A WordNet-Based Approach to Named Entities Recognition

Bernardo Magnini, Matteo Negri, Roberto Prevete and **Hristo Tanev**
ITC-Irst, Centro per la Ricerca Scientifica e Tecnologica
[magnini,negri,prevete,tanev]@itc.it

Abstract

This paper presents a Named Entities (NE) recognition system for the English written language, which combines the wealth of the WordNet taxonomy and the effectiveness of traditional rule-based approaches. The core of the system relies on the combination of approximately 200 language-dependent rules with a set of predicates, defined on the WordNet hierarchy, for the identification of both proper nouns and *trigger words*. The strengths of this approach are twofold. First, the use of a semantic network allows it to cope with the difficulty of building and maintaining extensive gazetteers. Second, considering the recent spread of WordNet-like semantic networks for languages other than English and aligned with the English version, the use of language-independent predicates offers a useful basis for achieving multilinguality.

1 Introduction

Named Entities (NE) recognition represents a crucial aspect in the process of natural language understanding. Since up to 10% of a newswire text may consist of proper names, dates, times, and similar expressions (Coates-Stephens, 1992), their effective identification is required both in Information Extraction and Information Retrieval related tasks.

The NE recognition task has been defined in the context of the Message Understanding Conferences (MUC) (Chinchor, 1998) as the capability 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 represent a possible solution to the NE recognition problem. Such techniques 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 the context of this framework, dictionaries and extensive gazetteer lists of first names, company names, and corporate suffixes are often claimed to be a very useful resource. Although the recourse to list lookup seems to be a straightforward method to achieve reasonable performance, several works pointed out that the compilation of gazetteers represents a bottleneck in the design of an NE recognition system and that their availability for languages other than English is rather limited (Cucchiarelli et al., 1998). (Mikheev et al., 1999) presents an exhaustive discussion about the drawbacks related to the pure list lookup approach. Such problems mainly depend on the required dimensions of reliable gazetteers, on the difficulty of maintenance of this kind of resource (proper names form an open class, making the incompleteness of gazetteers an obvious problem), and on the possibility of overlaps among the lists (for instance, the name "Washington" could refer either to a person or to a location).

The difficulties related to the construction and maintenance of reliable gazetteers can be overcome by focusing attention on the presence in the text of *trigger words*, i.e. predicates and constructions typically associated with named entities (Wakao et al., 1996). Trigger words often provide a sufficient contextual information to determine the class of candidate proper nouns in their proximity. As an example, sys-

tems designed to deal with this kind of contextual information usually access more or less complete hand-crafted word lists containing expressions like "President", "Corporation", and "County" in order to recognize respectively person, organization, and location names into a given text.

According to (McDonald, 1996), the identification and the classification of a candidate named entity can be tackled by considering two kinds of information: *internal* evidence and *external* evidence. The former is provided by the candidate string itself, while the latter is provided by the context into 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[1] 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 sufficient for disambiguation.

In light of the above considerations, the distinguishing features of our approach to NE recognition are the following:

(i) We avoid the use of gazzeteers; instead, we exploit WORDNET, a well known resource freely available (Fellbaum, 1998).

(ii) As a consequence, trigger words assume a crucial role. In order to identify trigger words relevant to the NEs categories defined in the context of the DARPA/NIST HUB4 evaluation exercise (Chinchor et al., 1998), and to separate them from words bringing internal evidence, we resort on the distinction between *Word_Class* and *Word_Instance* in WORDNET. Word_Classes often individuate trigger words, while Word_Instances provide internal evidence. Our hypothesis is that the huge number of possible trigger words that can be extracted from WORDNET compensates for the relatively limited availability of proper nouns, thus forming a reliable basis to accomplish NE recognition without the further use of gazetteer lists.

(iii) The NE recognition system is rule-based. Rules are designed to take advantage of the WORDNET taxonomy, in particular of abstractions induced by the IS-A hierarchy.

An additional advantage (that we have not yet experimented with) of a WORDNET-based approach to NE recognition is that, as multilingual semantic networks aligned with the English version of WORDNET are available, the construction of multilingual NE recognition systems becomes a much easier challenge.

This paper is structured as follows: Section 2 introduces our WORDNET-based approach. Section 3 describes how named entities are recognized and how high-level rules can be used to cope with tagging ambiguities and co-reference resolution. Section 4 shows experimental results. Section 5 concludes the paper with a discussion about the presented approach.

