the LR nouns to Wikipedia categories by matching their lemmas. For example, the noun "country" would be mapped to the category "Countries" as the PoS-tagger would obtain the same lemma for both words.

4.2 Disambiguation

Once the nouns of the LR have been mapped to categories, a further mandatory step must be carried out for those nouns that are polysemous: the sense that corresponds to the mapped category should be identified. Other approaches have neglected this step, e.g. YAGO (Suchanek et al, 2007) chooses the most frequent sense as the correct one and, subsequently, errors produced by this assumption are manually corrected.

We have devised two different approaches to do this automatically. The first looks for common instances in the hyponym trees of both the noun senses and the category, while the second performs text similarity between the definitions of the noun senses and the category. The following subsections present both approaches.

[23] the set of nouns that can be instantiated by means of a NE, e.g. "country" has instances such as "France".

4.2.1 Instances intersection

We hypothesise that instances could be useful to disambiguate WordNet polysemous words with respect to Wikipedia categories. E.g. the English word "obelisk" is mapped to the category "Obelisks". It has two senses in WordNet (1. stone pillar, 2. character used in printing). The first sense has one instance ("Washington Monument") while the second has none. In the Wikipedia category "Obelisks" we find the instance "Washington Monument". Thus, the sense chosen for the mapping would be the initial one.

As the taxonomy of Wikipedia is usually deeper than that of WordNet, we not only consider looking for instances in the mapped categories but also in their hyponyms (subcategories). However, the subcategory relation in the categories taxonomy of Wikipedia does not always follow the hyponymy relation[24]. Therefore, in order to exploit subcategories, we need to identify whether they are hyponyms or not. We propose to apply regular expression patterns which can hold both lexical and Part-of-Speech elements. If a subcategory matches a pattern then it is considered as a hyponym. From studying the category structure of Wikipedia, we come up with the following patterns for English (for each pattern we provide an example of matching subcategory for the category "Philosophers"):

- `^category " by|in|from|of "`, e.g. "philosophers of mind"
- `^ category " stubs" $`, e.g. "philosophers stubs"
- `^ (JJ|JJR|NN|NP)+ (CC(JJ|JJR|NN|NP)+)* " " category $`, e.g. "Spanish philosophers"

As an example, we show how the word philosopher (1. specialist in philosophy, 2. wise person who is calm and rational) is disambiguated with respect to the category "Philosophers". The first sense contains several instances such as "Averroes" while the second contains none. "Averroes" is not present in the mapped category but it is found in a subcategory that follows the hyponymy relation ("Philosophers" `->` "philosophers by nationality" `->` "Spanish philosophers").

Equivalent patterns have been also built for the other languages considered, i.e. Spanish:

- `^ category " por|de|del|en "`, e.g. "Filósofos de la Edad Antigua"
- `^ "Wikipedia:esbozo " category $`, e.g. "Wikipedia:Esbozo filósofos"
- `^ category " " (AQ[0-9A-Z]+ |N[0-9A-Z]+)+(CC|SP[0-9A-Z]+\ (AQ[0-9A-Z]+|N[0-9A-Z]+)+)* $`, e.g. "Filósofos árabes"

and for Italian:

- `^ category " per|di|del|dell'|della|delle|degli "`, e.g. "Filosofi del XX secolo"

[24] E.g. In the category "Philosophers" there are subcategories that follow the hyponymy relation (e.g. "Philosophers by country") but there are also others that do not (e.g. "Philosophy academics").

– `^ "stub " category $`, e.g. “stub Filosofi”
– `^ category " " (AQ[0-9A-Z]+ |N[0-9A-Z]+)+(CC|SP[0-9A-Z]+\ (AQ[0-9A-Z]+|N[0-9A-Z]+)+)* $`, e.g. “Filosofi atei”

4.2.2 Text similarity

The second disambiguation approach relies on the definitions of the mapped elements, it applies text similarity to disambiguate the correct sense of the polysemous nouns mapped to Wikipedia categories. For each such noun, it computes the similarity between the gloss of each of its senses and the abstract of the mapped category.

As there are different approaches to compute text similarity, we have decided to consider a set of representative methods in order to find out which works best for the current task:

– Semantic Vectors, a Latent Semantic Analysis like algorithm based on random projection (Widdows and Ferraro, 2008)[25]. It relies on Apache Lucene[26] for tokenisation and indexing in order to create a term document matrix. At that point, Semantic Vectors creates a WORDSPACE model by applying random projection. Semantic Vectors provides a class (CompareTerms) that calculates the similarity between two terms (which can be words or texts).
  For the current task we have gathered a corpus made up of WordNet glosses and Wikipedia abstracts. On one hand, it contains the glosses of all the synsets present in WordNet 2.1. On the other hand, it contains the abstracts of all the entries present in a Wikipedia dump obtained in January 2008. The final corpus has 1,292,447 terms.
– A Textual Entailment system (Ferrández et al, 2007a; Balahur et al, 2008) which implements several inferences aimed at solving entailment relations. On one hand, lexical inferences based on distance measures (Levenshtein, Smith-Waterman, etc.). On the other hand, semantic inferences focused on semantic distances between concepts (WordNet-based similarity measures, verb similarities according to relations encoded in VerbNet and VerbOcean and reasoning about NE correspondences between texts.
  For the application of the system to the current target task, we adapted it in order to manage bidirectional meaning relations. Linking WordNet glosses to Wikipedia categories is not a clear entailment phenomenon. It can occur that the gloss is implied by the category, the category is deducted by the gloss or the entailment appears in both directions. Therefore, to control these situations we opted for computing the average of the two system outputs regarding each unidirectional relation (as shown in equation 1).

[25] http://code.google.com/p/semanticvectors
[26] http://lucene.apache.org

$$BiSim(Gloss_i, Catg_j) = \frac{sim(Gloss_i \rightarrow Catg_j) + sim(Catg_j \rightarrow Gloss_i)}{2} \quad (1)$$

– A LR-based algorithm which applies Personalised PageRank to WordNet (Agirre and Soroa, 2009). The LR is represented as a graph; nodes represent concepts and dictionary words while relations among concepts are represented by undirected edges. Dictionary words are linked to the concepts associated to them by directed edges.
Given a pair of texts and a graph-based representation of a LR, this method has basically two steps: it first computes the personalised PageRank over the LR separately for each of the texts, producing a probability distribution over LR concepts. It then compares how similar these two discrete probability distributions are by encoding them as vectors and computing the cosine between the vectors.

4.3 Extraction

Once the mapping has been carried out, NEs can be extracted from the mapped categories. For each category mapped we extract all its subcategories which are hyponyms (see section 4.2.1). From the resulting set of categories (i.e. the mapped category plus all its hyponyms), we obtain the articles they contain and identify which are NEs, as explained in 4.4. Thus we obtain the set of articles which are NEs. From them we gather further relevant information such as their abstracts and their redirects (this information is explicitly available in a structured form from the Wikipedia database dumps and thus obtaining it is straight-forward). Finally, all this information is uploaded to the NE lexicon (see section 4.6).

Let us take the example of the mapped category "Countries". First, the procedure would obtain all the hyponym subcategories: "Fictional countries", "Countries by language", "Arabic-speaking countries", etc. Subsequently, all the articles from the resulting category set would be extracted: "Neverland", "Algeria", "Fictional country" etc. and only those being NEs are considered (in this example "Fictional country" would be discarded). From the articles that are NEs we gather other information; from "Neverland" we would get the redirects "Never Land" and "Never Never Land" and the abstract "Neverland (also spelled Never Land or expanded as Never Never Land) is a fictional world featured in the works of J. M. Barrie and those based on them".

4.4 NE identification

We have explored three different possibilities in order to identify which of the extracted articles are NEs. The first relies on a web search engine, the second on the content of Wikipedia entries while the third combines the two. All three

share a common aspect though; they exploit the capitalisation norms followed in some languages, i.e. that proper nouns begin with uppercase while common nouns begin with lowercase. A detailed explanation on each of them follows.

