Using WordNet Predicates for Multilingual Named Entity Recognition

Matteo Negri and Bernardo Magnini

ITC-Irst, Centro per la Ricerca Scientifica e Tecnologica,
Via Sommarive, 38050 Povo (TN), Italy
Email: negri@itc.it, magnini@itc.it

Abstract. *WordNet predicates* (WN-PREDS) establish relations between words in a certain language and concepts of a language independent ontology. In this paper we show how WN-PREDS can be profitably used in the context of multilingual tasks where two or more wordnets are aligned. Specifically, we report about the extension to Italian of a previously developed Named Entity Recognition (NER) system for written English. Experimental results demonstrate the validity of the approach and confirm the suitability of WN-PREDS for a number of different NLP tasks.

1 Introduction

WORDNET predicates (WN-PREDS) are defined over a set of WORDNET synsets which express a certain concept. A WN-PRED takes as input a word *w* and a language *L* in which the word is expressed, and returns TRUE if at least one sense of *w* in *L* is subsumed by at least one of the synsets defining the predicate, and FALSE otherwise. As an example, a WN-PRED "*location-p*" can be defined over the high-level synsets `location#1`, `mandate#2`, `road#1`, `solid_ground#1`, `body_of_water#1`, `geological_formation#1`, and `celestial_body#1`[1]. According to the previous definition:

location-p [<capital>, <English>]

returns `capital#3` (i.e.TRUE) since this sense of "capital" in the English WORDNET is subsumed by at least one of the synsets defining the predicate (*i.e.* `location#1`). On the other hand:

location-p [<computer>, <English>]

returns FALSE since none of the senses of "computer" is subsumed by one of the synsets defining the concept of location.

WORDNET predicates establish relations between a single word in a language and a general concept in a language independent ontology. However, WORDNET predicates are context independent *i.e.* they produce the same result for the same word, independently of the context in which the word occurs. As a consequence, their practical use is limited to applications (such as the one proposed in this paper) in which predicates are coupled with contextual information.

[1] Throughout the paper WORDNET word senses are reported with this `typeface#1`, where #1 is the corresponding sense number in WORDNET 1.6, while Named Entity categories are indicated with this TYPEFACE.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 169–174.

While the use of WORDNET predicates has been proposed in several NLP tasks, including Named Entity Recognition (NER) [3] and Question Answering (QA) [6], this paper addresses their more specific use in a multilingual scenario, where two or more wordnets are aligned. Starting from the WORDNET predicates used in an NER system for written English (overviewed in Section 2), we experimented the portability of the approach building an Italian system without any change in the predicates (Section 3). Results (Section 4) are highly encouraging, and demonstrate the suitability of the proposed methodology both in term of performance and in term of time required for system development.

2 Using WORDNET Predicates for NER

NER is the task of identifying and categorizing entity names (such as persons, organizations, and locations names), temporal expressions (dates and times), and certain types of numerical expressions (monetary values and percentages) in a written text. Knowledge-based approaches, which represent a possible solution to the NER problem, usually rely on the combination of a wide range of knowledge sources (for example, lexical, syntactic, and semantic features of the input text as well as world knowledge and discourse level information) and higher level techniques (*e.g.* co-reference resolution). In this framework, dictionaries and extensive gazetteer lists of first names, company names, and corporate suffixes are often claimed to be a useful resource. Nevertheless, several works (see, for example, [5]) pointed out some drawbacks related to the pure list lookup approach, which mainly depend on the required dimensions of reliable gazetteers, on the difficulty of maintenance of this kind of resource, and on the possibility of overlaps among the lists. Moreover, their availability for languages other than English is rather limited.

An effective solution to these problems has been recently proposed in [3], and relies on a rule-based approach which avoids the difficulties related to the construction and maintenance of reliable gazetteers by making the most of the information stored in the WORDNET hierarchy. The starting point, as also suggested by [4], is that the identification and classification of a candidate named entity can be tackled by considering two kinds of information, namely *internal* and *external* evidence. The former is provided by the candidate string itself, while the latter is provided by the context in which the string appears. As an example, in the sentence, “Judge Pasco Bowman II, who was appointed by President Ronald Reagan ...”, the candidate proper names “Pasco Bowman II” and “Ronald Reagan” can be correctly marked with the tag PERSON either by accessing a database of person names (*i.e.* considering their internal evidence) or by considering the appositives “Judge”, “II” and “President”, or the pronoun “who” as external evidence for disambiguation.

While internal evidence is mostly conveyed by proper nouns, external evidence can be conveyed by the presence in the text of *trigger words*, *i.e.* predicates and constructions providing sufficient contextual information to determine the class of candidate proper nouns in their proximity [9]. For instance, systems designed to deal with this kind of information usually access more or less complete hand-crafted word lists containing expressions like “director”, “corporation”, and “island” in order to recognize respectively person, organization, and location names into a given text.