2 Exploiting WordNet hierarchy for NE recognition

WORDNET is a lexical database whose basic building block is the synset, a set of synonym words representing an underlying lexical concept. The 1.6 English version contains 129,505 words organized into 99,642 synsets. In WORDNET, two kinds of relations are distinguished: semantic relations (e.g. IS-A, part-of, cause, etc.), which hold among synsets, and lexical relations (e.g. synonymy, antonymy), which hold among words. According to the IS-A (or hyperhyponymic) semantic relation, nouns and verbs are hierarchically organized in a sequence of levels going from generic concepts at the top to specific concepts at the lower levels.

2.1 Word Classes and Word Instances

Our criterion for distinguishing between trigger and entity words relies on the distinction between Word_Class and Word_Instance in the WORDNET hierarchy. This distinction has been first introduced in the EuroWordNet model (Alonge et al., 1998) to capture the difference between concepts (*e.g.* `river#1`) and particular instances of those concepts (*e.g.* `Mississippi#1`). The two EuroWordNet relations "has_instance" and "belongs_to_class" provide a connection between Word_Classes and Word_Instances.

Unfortunately, the WORDNET hierarchy does not provide such a distinction. The ontological confusion among classes and individuals in

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

WordNet has been pointed out by (Gangemi et al., 2002). As an example, they report the hyponyms of the synset `person#1`, which are a mixture of concepts (*e.g.* "astronomer", "philosopher", "socialist", etc.) and individuals (*e.g.* "Galileo Galilei", "Ludwig Wittgenstein", "Karl Marx", etc.).

In order to cope with the lack of expressivity in WordNet, we devised a semi-automatic procedure which exploits the IS-A relation to distinguish between Word_Classes and Word_Instances. This distinction has been accomplished in two steps. First, a set of predicates has been defined for the extraction of the hyponyms of several high-level synsets, such as `person#1`, `social_group#1`, `location#1` and `measure#3`. Second, instances have been separated from classes via simple heuristics (*e.g.* capitalized words have been considered as instances, while lower case words have been considered as Word_Classes) and then manually checked. Table 1 shows the distribution of Classes and Instances over the WordNet hierarchy with respect to the NE categories Person, Location, Organization, Measure, Money, Duration, Date, Time, Percent and Cardinal, which are the categories considered in the design of our system.

	#Classes	*#Instances*
Person	6775	1202
Location	1591	2173
Organiz.	1405	498
Measure	622	-
Money	265	-
Duration	1054	-
Date	363	3
Time	60	-
Cardinal	124	-
Percent	-	-
TOTAL	12259	3876

Table 1: Distribution of Word Classes and Word Instances in WordNet 1.6

2.2 Capturing external evidence using the WordNet hierarchy

Given an input text, our rule system captures external evidence considering all the 12259 words belonging to the relevant Word_Classes (see Table 1) as possible trigger words.

As an example, among the 6086 hyponyms of the synset `person#1` {person, individual, someone, somebody, mortal, human, soul}, a class of 6775 trigger words has been extracted. Among these, words like "astronomer", "physicist", "Norwegian", and "professor" provide our system with the essential contextual information for the recognition of a person name (i.e. "Christopher Hansteen") in the text fragments reported in Table 2.

1. "<b_enamex type="PERSON">Christopher Hansteen<e_enamex> was an astronomer who devoted his time to the study of geomagnetism"
2. "<b_enamex type="PERSON">Christopher Hansteen<e_enamex>, a physicist, mounted an expedition to <b_enamex type="LOCATION"> Siberia<e_enamex> to study magnetic declinations"
3. "In the same year, the Norwegian professor <b_enamex type="PERSON">Christopher Hansteen<e_enamex>, (<b_timex type="DATE">1784<e_enamex>-<b_timex type="DATE">1873<e_enamex>), wrote an atlas of magnetic strength and declination."

Table 2: Text fragments tagged with NE

2.3 Capturing internal evidence from the WordNet hierarchy

Internal evidence is captured, via pure list lookup, not only from the Word_Instances lists, but also from the Word_Classes lists.

All the 3876 Word_Instances mined from the WordNet hierarchy are supposed to provide internal evidence. As an example, instances like "Galileo", "New York", "Federal Home Loan Mortgage Corporation" and "6 June 1944" (which are hyponyms of `person#1`, `location#1`, `social group#1` and `time unit#1` respectively) are marked as entity words also without any contextual information.

Moreover, many words belonging to the Word_Classes lists can be considered as entities by virtue of their internal evidence. As an ex-

ample, most of the hyponyms of `time unit#1` (*e.g.* "Monday", "mid-January", "seventies", "Christmas", etc.) can be correctly considered either as classes (i.e. trigger words for complex multiword expressions) or as entity words.