4.4.1 Web search

The article's title is searched in the World Wide Web by using a web search engine. The first 50 results where the title is found are returned and an algorithm calculates the number of times the article's title appears in the website's description (i) with all the words beginning with capital letters, (ii) with some words beginning with capital letters and (iii) with no word beginning with capital letters. Besides, a threshold is established in order to discard between articles being instances and non-instances according to the different models of capitalisation.

4.4.2 Wikipedia search

This approach also takes advantage of capitalisation norms, but instead of looking for entry occurrences in the World Wide Web, it looks for them in the body article of the entry, following (Bunescu and Pasca, 2006). The difference is that our method is language independent due to the use of Wikipedia's interlingual links. For a given Wikipedia article title, whatever its language, we obtain its equivalents in a set of ten languages that follow the aforementioned capitalisation rules (Catalan, Dutch, English, French, Italian, Norwegian, Portuguese, Romanian, Spanish and Swedish). Apart from the language independence, considering the entry in ten languages presents another important advantage: the text size where we look for occurrences of the entry is bigger, hence the results are more representative.

In order to obtain the entry title for each of these languages we use the interlingual links of Wikipedia that connect equivalent entries in different languages (translations). We look for occurrences of the article title in the body of each translation and compute the percentage of times it begins with uppercase. Finally, as in the previous approach, if the percentage is higher than a threshold then the article title is classified as a NE.

From a technical point of view it is worth mentioning that the body articles of Wikipedia are not in plain text but in the mediawiki mark-up format and thus are not directly processable by text tools. In order to carry out the current procedure, we first transformed the body articles into plain text by using two perl modules, Text::MediawikiFormat[27] and html2text[28].

All in all, this approach presents two advantages over the previous one:

– Language independence. Whatever the language we apply these procedures to, we can obtain the Wikipedia entry titles for languages which follow the aforementioned capitalisation norms.

[27] `http://search.cpan.org/~dprice/Text-MediawikiFormat-0.05/lib/Text/MediawikiFormat.pm`

[28] `http://search.cpan.org/~awrigley/html2text-0.003/html2text.pl`

– Avoidance of sense variation. A problem of the previous method is related to the fact that some nouns have senses in which they are instances and others in which they are classes. If an extracted entry is a NE but it also has a class sense the method could fail to classify it as a NE as in the Web we would find both senses. E.g. the Wikipedia entry "Children's Machine" is a NE referring to a laptop developed by the OLPC (acronym of One Laptop Per Child). However, this term can also be found in the string "The children's machine", the title of book from Seymour Papert in which "children" and "machine" are classes. With the new method we look for "Children's Machine" in the body of its article and so it is really unexpected to find this string referring to the book.

4.4.3 Combining Wikipedia and the Web

While searching for occurrences of the title in Wikipedia avoids noise due to eventual sense variation, the web method presents an important advantage: the amount of text available is considerably bigger and therefore more occurrences can be found.

We conclude then that the advantages of these two approaches could be combined by extracting salient terms from the entry body text in Wikipedia (the tf-idf measure is applied) and then searching in the Web pages where the entry title and these terms appear. Therefore, our combination method consists of the web search method refined with significant terms from the Wikipedia entry.

Following the example presented for the previous method (the entry "Children's Machine"), of the first ten results from Google, six correspond to the computer and the remaining four to Papert's book. However, if we extract the two more significant terms from the body text of the Wikipedia entry according to tf-idf ("OLPC" and "$100 laptop"), and then search the three terms (the title plus these two terms) in Google, then all of the first ten results correspond to the computer.

4.5 Postprocessing

The aim of the postprocessing phase is to improve and increment the information extracted and introduced into the NE lexicon. Two different actions are carried out: (i) introducing additional NEs and (ii) linking NEs to ontologies.

Additional NEs are introduced into the lexicon by exploiting Wikipedia multilingual links. If a NE has been extracted for language a but its equivalent in language b has not, we gather the NE for language b and add it to the lexicon.

Links to ontologies present in some of the LRs could be exploited to connect the extracted NEs to them. On one hand, the English WordNet has been linked to the SUMO ontology (Niles and Pease, 2003). On the other hand, the Italian PSC contains an ontology itself. Therefore, the extracted NEs are connected to these two ontologies.

4.6 The Named Entity Lexicon

In order to make the procedures independent of specific LRs and to facilitate interoperability with other LRs, we provide a standard-compliant output format. The elements that are part of this output are mainly NEs, orthographic variants of these NEs and classes to which these NEs belong (by means of "instance of" relations). Due to the fact that this data could be naturally represented by means of a LR and because we build with it a LR we have decided to follow the Lexicon Markup Framework (LMF), an ISO standard for the representation of lexicons (Francopoulo et al, 2008) (ISO 24613, 2008), in order to encode the output.

A description of the LMF elements we have considered and their role in our lexicon follows. The "Lexicon" element holds each NE monolingual dictionary. "LexicalEntry" acts as a container for all the information that regards each NE and contains two child elements. The first, "Lemma", contains the lemma of the NE and its orthographic variant/s (by making use of "FormRepresentation"). The second, "Sense", holds semantic information which can be one of two types: (i) relations to other lexical entries ("SenseRelation") and (ii) links to other resources ("MonolingualExternalRef").

Furthermore, we make use of the NLP multilingual notations extension of LMF to create a multilingual lexicon where NEs for different languages might be related by means of interlingual links. The element of LMF employed for this purpose is the "SenseAxis"; it represents the relationships between different closely related senses in different languages (each of these senses is contained in a "SenseAxisElements"). This element groups together monolingual senses that correspond to one another. Within a "SenseAxis" element we use the "InterlingualExternalRef" in order to link its elements to ontologies.

As for the output structure, we have designed a NE lexicon as a database whose structure is compliant (isomorphic) with LMF. Figure 2 presents the ER diagram of this database.

As an example of the information extracted, we provide the LMF compliant XML notation and the corresponding database entries for a NE for English and Italian (see appendix A).

5 Results and Discussion

This section introduces the experimental setting and presents the evaluation and subsequent discussion of the different phases of the methodology described in the previous section.

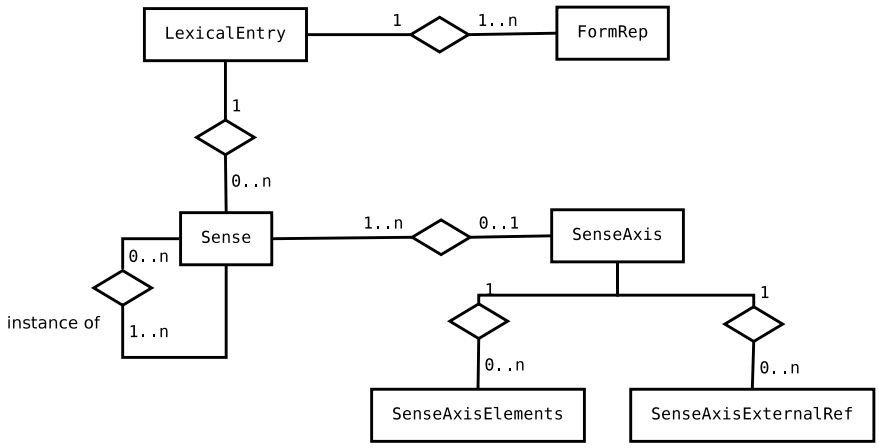

Fig. 2 ER diagram of the NE lexicon database

5.1 Data

In the current research we use database dumps of English, Italian and Spanish Wikipedia from January 2008[29]. From these dumps, we have used the page, pagelinks, categorylinks, text and abstract data. Table 2 shows the number of categories and articles for each of the Wikipedia dumps. Concerning LRs, we have used WordNet 2.1, Spanish WordNet 1.6 and PSC.