In light of these considerations, the basic assumption underlying the approach suggested by [3] is that the huge number of possible trigger words that can be extracted from WORD-

NET compensates for the relatively limited availability of proper nouns, thus forming a reliable basis to accomplish NER without the further use of gazetteer lists. In this framework, they propose a semi-automatic procedure to extract trigger words from WORDNET, and to separate them from proper nouns bringing internal evidence. This procedure exploits the IS-A relation to distinguish between *Word_Classes* (*i.e.* concepts bringing external evidence, such as `river#1`) and *Word_Instances* (*i.e.* particular instances of those concepts, such as `Mississippi#1`, which can be marked as entity words also without any contextual information) present in WORDNET. For instance, as for the NE category LOCATION, starting from the high level synsets already listed in Section 1, and considering as proper nouns their capitalized hyponyms, they obtain 1591 English Word_Classes and 2173 Word_Instances. Once the relevant high level synsets have been selected, and the corresponding Word_Classes and Word_Instances have been mined from the WORDNET hierarchy, WORDNET predicates relevant to each NE category (*e.g.* "person-p", "person-name-p" "location-p", "location-name-p", "organization-p", etc.) are used to access this information in the NER process. The task is accomplished by means of simple rules that check for different features of the input text, detecting the presence of particular word senses satisfying the WORDNET predicates, as well as word lemmas, parts of speech or symbols.

3 Porting to Italian

The construction of an NER system for written Italian represented an ideal opportunity to test the portability of the above outlined approach, which [3] has claimed to be well-suited to multilingual extensions. In fact, in addition to its effectiveness in the NER task, mining information from WORDNET also offers a practicable way to address multilinguality. This is due to the recent spread of multilingual semantic networks aligned with WORDNET, a necessary condition for the complete reusability of the predicates defined on the English taxonomy.

Our extension to Italian takes advantage of MULTIWORDNET [8], a multilingual lexical database developed at ITC-Irst which includes information about English and Italian words. MULTIWORDNET is an extension of the English Princeton WORDNET, keeping as much as possible of the original semantic relations. Italian synsets have been created in correspondence with English synsets, whenever possible, by importing lexical and semantic relations from the corresponding English synsets. The Italian part of MULTIWORDNET currently covers about 43,000 lemmas, completely aligned with English WORDNET 1.6.

Exploiting the alignment between the two languages, Italian Word_Classes and Word_Instances have been mined from MULTIWORDNET starting from the high-level synsets defined on the English taxonomy and collecting their Italian equivalents as well as their hyponyms. Table 1 shows their distribution with respect to the NE categories we used in our experiments (namely PERSON, LOCATION, and ORGANIZATION), compared to the distribution of the English words. It's worth noting that, in order to improve the system performance, all the English Word_Instances have been also used in our extension since most of them (*e.g.* proper nouns like "William Shakespeare", "Beverly Hills", and "UNESCO") usually are not translated into Italian. The same holds for some of the English Word_Classes (*e.g.* "anchorman", "checkpoint", and "corporation"), which can be considered as trigger words also when they are encountered within an Italian text. This way, even though the over-

Table 1. Distribution of Word_Classes and Word_Instances in MULTIWORDNET

	#ENG Classes	*#ENG Instances*	*#ITA Classes*	*#ITA Instances*
PERSON	6775	1202	5982	348
LOCATION	1591	2173	979	950
ORGANIZ.	1405	498	890	297
TOTAL	9771	3873	7851	1595

all number of Italian words is lower, both internal and external evidence are still effectively captured by the system.

Using the information mined from the MULTIWORDNET hierarchy, and taking advantage of the complete reusability of the English WORDNET predicates, the process of recognition and identification of NEs is carried out in three phases.

Preprocessing. In the first phase, the input text is tokenized and words are disambiguated with their lexical category by means of a statistical part of speech tagger developed at ITC-Irst. Also, multiwords recognition is carried out in this phase: about seven thousand Italian multiwords (*i.e.* collocations, compounds, and complex terms) have been automatically extracted from MULTIWORDNET and are recognized by pattern matching rules.

Basic rules application. In the second phase, a set of approximately 400 *basic rules* is in charge of finding and tagging all the possible NEs present in the input text. Most of these rules capture internal and external evidence by means of the WORDNET predicates used to mine the Italian taxonomy. As an example, Table 2 describes a rule containing the WORDNET predicate "location-p", which is satisfied by any of the 979 Italian Word_Classes of the category LOCATION extracted from MULTIWORDNET. This rule captures contextual evidence matching with sentences formed by a capitalized noun followed by a verb whose lemma is "essere" (*i.e.* "to be"), a determiner, and any of those trigger words, like "capitale" in "Roma e' la capitale italiana" (*i.e.* "Rome is the Italian capital").