Table 2 Number of categories and articles in Wikipedia per language

Language	Number of categories	Number of articles
English	312,948	2,183,497
Italian	39,019	388,717
Spanish	45,888	305,366

Apart from the aforementioned Wikipedia dumps and LRs, some of the experiments rely on specific test data. In that case, datasets are described with the experiments they were created for.

5.2 Mapping

In order to map the English WordNet, we departed from its noun classes that contain instances (as we are interested in extending a LR with NEs we decided to consider a set of instantiable nouns; clearly if a noun is instantiated it is instantiable). Apart from this set of noun classes, following the inheritance principle of the hyponymy relation (i.e. if a noun class is instantiable, also its

[29] Downloaded from http://download.wikimedia.org

hyponyms are), we also consider the noun classes that are hyponyms of this set.

For Spanish, as the EuroWordNet model does not include a specific type of relation for instantiation, rather than begin with the Spanish WordNet we are forced to start with the English one. From it we extract the nouns that contain instances and their hyponyms. We obtain the equivalent synset offsets in WordNet 1.6 by using the mapping sets between WordNet versions provided by (Daudé et al, 2003).

From the set of 15,906 synsets with instances (and its hyponyms) present in WordNet 2.1, 2,140 cannot be mapped to WordNet 1.6 because no mapping is found, 13,640 have 1-to-1 mappings and 126 1-to-n ($n > 1$) mappings. In the last case the mapping with the highest confidence score is preserved. We end up with a set of 13,278 instantiable synsets of WordNet 1.6. From these, exploiting the ILI, we obtain the corresponding 15,094 variants of the Spanish WordNet. 2,966 of them are proper nouns (e.g. "África", "Nuevo Testamento") and therefore discarded. This leads us to a set of 12,128 variants. From these, 7,739 are monosemous and the remaining 4,389 polysemous. Subsequently, these variants can be mapped to categories of the Spanish Wikipedia.

For Italian, we proceed in an analogous manner. We departed from the set of instantiable nouns of the English WordNet. From this we obtained the equivalent synsets in WordNet 1.5, which is connected to the Italian WordNet through the ILI. Finally, from the Italian WordNet, we obtained the equivalent entries of PSC. From the set of 15,906 synsets with instances (and its hyponyms) present in WordNet 2.1, 2,806 cannot be mapped to WordNet 1.5 because no mapping is found, 12,946 have 1-to-1 mappings and the remaining 154 1-to-n ($n > 1$) mappings. This leads to a set of 13,183 synsets of WordNet 1.5. Following the ILI we gather 12,488 corresponding synsets from ItalWordNet. Finally, exploiting the ItalWordNet-PSC mapping, we obtain 10,498 variants of PSC. After discarding instances we have 6,977 monosemous and 3,067 polysemous nouns.

Table 3 shows the mapping results obtained for English, Spanish and Italian. The table presents two types of results (columns *nh* without considering hyponyms of the initial synsets and columns *h* considering them).

Table 3 Mapping for English, Spanish and Italian

		English		Spanish		Italian	
		nh	h	nh	h	nh	h
Monosemous	Total	1,012	14,855	627	7,739	777	6,977
	Mapped	557	2,860	195	446	159	529
	Percentage	55.03	19.25	31.10	5.76	20.46	7.58
Polysemous	Total	628	6,903	473	4,389	386	3,067
	Mapped	282	1,429	103	490	103	358
	Percentage	44.90	20.70	21.77	11.16	26.68	11.67

The amount of mapped nouns is notably lower for Spanish and Italian than for English both for monosemous and polysemous nouns, this is expected because both the number of total monosemous and polysemous nouns and the number of Wikipedia categories (45,796 and 39,019 vs. 312,941) are substantially lower for these two languages.

The percentages are considerably lower when considering hyponyms. This is expected as in doing so we map very specific nouns from deep nodes of the LR taxonomy for which is less probable that a correspondent Wikipedia category exists (e.g. it is expected to find a category for the noun "sword" but more unlikely for a more specific hyponym such as "rapier"). However, considering hyponyms boosts the total amount of nouns mapped in all cases.

5.3 Mapping Analysis

We present an analysis of the mapping results for English (column *nh* in table 3). Table 4 shows the percentages of monosemous words, polysemous words and synsets that get mapped to Wikipedia categories for three different dumps from April 2007, November 2007 and January 2008. As it can be seen, the continuous growth of Wikipedia allows us to increase the mapping percentage. 57.44% of the synsets were mapped to the April 2007 dump. This percentage increases to 60.02% for the November 2007 dump and 65.39% for the January 2008 dump (the one we are currently working with).

Table 4 Mapping percentages for different Wikipedia dumps

		Wikipedia dump date		
		200704	200711	200801
Monosem. Nouns	Total		1012	
	Mapped	491	509	557
	Percent.	48.51%	50.29%	55.03%
Polysemous Nouns	Total		628	
	Mapped	249	265	282
	Percent.	39.64%	42.19%	44.90%
Synsets	Total		893	
	Mapped	513	536	584
	Percent.	57.44%	60.02%	65.39%

In order to get a better understanding from the mapping procedure, we have manually analysed a randomly selected set of WordNet classes which do not get mapped to any Wikipedia category. In most of the cases (75%), although there is not a matching category, there is a matching article in Wikipedia to which the class could be mapped. E.g. "oracle" could be mapped to the article "Oracle". In 13% of the cases there is neither a matching category nor a matching article (e.g. "formal garden"). 10% of times there is a matching category but the class is not mapped to it due to a PoS tagger error. E.g.

the class “aquarium” is not mapped to the category “Aquaria” because the tagger fails to obtain “aquarium” as the lemma. The remaining 2% is due to having the class and matching category in different English variants. E.g. the class “railroad tunnel” (British) should be mapped to the category “railway tunnels” (American) but is not mapped as their lemmas do not match.

5.4 Disambiguation

We have evaluated the two automatic methods (instances intersection and semantic similarity, described in 4.2.1 and 4.2.2 respectively) for English.

In order to evaluate these methods we took a set of 207 mappings of polysemous words from WordNet to Wikipedia categories. For these words we manually selected the sense/s that correspond(s) to the mapped category. In most of the cases (154, 74,4%) there is a one to one correspondence. For 37 (17,9%) mappings, more than one sense corresponds to the mapped category, this usually occurs because the WordNet senses tend to be finer-grained than the Wikipedia categories. Concerning the remaining 16 (7,7%) mappings, no sense corresponds to the mapped category. Additional information is provided for the semantic similarity method; the glosses of the nouns and the abstracts of the categories. This evaluation set is publicly available at `http://computing.dcu.ie/~atoral/#Resources`

5.4.1 Instance Intersection

This algorithm disambiguates 39% of the words. This low recall, which is due to the low number of instances present in WordNet, is compensated by a very high precision. In fact, all the disambiguated entries were correct. We analysed the reasons why 61% of the words were not disambiguated. There are two main causes:

- One of the senses from WordNet corresponds to the category but no common instance is found. This happens for 78% of the cases. For 74% of the words there is simply no common instance in both resources. For the remaining 4% a common instance does exist but it is in a subcategory that although being a hyponym of the mapped category, the hyponymy patterns are not able to identify as such. E.g. “Colosseum, Amphitheatrum Flavium” is an instance of the second sense of “amphitheater”, which is mapped to the category “Amphitheaters”. “Colosseum” is present in the category “Roman amphitheatre buildings” which is a subcategory of “Amphitheaters”. However, the aforementioned patterns do not identify “Roman amphitheatre buildings” as a hyponym of “amphitheater”.
- No sense from WordNet corresponds to the category or the category has been changed. This occurs for the remaining 22%. An example of no sense corresponding to the mapped category happens for the word “assemblage” which has four senses: “a group of persons together in one place”, “a system

of components assembled together for a particular purpose", "the social act of assembling" and "several things grouped together or considered as a whole". The mapped category, "Assemblage", is "for assemblage artists". As an example of a category change, the word "college" is mapped to the category "Colleges" but it has been moved to "Universities and colleges". Obviously we cannot map "college" to "College and Universities" as by doing so we would end up with instances of universities under the class college.

5.4.2 Semantic similarity

We have evaluated the systems presented in section 4.2.2 together with two baselines:

- First Sense, it follows the assumption that senses in WordNet are ordered according to their usage predominance (i.e. the first sense is the most general). First Sense chooses always the first sense of WordNet as being the correspondent to the mapped Wikipedia category.
- Word Overlap, calculates similarity between two texts by counting the number of overlapping words. In order to do this we have used the software package Text::Similarity[30].

Hypothesising that the different nature of the considered systems might make their results complementary we have explored also with their combination; we present three strategies:

- Voting. For each mapping it ranks senses according to the number of times they are returned by the different systems which are combined. Finally, it outputs the first ranked sense. Voting returns more than one sense if two or more senses are ranked first with the same score.
- Unsupervised combination. Within this combination, the methods taken into account have the same relevance computing a simple average function among the outputs of the considered methods.
- Supervised combination. The whole set of inferences carried out by the Textual Entailment system together with the scores returned by the other methods are computed as features for a machine learning algorithm. We have used the BayesNet implementation provided by Weka (Witten and Frank, 2005), and we obtained the 10-fold cross validation results over our gold standard corpus.

Table 5 presents the scores obtained by the different systems, the baselines and the combinations[31].

The first element that appears is the high score obtained by the First Sense baseline (64.7%). In fact, leaving aside supervision, only one system is able to reach its score, Textual Entailment. Regarding combinations, the three

[30] http://text-similarity.sourceforge.net

[31] The combination strategies use Textual Entailment, Personalised PageRank and Word Overlap

Table 5 Semantic Similarity Results

Run	Accuracy
Baseline 1st sense	64.7%
Baseline Word Overlap	62.7%
Semantic Vectors	54.1%
Personalised PageRank	64.3%
Textual Entailment	64.7%
Voting	68%
Unsupervised combination	65.7%
Supervised combination	77.11%

of them outperform the best system; the improvement is slight both for the unsupervised (65.7% vs. 64.7%) and for the voting approaches (68% vs. 64.7%) while it is more significant for the supervised combination (77.11% vs. 64.7%).

5.5 Extraction

We have extracted NEs for the mapped nouns (see table 3) for each language. Table 6 provides quantitative data about the NEs extracted. We not only show the number of NEs which are added to the lexicon but also the amount of orthographic variants (written forms) of these NEs and the number of instance relations extracted that are linked to the LRs used.

Table 6 Extracted NEs

	English	Spanish	Italian
NEs	948,410	99,330	78,638
Written forms	1,541,993	128,796	104,745
Instance relations	1,366,899	128,796	139,190

The number of NEs extracted for Spanish and Italian is notably lower than the number of NEs for English. This result was expected because both the number of pages in Wikipedia (305,000 and 388,000 vs. 2,100,000) and the number of mapped categories (see section 5.2) are significantly lower.

Table 7 provides results about the nature of the NEs for English added to the lexicon. It shows the number of instances added according to the different noun lexicographic files of WordNet. For each lexicographic file where a substantial amount of instances is added we include an example of such instance together with the synset it is attached to.

5.6 NE identification

A set of articles from the English Wikipedia was randomly selected, and these articles manually tagged as being instances or classes. The set contains 278

Table 7 Number of English NEs per lexicographic file

Lex. File	NEs	Example
act	43,005	Project_Pluto instanceOf project0_4
artifact	55,454	Akinada_Bridge instanceOf suspension_bridge0_6
communication	18,361	Flower_of_Scotland instanceOf national_anthem0_10
event	2,146	Sino-Soviet_split instanceOf schism0_11
group	81,373	Medici instanceOf family0_14
location	111,564	Incense_Route instanceOf trade_route0_15
object	39,321	Pyxis instanceOf constellation0_17
person	520,422	Vladimir_Kotelnikov instanceOf electrical_engineer0_18
time	1,169	Black_Saturday_(France) instanceOf en_s_days0_28
ambiguous	485,542	Barachiel instanceOf archangel?_?

articles and was used to evaluate the different methods we applied to NE identification, the web-based, wikipedia-based and combination methods (see section 4.4). Concerning the capitalisation model, we chose the one that considers the number of times that the first word of the string begins with capital letters. A threshold, minimum percentage of occurrences in which the article title begins with capital letters to be considered a NE, is used. The next paragraphs report on the results obtained by these methods.

5.6.1 Web

Table 8 shows the results obtained by the web method. For several values of the threshold (Thr), precision (P), recall (R) and F-measure ($_{\beta=1}$ and $_{\beta=0.5}$) are included. $F_{\beta=1}$ weights evenly precision and recall whereas $F_{\beta=0.5}$ weights precision twice as much as recall.

Table 8 NE identification results using the web

Thr	P	R	$\mathbf{F}_{\beta=1}$	$\mathbf{F}_{\beta=0.5}$
0.81	74.73	92.52	82.67	77.71
0.83	75.84	91.84	83.08	78.58
0.85	76.74	89.80	82.76	**79.04**
0.87	76.74	89.80	82.76	**79.04**
0.89	76.65	87.07	81.53	78.53
0.91	77.12	80.27	78.67	77.73
0.93	76.81	72.11	74.39	75.82
0.95	76.92	61.22	68.18	73.17

It can be seen that the highest $F_{\beta=0.5}$ is obtained when the threshold is set to 0.85 and 0.87, reaching 79.04% and precision 76.74%. Although other values of the threshold provide higher values of $F_{\beta=1}$, as the aim of the approach is to extend a knowledge resource, we consider precision more important than recall as we think that it is better to link a lower number of NEs to LRs while making sure that the quality of the final resource is good enough.

5.6.2 Wikipedia

We have evaluated this new approach by both looking for entry occurrences only in the English Wikipedia and then again in the English Wikipedia plus the other nine aforementioned Wikipedias. The aim is to increase the precision (76.74%) of the web-based method without causing negative effects in recall (89.80%). Table 9 presents the results obtained for each of the scenarios.

Table 9 NE identification results using Wikipedia

Thr	only English				ten languages			
	P	**R**	$\mathbf{F}_{\beta=1}$	$\mathbf{F}_{\beta=0.5}$	**P**	**R**	$\mathbf{F}_{\beta=1}$	$\mathbf{F}_{\beta=0.5}$
0.81	70.83	92.52	80.24	74.32	74.30	90.48	81.60	77.06
0.83	70.68	91.84	79.88	74.09	74.72	90.48	81.85	77.42
0.85	70.68	91.84	79.88	74.09	75.57	90.48	82.35	78.14
0.87	71.43	91.84	80.36	74.75	75.57	90.48	82.35	78.14
0.89	71.43	91.84	80.36	74.75	76.16	89.12	82.13	78.44
0.91	71.81	91.84	80.60	**75.08**	76.16	89.12	82.13	78.44
0.93	71.81	91.84	80.60	**75.08**	76.02	88.44	81.76	78.22
0.95	71.81	91.84	80.60	**75.08**	76.47	88.44	82.02	**78.60**

The best $F_{\beta=0.5}$ is obtained for the thresholds 0.91 to 0.95 when only using the English Wikipedia (75.08%) and for the threshold 0.95 when using ten Wikipedias (78.60%). For this threshold, using more text allows us to obtain 6% better precision (76.47% vs. 71.81%) while losing 3.7% recall (88.44 vs. 91.84%), which supports our hypothesis of using different Wikipedias to increase the text size. Compared to the web search approach, the current one obtains 1.5% lower recall (88.44% vs. 89.80%) and practically the same precision (76.47% vs. 76.74%). By analysing the results, we have found a drawback of the current approach compared to web search. The number of occurrences found per article is quite low: 7.97 when only using the English Wikipedia and 13.59 when using also the others. These values contrast with those obtained for the web search. In fact, for that experiment we set the number of occurrences per article to 100 and found such a high number for all the articles of the evaluation set.