Table 2. A rule matching with "Roma e' la capitale italiana"

PATTERN	*t1 t2 t3 t4*
t1	[pos = "NP"] [ort = Cap]
t2	[lemma = "essere"]
t3	[pos = "DT"]
t4	[sense = (location-p *t4* Italian)]
OUTPUT	<LOCATION>t1<\LOCATION>

Composition rules application. Besides the application of the basic rules, a correct NER procedure requires the application of higher level rules in charge of resolving co-references between recognized entities and proper names not yet disambiguated, as well as handling tagging ambiguities, tag overlaps and inclusions. For instance, considering the start/end position of the tags, the content, and the tag type of the candidate entities, these rules handle inclusions which may occur when a recognized entity contains other more specific entities, as in "Università di Napoli" (*i.e.* "Naples University"), where a proper noun belonging to the category LOCATION (*i.e.* "Napoli") is included into an entity belonging to the more general category ORGANIZATION.

4 Results and Conclusion

System performance was evaluated using the scoring software provided in the framework of the DARPA/NIST HUB4 evaluation exercise [1]. Scores (i.e. F-measure, Precision and Recall) have been computed by comparing a 77 Kb reference tagged corpus[2] with an automatically tagged corpus according to *type*, *content* and *extension* of the NE categories PERSON, LOCATION, and ORGANIZATION. Table 3 illustrates the results achieved by our system, compared with the performance of the English version described by [3].

Table 3. Overall Precision, Recall and F-Measure scores

	Recall	*Precision*	*F-Measure*
PERSON	91.48 (87.29)	85.08 (88.38)	88.16 (87.83)
LOCATION	97.27 (92.16)	80.45 (81.17)	88.07 (86.32)
ORGANIZATION	83.88 (82.71)	72.70 (83.02)	77.89 (82.87)
All categories	91.32 (87.28)	74.75 (82.99)	82.21 (84.12)

As can be seen from Table 3, even though MULTIWORDNET is smaller than WORDNET, our results compare well with the ones achieved by the English version. For instance, considering the category LOCATION, even if for WORDNET 1.6 provides about 600 Word_Classes more than the Italian part of MULTIWORDNET, the difference between the two F-Measure scores is rather small (*i.e.* 0.67). The suitability and the portability to other languages of the WORDNET-based approach to NER are also confirmed by the relatively limited amount of time required for system development. In fact, since the WORDNET predicates defined on the English taxonomy were reused without any change, all the effort was concentrated on the creation of the Italian rules, which took approximately one person month.

As a final remark, it's worth noting that while in the present work WORDNET predicates have been defined according to the concepts that are relevant for the NER task (*i.e.* PERSON, LOCATION, and ORGANIZATION), a wider set of such predicates can be easily realized by taking advantage of the concepts defined in already available upper-level ontologies and their mappings to WORDNET. Among these ontologies, an important role in the framework of approaches similar to the one described in this paper could be played by the SUMO ontology [7], with about 1100 concepts completely mapped against WORDNET, and the DOLCE ontology [2], whose mapping to WORDNET is, however, still under development.

[2] Reference transcripts of two Italian broadcast news shows, including a total of about 7,000 words and 322 tagged named entities, were manually produced for evaluation purposes

References

1. Chinchor, N., Robinson, P., Brown, E.: Hub-4 Named Entity Task Definition (version 4.8). Technical Report, SAIC. `http://www.nist.gov/speech/hub4_98` (1998).
2. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening Ontologies with DOLCE. Proceedings of EKAW 2002. Siguenza, Spain (2002).
3. Magnini, B., Negri M., Prevete R., Tanev H.: A WordNet-Based Approach to Named Entities Recognition. Proceedings of SemaNet '02: Building and Using Semantic Networks Taipei, Taiwan (2002) 38–44.
4. McDonald, D.: Internal and external evidence in the identification and semantic categorization of proper names. In: Boguraev, I., Pustejovsky, J. (eds.): Corpus Processing for Lexical Acquisition, Chapter 2. The MIT Press, Cambridge, MA (1996).
5. Mikheev, A., Moens, M., Grover, C.: Named Entity recognition without gazetteers. Proceedings of EACL-99, Bergen, Norway (1999).
6. Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lacatusu, F., Novischi, A., Badulescu, A., Bolohan, O.: LCC Tools for Question Answering. Proceedings the TREC-2002 Conference, NIST, Gaithersburg, MD (2002), Bergen, Norway (1999).
7. Niles, I., Pease, A.: Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03), Las Vegas, Nevada, (2003).
8. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: Developing an Aligned Multilingual Database. Proceedings of the 1st International Global WordNet Conference, Mysore, India (2002).
9. Wakao, T., Gaizauskas, R., Wilks, Y.: Evaluation of an Algorithm for the Recognition and Classification of Proper Names. Proceedings of the 16th Conference on Computational Linguistics (COLING '96), Copenhagen, Denmark (1996).