5.6.3 Combining Wikipedia and the Web

Finally we present the results obtained when combining both methods. We have refined the web method by adding to the query salient words from the Wikipedia article, table 10 presents the results of adding one, two and three words from the article to the query.

From the three configurations, the best $F_{\beta=0.5}$ is obtained when considering two additional words from the body article (81.20% with threshold 0.89). The best results obtained with one and three words are slightly lower, 80.37% (threshold 0.87) and 79.89% (threshold 0.93) respectively. Compared to the

Table 10 NE identification results using the combination method

Thr	1 word			2 words			3 words		
	P	**R**	$\mathbf{F}_{\beta=0.5}$	**P**	**R**	$\mathbf{F}_{\beta=0.5}$	**P**	**R**	$\mathbf{F}_{\beta=0.5}$
0.81	76.67	93.88	79.59	76.80	94.56	79.79	76.16	89.12	78.44
0.83	77.10	93.88	79.95	77.53	93.88	80.32	76.33	87.76	78.37
0.85	77.71	92.52	80.28	77.14	91.84	79.70	76.79	87.76	78.75
0.87	78.44	89.12	**80.37**	77.46	91.16	79.86	77.58	87.07	79.31
0.89	77.91	86.39	79.47	79.17	90.48	**81.20**	77.5	84.35	78.78
0.91	77.99	84.35	79.18	79.63	87.76	81.13	78.21	82.99	79.12
0.93	78.57	82.31	79.29	79.87	86.39	81.10	79.47	81.63	**79.89**
0.95	78.08	77.55	77.98	80.82	80.27	80.71	78.47	76.87	78.15

other methods, this obtains both better precision (79.17% vs. 76.47% and 76.74%) and recall (90.48 vs. 88.44% and 89.80%).

Figure 3 shows the values of $F_{\beta=0.5}$ for the different identification methods in the threshold range [0.81-0.95]. The combination method (with two extra words) obtains better results than the web and Wikipedia methods for any value of the threshold in the whole range.[32]

5.7 Postprocessing

To close the results section, we present the results on the added NEs by exploiting multilingual links and the links to the SUMO and SIMPLE ontologies.

By exploiting Wikipedia's multilingual links we are able to extract 26,157 additional NEs for English, 38,253 for Spanish and 47,168 for Italian. Therefore, the lexicon after this step contains 974,567 English NEs, 137,583 for Spanish and 125,806 for Italian.

In this step we also connect sets of equivalent NEs in different languages (encoded in the "SenseAxis" element) to two ontologies through the "InterlingualExternalRef" element. 814,251 such sets are linked to SUMO while 42,824 get linked to SIMPLE. The substantial difference of the number of entities connected to these ontologies (roughly by a 20 to 1 factor) is due to the fact that in order to connect a set of NEs to SIMPLE, it has to contain an Italian NE linked to PSC while to connect a set to SUMO it needs to contain an English NE linked to WordNet. The results are expected then as the NE lexicon contains much more NEs linked to WordNet (1,366,899) than to PSC (139,190).

Table 11 shows the number of NE sets linked to the different nodes of the SIMPLE ontology. It shows for each ontology node for which a substantial number of NE sets are linked the actual number of NEs, an example of an Italian NE and the PSC wordsense to which this NE is connected.

[32] Despite these results, the Wikipedia method is used for building the NE lexicon because of the limitation of the amount of daily queries imposed by web search engines.

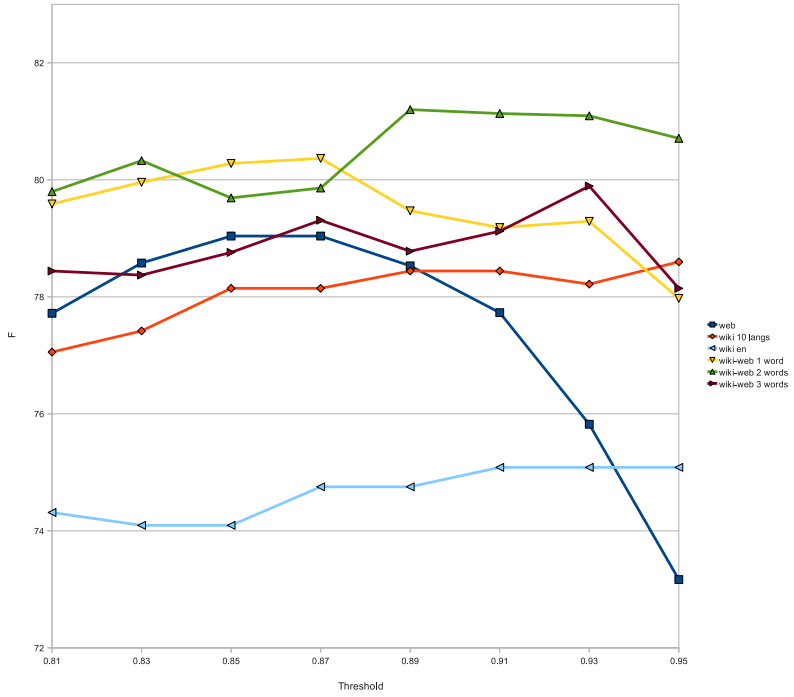

Fig. 3 $F_{\beta=0.5}$ values for the different identification methods

6 Question Answering Application

With the aim of applying our NE lexicon to a real-world NLP task and to validate its usefulness, we have added the knowledge encoded in our lexicon to a QA process.[33] The main idea is to plug the lexicon into a QA system and to use its knowledge to validate the answers given by the system.

For this propose, we have used the BRILIW (Spanish acronym for "QA using Inter Lingual Index module of EuroWordNet and Wikipedia") system (Ferrández et al, 2007b). It was designed to localise answers from documents, where answers and input questions are written in different languages. BRILIW was presented at CLEF 2006 being ranked first in the bilingual English-Spanish QA task (Magnini et al, 2006; Ferrández et al, 2006).

BRILIW architecture is built on three main pillars which stand out among other state–of–the–art Cross–Lingual QA systems: (i) the use of several multilingual knowledge resources to reference words between languages (the ILI module of EuroWordNet and the multilingual knowledge encoded in Wikipedia);

[33] The NE lexicon has also been applied to Machine Translation yielding notable results (Toral and Way, 2011)

Table 11 Number NE sets linked to the SIMPLE ontology

Ontology node	NEs	Example
Artwork	1,221	Las_Meninas_(Velazquez) instanceOf USem837dipinto
Agent_of_persistent _activity	2,890	Carl_Lewis instanceOf USem2018atleta
Building	748	Arena_di_Verona instanceOf USem70845anfiteatro
D_3_location	1,023	Eufrate instanceOf USem5089fiume
Domain	1,804	Martini_Racing instanceOf USem77024automobilismo
Ideo	596	Henri_Bergson instanceOf Usem08517esistenzialista
Institution	2,126	Paramount_Pictures instanceOf USem61226azienda
Instrument	751	Intel 80286 instanceOf USem75625microprocessore
Metalanguage	586	ENIAC instanceOf USem67411acronimo
Profession	18,383	Lukas_Moodysson instanceOf USem3641regista
Purpose_act	689	Coppa_UEFA instanceOf USemD6042competizione
Social_status	6,749	Franco_Turigliatto instanceOf USem3581senatore
Vehicle	1,667	Toyota_Prius instanceOf USem843automobile

(ii) the consideration of more than only one translation per word in order to search candidate answers; and (iii) the analysis of the question in the original language without any translation process.

The architecture of BRILIW is organised as a sequential set of modules. First, the language of the input question is detected. Next, the NEs of the input question are identified and classified with a NE recognition tool and then translated by using Wikipedia. This is followed by an analysis of the input question, where its answer type and its main syntactic blocks are detected. Later on, the equivalents in the target language for the words of the input question are extracted by exploiting the Inter Lingual Index (ILI) module of EuroWordNet. This is done for common nouns and verbs, but not for NEs as these have been previously translated. Subsequently, using as input the translations from ILI (common nouns and verbs) and Wikipedia (NEs), the relevant passages to the input question are fetched by using an Information Retrieval tool. Finally, an ordered list of answers is extracted from the set of relevant passages by applying syntactic patterns.

At this point, we have added a *validation module* which uses the knowledge encoded in the NE lexicon to validate the correctness of the answers, with the possibility of reordering the list of answers provided by BRILIW with the aim of improving the effectiveness of the whole system.

Using the NE lexicon, the *Validation module* is able to validate two types of questions: i) those that expect a NE as the answer type (e.g. Who is the General Secretary of Interpol?); and ii) those which ask for definitions of NEs (e.g. Who is Vigdis Finnbogadottir?). With this objective the model assesses the answer as:

– UNKNOWN: if the expected NE as the answer (type i) or the NE of the question (type ii) are not present in the lexicon.

- CORRECT: if the expected NE as the answer or the NE of the question are present in the lexicon and their types (person, location, etc ...) match with the type tagged by BRILIW.
- INCORRECT: if the expected NE as the answer or the NE of the question are present in the lexicon and their types do not match with the type tagged by BRILIW.

Once the answers are tagged, the *Validation module* reorders the list of answers provided by the system according to the next preferential ranking: CORRECT, UNKNOWN and INCORRECT. Using an official question of CLEF 2006, we show an example of the process in table 12. This example shows how, by using the knowledge encoded in the NE lexicon, the correct answer is returned in the first place, therefore improving the whole accuracy of the system.

Table 12 Example of the Validation module in QA

Question 072 at CLEF 2006: Who is the General Secretary of Interpol?		
Answer	Validation tag	Validation Ranking
Organización Internacional de Policía Criminal	UNKNOWN	2
Enrique Gómez	CORRECT	1
Jefe de la Policía Interna	UNKNOWN	3
Policía Internacional	UNKNOWN	4

We have evaluated the effectiveness of the Validation module and how its knowledge improves the whole precision of the system. For this purpose we have used the CLEF 2006 set of questions, the EFE corpora, the evaluation measures[34] proposed by the CLEF organization (Magnini et al, 2006) and our official results in this competition. In this campaign the CLEF organization decided to use the accuracy, as the main evaluation score, defined as the average of score over all 200 questions. We have used this metric to calculate the overall improvement provided by the Validation module. The results obtained are very promising (see Table 13). BRILIW obtains an improvement of 28.1% compared to the former official results (Ferrández et al, 2006).

Table 13 QA results

Experiment	Overall Accuracy (%)	Improvement (%)
BRILIW	22.5	-
BRILIW + Validation Module	27.5	28.1

[34] The exact answers are assessed as: (1) Right: if correct; (2) Wrong: if incorrect; (3) Inexact: if contained less or more information than that required by the query; or (4) Unsupported: the supporting snippet did not contain the exact answer.

7 Conclusions and Future Work

This paper has presented a generic methodology to automatically create a NE lexicon by combining the complementary views in community-driven and authoritative sources. We have motivated and demonstrated that lexical and semantic knowledge acquisition could benefit from exploiting *New Text* sources such as wikis by showing the potential advantages over common approaches that rely on unrestricted corpora and MRDs.

An important feature of the proposed approach is its high degree of language independence. This method can be directly applied to any language if there is a version of Wikipedia, a LR with a noun taxonomy and a lemmatiser. In fact, we have applied it to LRs based on different theories and covering three languages (English, Spanish and Italian).

The different phases regarding the construction of this resource have been discussed in detail and have been evaluated. These include an initial mapping procedure, the treatment of polysemous nouns, the extraction and identification of NEs and a post-processing step. Finally, we have built a lexicon of NEs that holds the extracted information and whose representation is compliant with the LMF standard (ISO 24613:2008).

The resulting resource contains 974,567, 137,583 and 125,806 NEs for English, Spanish and Italian respectively and 1,366,860, 141,055 and 139,190 "instance of" relations. This resource, together with two APIs (C++ and PHP), is publicly available at `http://computing.dcu.ie/~atoral/#Resources`.

While there exist other previous approaches to build NE repositories (see section 2.2), our proposal clearly represents a step forward in terms of automation, language independence, amount of entities acquired and richness of the information represented in the resulting repository (a comparison of our approach to previous ones across a set of features is shown in Table 14). Therefore, we think that this innovative approach could be applied to other types of linguistic phenomena and lead to important advances in the automatic creation and extension of LRs.

Table 14 Comparison of the NE lexicon to previous approaches

System	languages	LR	size	population
our proposal	en, es, it[1]	3 lexica, 2 ontologies	1,2M	automatic
Sheremetyeva et al (1998)	n/a	ad-hoc	n/a	manual
Mann (2002)	en	none	113K	automatic
Fleischman et al (2003)	en	none	500K	automatic
De Loupy et al (2004)	en	WordNet	130K	semi-automatic
Sarmento et al (2006)	pt	ad-hoc	450K	semi-automatic
Maurel (2008)	fr, de, sr	ad-hoc, EWN	100K	manual

[1] The method allows to acquire NEs for any language present in Wikipedia

We have tested the usefulness of the created resource for real world applications, by applying it to validate the answers produced by a state-of-the art

QA system. With the knowledge of the NE lexicon, the performance of the system increases by 28.1%. The lexicon could be exploited by systems that attempt to classify NEs across a high number of categories. Also, as we provide a classification of entities in nodes of a taxonomy instead of isolated lists of entities for each category, the resource can be used with different levels of granularity for entity recognition.

As it has been said, the methodology introduced has a high degree of language independence. This has been demonstrated by applying it to a set of Indo-European languages, including two Romance languages (Spanish and Italian) and a Germanic language (English). A step forward in order to prove this fact could be assessed by applying our approach to a language that belongs to a different family. In this direction, on-going work is being carried out to exploit the methodology introduced in order to extract Arabic NEs.

It would be also interesting to extract additional types of information in order to enrich the resulting lexicon. For example, it might be interesting to extract relations between NEs. In this case, we plan to identify relations between pairs of Wikipedia articles and to detect their types.

References

Agichtein E, Gravano L (2000) Snowball: extracting relations from large plain-text collections. In: Proceedings of the fifth ACM conference on Digital libraries, ACM, New York, NY, USA, pp 85–94

Agirre E, Soroa A (2009) Personalizing PageRank for word sense disambiguation. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Association for Computational Linguistics, Athens, Greece, pp 33–41

Ahn D, Jijkoun V, Mishne G, de Rijke KMM, Schlobachz S (2005) Using Wikipedia at the TREC QA Track. In: Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004)

Alonge A, Bertagna F, Calzolari N, Roventini A (1999) The Italian Wordnet, EuroWordNet Deliverable D032D033 part B5. Tech. rep.

Alshawi H (1987) Processing dictionary definitions with phrasal pattern hierarchies. Computational Linguistics 13(3-4):195–202

Aristotle (1908) Metaphysics. In: Ross WD (ed) The Works of Aristotle translated into English, Volume VIII, Oxford University Press, Oxford

Atserias J, Villarejo L, Rigau G, Agirre E, Carroll J, Magnini B, Vossen P (2004) The meaning multilingual central repository. In: Proceedings of GWC

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2008) DBpedia: A Nucleus for a Web of Open Data. pp 722–735

Balahur A, Lloret E, Ferrández Ó, Montoyo A, Palomar M, Muñoz R (2008) The DLSIUAES Team's Participation in the TAC 2008 Tracks. In: Notebook Papers of the Text Analysis Conference, TAC 2008 Workshop, National Institute of Standards and Technology, Gaithersburg, Maryland, USA

Bunescu RC, Pasca M (2006) Using Encyclopedic Knowledge for Named entity Disambiguation. In: EACL, The Association for Computer Linguistics

Buscaldi D, Rosso P (2006) Mining Knowledge from Wikipedia for the Question Answering Task. In: Proceedings of The fifth international conference on Language Resources and Evaluation

Calzolari N (1992) Acquiring and representing semantic information in a lexical knowledge base. In: Proceedings of the First SIGLEX Workshop on Lexical Semantics and Knowledge Representation, Springer-Verlag, London, UK, pp 235–243

Daudé J, Padró L, Rigau G (2003) Making wordnet mappings robust. In: Proceedings of the 19th Congreso de la Sociedad Española para el Procesamiento del Lenguage Natural, SEPLN, Universidad Universidad de Alcalá de Henares. Madrid, Spain

De Loupy C, Crestan E, Lemaire E (2004) Proper Nouns Thesaurus for Document Retrieval and Question Answering. In: Atelier Question-Réponse, Traitement Automatique des Langues Naturelles (TALN)

Etzioni O, Banko M, Soderland S, Weld DS (2008) Open information extraction from the web. Communications of the ACM 51(12):68–74

Ferrández Ó, Micol D, Muñoz R, Palomar M (2007a) A Perspective-Based Approach for Solving Textual Entailment Recognition. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Association for Computational Linguistics, Prague, pp 66–71

Ferrández S, López-Moreno P, Roger S, Ferrández A, Peral J, Alvarado X, Noguera E, Llopis F (2006) Monolingual and Cross-Lingual QA Using AliQAn and BRILI Systems for CLEF 2006. In: Peters et al (2007), pp 450–453

Ferrández S, Toral A, Ferrández Ó, Ferrández A, Muñoz R (2007b) Applying Wikipedia's Multilingual Knowledge to Cross-Lingual Question Answering. In: Kedad Z, Lammari N, Métais E, Meziane F, Rezgui Y (eds) NLDB, Springer, Lecture Notes in Computer Science, vol 4592, pp 352–363

Fleischman M, Echihabi A, Hovy E (2003) Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked. In: Proceedings of the ACL Conference. Sapporo, Japan

Francopoulo G, Bel N, George M, Calzolari N, Monachini M, Pet M, Soria C (2008) (forthcoming) Multilingual resources for NLP in the Lexical Markup Framework (LMF). Language Resources and Evaluation Journal

Gabrilovich E, Markovitch S (2007) Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of The Twentieth International Joint Conference for Artificial Intelligence, Hyderabad, India, pp 1606–1611

Giles J (2005) Internet encyclopaedias go head to head. Nature 438(7070):900–901, DOI 10.1038/438900a

Gregorowicz A, Kramer MA (2006) Mining a Large-Scale Term-Concept Network from Wikipedia. Tech. rep., MITRE

Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. In: COLING, pp 539–545

Hearst MA (1998) Automated Discovery of WordNet Relations, MIT Press, Cambridge, MA

ISO 24613 (2008) Languages Resources Management – Lexical Markup Framework (LMF), rev.15 ISOTC37SC4 FDIS. [Online; accessed 25-March-2008]

Jijkoun V, Sang ETK, Ahn D, Möller K, de Rijke M (2005) The University of Amsterdam at QA@CLEF 2005. In: Working Notes of the CLEF 2005 Workshop

Jones G, Fantino F, Newman E, Zhang Y (2008) Domain-Specific Query Translation for Multilingual Information Access Using Machine Translation Augmented With Dictionaries Mined From Wikipedia. In: 2nd International Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies

Karlgren J (ed) (2006) NEW TEXT, Wikis and blogs and other dynamic text sources, Trento, Italy

Krstev C, Vitas D, Maurel D, Tran M (2005) Multilingual Ontology of Proper Names. In: Proceedings of the Language and Technology Conference, pp 116–119

Lenat D (1998) From 2001 to 2001: Common sense and the mind of HAL, MIT Press, Cambridge, MA, pp 193–208

Lenci A, Bel N, Busa F, Calzolari N, Gola E, Monachini M, Ogonowski A, Peters I, Peters W, Ruimy N, Villegas M, Zampolli A (2000) SIMPLE: A general framework for the development of multilingual lexicons. International Journal of Lexicography 13(4):249–263

Magnini B, Giampiccolo D, Forner P, Ayache C, Jijkoun V, Osenova P, Peñas A, Rocha P, Sacaleanu B, Sutcliffe RFE (2006) Overview of the CLEF 2006 Multilingual Question Answering Track. In: Peters et al (2007), pp 223–256

Mann G (2002) Fine-grained proper noun ontologies for question answering. In: Proceedings of SemaNet'02: Building and Using Semantic Networks

Maurel D (2008) Prolexbase: a Multilingual Relational Lexical Database of Proper Names. In: (ELRA) ELRA (ed) Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco

Medelyan O, Legg C (2008) Integrating Cyc and Wikipedia: Folksonomy Meets Rigorously Defined Common-sense. In: AAAI 2008 workshop Wikipedia and Artificial Intelligence: An Evolving Synergy, Chicago, United States

Miller GA (1995) WORDNET: A Lexical Database for English. Communications of ACM (11):39–41

Miller GA, Hristea F (2006) WordNet Nouns: Classes and Instances. Computational Linguistics 32(1):1–3

Milne D, Witten IH (2008) An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: AAAI 2008 workshop Wikipedia and Artificial Intelligence: An Evolving Synergy, Chicago, United States

Milne D, Medelyan O, Witten IH (2006) Mining Domain-Specific Thesauri from Wikipedia: A Case Study. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, Washington, DC, USA, pp 442–448

Nakamura J, Nagao M (1988) Extraction of semantic information from an ordinary english dictionary and its evaluation. COLING-88 pp 459–464

Niles I, Pease A (2003) Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In: Proceedings of the 2003 International Conference on Information and Knowledge Engineering, pp 23–26

Nothman J, Murphy T, Curran JR (2009) Analysing Wikipedia and gold standard corpora for NER training. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics

Pedro V, Niculescu S, Lita L (2008) Okinet: Automatic Extraction of a Medical Ontology From Wikipedia. In: AAAI 2008 workshop Wikipedia and Artificial Intelligence: An Evolving Synergy, Chicago, USA

Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) (2007) Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers, Lecture Notes in Computer Science, vol 4730, Springer

Philpot A, Hovy E, Pantel P (2005) The omega ontology. In: IJCNLP Workshop on Ontologies and Lexical Resources (OntoLex-05), Jeju Island, South Korea, pp 59–66

Ponzetto SP, Strube M (2007) Knowledge Derived from Wikipedia for Computing Semantic Relatedness. Journal of Artificial Intelligence Research 30:181–212

Pustejovsky J (1991) The generative lexicon. Computational Linguistics 17(4):409–441

Richardson SD, Dolan WB, Vanderwende L (1998) MindNet: Acquiring and Structuring Semantic Information from Text. In: COLING-ACL, pp 1098–1102

Rigau G (1998) Automatic Acquisition of Lexical Knowledge from MRDs. PhD thesis, Universitat Politècnica de Catalunya

Roventini A, Ruimy N (2008) Mapping Events and Abstract Entities from PAROLE-SIMPLE-CLIPS to ItalWordNet. In: (ELRA) ELRA (ed) Proceedings of the Sixth International Language Resources and Evaluation, Marrakech, Morocco

Roventini A, Ruimy N, Marinelli R, Ulivieri M, Mammini M (2007) Mapping Concrete Entities from PAROLE-SIMPLE-CLIPS to ItalWordNet: Methodology and Results. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Prague, Czech Republic, pp 161–164

Ruimy N, Corazzari O, Gola E, Spanu A, Calzolari N, Zampolli A (1998) The European LE-PAROLE Project: The Italian Syntactic Lexicon. In: Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain

Ruimy N, Monachini M, Distante R, Guazzini E, Molino S, Ulivieri M, Calzolari N, Zampolli A (2002) CLIPS, a Multi-level Italian Computational lexicon: A Glimpse to Data. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), Las Palmas de

Gran Canaria, Spain
Ruiz-Casado EAM, Castells P (2006) From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach. In: Proceedings of ESWC2006
Sarmento L, Pinto AS, Cabral L (2006) REPENTINO - A wide-scope gazetteer for entity recognition in Portuguese. In: Vieira R, Quaresma P, da Graças Volpes Nunes M, Mamede N, Oliveira C, Dias MC (eds) Proc. of the 7th Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006, Springer, Itatiaia, Rio de Janeiro, Brazil, pp 31–40
Sekine S, Sudo K, Nobata C (2002) Extended Named Entity Hierarchy. In: Proceedings of Third International Conference on Language Resources and Evaluation
Sheremetyeva S, Cowie J, Nirenburg S, Zajac R (1998) Multilingual Onomasticon as a Multipurpose NLP Resource. In: Proceedings of the First International Conference on Language Resources and Evaluation
Snow R, Jurafsky D, Ng AY (2006) Semantic taxonomy induction from heterogenous evidence. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp 801–808
Suchanek FM, Kasneci G, Weikum G (2007) Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web, ACM Press, New York, NY, USA, pp 697–706, DOI 10.1145/1242572.1242667
Sundheim BM, Mardis S, Burger J (2006) Gazetteer Linkage to WordNet. In: Proceedings of the Third International WordNet Conference, pp 103–104
Tjong Kim Sang EF (2002) Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2002, Taipei, Taiwan, pp 155–158
Toral A, Way A (2011) Automatic acquisition of Named Entities for Rule-Based Machine Translation. In: Second International Workshop on Free/Open-Source Rule-Based Machine Translation
Tran M, Grass T, Maurel D (2004) An Ontology for Multilingual Treatment of Proper Names. In: Proceedings of OntoLex 2004
Verdejo MF (1999) The Spanish Wordnet, EuroWordNet Deliverable D032D033 part B3. Tech. rep.
Vossen P (1998) EuroWordNet A Multilingual Database with Lexical Semantic Networks. Kluwer Academic publishers
Widdows D, Ferraro K (2008) Semantic vectors: a scalable open source package and online technology management application. In: (ELRA) ELRA (ed) Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco
Wiebe J, Riloff E (2005) Creating subjective and objective sentence classifiers from unannotated texts. In: Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics., Springer-Verlag, Mexico City, MX, Lecture Notes in Computer Science, vol

3406, pp 475–486
Wiebe JM, Wilson T, Bruce RF, Bell M, Martin M (2004) Learning subjective language. Computational Linguistics 30(3):277–308
Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco, United States of America
Wu F, Hoffmann R, Weld DS (2008) Augmenting wikipedia-extraction with results from the web. In: AAAI 2008 workshop Wikipedia and Artificial Intelligence: An Evolving Synergy, Chicago, United States
Zesch T, Müller C, Gurevych I (2008) Extracting lexical semantic knowledge from wikipedia and wiktionary. In: (ELRA) ELRA (ed) Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco

A LMF output

This appendix contains an output sample in LMF format and in the database. It is made up of three monolingual lexicons whose entries are linked by using the "SenseAxis" object of the LMF multilingual extension.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LexicalResource SYSTEM "DTD_LMF_REV_16.dtd">
<LexicalResource dtdVersion="16">
 <GlobalInformation>
  <feat att="label" val="Multilingual Named Entity Repository"/>
 </GlobalInformation>
 <Lexicon>
  <feat att="Language" val="it"/>
  <LexicalEntry id="it_le_città">
   <Lemma>
    <FormRepresentation>
     <feat att="writtenform" val="città"/>
     <feat att="VariantType" val="full"/>
    </FormRepresentation>
   </Lemma>
   <Sense id="it_s_città_1">
    <MonolingualExternalRef>
     <feat att="external_system" val="PSC"/>
     <feat att="external_reference" val="USem2234citta1"/>
    </MonolingualExternalRef>
   </Sense>
  </LexicalEntry>
  <LexicalEntry id="it_le_Firenze">
   <Lemma>
    <FormRepresentation>
     <feat att="writtenform" val="Firenze"/>
     <feat att="VariantType" val="full"/>
    </FormRepresentation>
   </Lemma>
   <Sense id="it_s_Firenze">
    <SenseRelation targets="it_s_città_1">
     <feat att="semanticrelation" val="instance_of"/>
    </SenseRelation>
    <MonolingualExternalRef>
     <feat att="external_system" val="ItWikipedia"/>
     <feat att="external_reference" val="1118816"/>
    </MonolingualExternalRef>
   </Sense>
```

```
  </LexicalEntry>
 </Lexicon>
 <Lexicon>
  <feat att="Language" val="en"/>
  <LexicalEntry id="en_le_city">
   <Lemma>
    <FormRepresentation>
     <feat att="writtenform" val="city"/>
     <feat att="VariantType" val="full"/>
    </FormRepresentation>
   </Lemma>
   <Sense id="en_s_city_1">
    <MonolingualExternalRef>
     <feat att="external_system" val="EnWordNet"/>
     <feat att="external_reference" val="noun.loc:city0"/>
    </MonolingualExternalRef>
   </Sense>
  </LexicalEntry>
  <LexicalEntry id="en_le_Florence">
   <Lemma>
    <FormRepresentation>
     <feat att="writtenform" val="Florence"/>
     <feat att="VariantType" val="full"/>
    </FormRepresentation>
   </Lemma>
   <Sense id="en_s_Florence">
    <SenseRelation targets="en_s_city_1">
     <feat att="semanticrelation" val="instance_of"/>
    </SenseRelation>
    <MonolingualExternalRef>
     <feat att="external_system" val="EnWikipedia"/>
     <feat att="external_reference" val="11525"/>
    </MonolingualExternalRef>
   </Sense>
  </LexicalEntry>
 </Lexicon>
 <SenseAxis id="sa_001" senses="en_s_Florence it_s_Firenze">
  <feat att="type" val="eq_syn"/>
  <InterlingualExternalRef>
   <feat att="external_system" val="SUMO"/>
   <feat att="external_reference" val="City"/>
   <feat att="external_reltype" val="at"/>
  </InterlingualExternalRef>
  <InterlingualExternalRef>
   <feat att="external_system" val="SIMPLE"/>
   <feat att="external_reference" val="Geopolitical_location"/>
   <feat att="external_reltype" val="at"/>
  </InterlingualExternalRef>
 </SenseAxis>
</LexicalResource>
```

Table 15 NE Repository. LexicalEntry table

LE id	LE pos
it_le_Firenze	PN
it_le_città	N
en_le_Florence	PN
en_le_city	N

Table 16 NE Repository. FormRepresentation table

LE id	written form	variant type
it_le_Firenze	Firenze	full
it_le_città	città	full
en_le_Florence	Florence	full
en_le_city	city	full

Table 17 NE Repository. Sense table

S id	LE id	ext. resource	resource id	definition
it_s_Firenze	it_le_Firenze	it_Wikipedia	1118816	...
it_s_città1	it_le_città	it_PSC	USem2234citta1	...
en_s_Florence	en_le_Florence	en_Wikipedia	11525	...
en_s_city1	en_le_city	en_WordNet	noun.loc:city0	...

Table 18 NE Repository. SenseRelation table

source id	target id	relation
it_s_Firenze	it_s_città1	instanceOf
en_s_Florence	en_s_city1	instanceOf

Table 19 NE Repository. SenseAxis table

SA id	type
1	eq synonym

Table 20 NE Repository. SenseAxisElements table

SA id	element
1	it_s_Firenze
1	en_s_Florence

Table 21 NE Repository. SenseAxisExternalRef table

SA id	resource	resource id	relation
1	SUMO	city	at
1	SIMPLE	Geopolitical_location	at