3 Recognition and Classification of Named Entities

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. Also multiwords recognition is carried out in this phase: about five thousand multiwords (i.e. collocations, compounds, and complex terms) have been automatically extracted from WordNet and are recognized by pattern matching rules.
- *Basic rules application.* In the second phase, a set of approximately 200 basic rules is used for finding and tagging all the possible NEs present in the input text.
- *Composition rules application.* Finally, a set of higher-level rules is used to resolve ambiguities between possible multiple tags as well as for co-reference resolution.

Basic rules and composition rules are described in the following two sections.

3.1 Basic rules

As stated before, our system has been designed for the recognition of the NE categories described in Table 1. Each category is associated with a set of basic rules that check for different features of the input text. These rules may detect the presence of particular word senses, lemmas, parts of speech or symbols. The whole set of basic rules has been manually created in a few weeks considering, as training data, a small corpus (about 350 Kb) of English newswire texts.

Most of our rules rely on the internal and external evidence captured via the predicates defined on the WordNet hierarchy. As an example, Tables 3 and 4 describe two rules containing predicates that can be satisfied only by particular word senses. In the first rule, the predicate "proper-person-name-p" is satisfied by any of the 1202 Instances of the category Person. The rule captures internal evidence matching any occurrence of these instances in the input text (*e.g.* "Galileo" in "Galileo invented the telescope").

PATTERN	t1
t1	[sense = proper-person-name-p]
OUTPUT	<PERSON>t1<\PERSON>

Table 3: A basic rule matching with "Galileo invented the telescope"

In the second rule, the predicate "person-p" is satisfied by any of the 6775 *Classes* of the category Person. This rule captures contextual evidence matching with sentences formed by a capitalized noun followed by a verb whose lemma is "be", a determiner, and any of these trigger words (*e.g.* "astronomer" in "Hansteen was an astronomer")

PATTERN	t1 t2 t3 t4
t1	[pos = "NP"] [ort = Cap]
t2	[lemma = "be"]
t3	[pos = "DT"]
t4	[sense = person-p]
OUTPUT	<PERSON>t1<\PERSON>

Table 4: A basic rule matching with "Hansteen was an astronomer"

In some cases, the presence of particular word senses is not required. In fact, external evidence can often be captured from the context even in the absence of trigger words. As an example, instead of checking the presence of particular word senses, the rule described in Table 5 considers the contextual information provided by the following sentence structure: a capitalized noun followed by a comma and the pronoun "who" (*e.g.* "Pangborn, who flew across the Pacific Ocean").

Percentages and cardinals are handled in the same way. The reason for this is that these numeric expressions, as well as many temporal expressions (*e.g.* "05/15/2002") in English texts have a fairly structured appearance which can be reliably captured by means of simple WordNet-independent rules.

PATTERN	t1 t2 t3
t1	[pos = "NP"] [ort = Cap]
t2	[lemma = ","]
t3	[lemma = "who"]
OUTPUT	<PERSON>t1<\PERSON>

Table 5: A basic rule matching with "Pangborn, who flew across the Pacific Ocean"

3.2 Composition-Rules

The output of the basic rules application phase is processed by a set of composition rules. These rules are in charge of handling inclusions between tagged entities, as well as resolving co-references between recognized entities and proper names not yet disambiguated. The final output is a version of the original text, in which all the detected NEs are marked up with a set of pre-defined SGML tags specifying their category (for an example, see the NE tagged text fragments reported in Table 2).

Tag inclusions and co-reference resolution are handled by rules considering the start/end position of the tags, the content and the tag type of the candidate entities.

Inclusions may occur when a recognized entity contains one or more other entities. As an example, consider the sentence "Boston is about 200 miles from New York". The basic rules application phase recognizes (besides the two locations) one entity belonging to the category CARDINAL (i.e. "200"), which is included into an entity belonging to the category MEASURE (i.e. "200 miles"). In order to cope with inclusions, a hierarchy of tag categories has been defined. According to this hierarchy, the categories PERSON, LOCATION and ORGANIZATION are all subsumed by the more general category NAMEX. In a similar way, the categories MONEY, MEASURE, TIME, DURATION, DATE and PERCENT are all subsumed by the more general category CARDINAL. In case of tag inclusions, our system always chooses the most specific ones. Table 6 shows an example of a composition rule for handling inclusions (the symbol $\dashv$ is used to indicate the subsumption relation between categories).

Co-reference resolution contributes to proper name classification recognizing, in the input text, parts of entities that have already been disambiguated. Often, in fact, the first reference to an entity includes a relatively full form of its name (*e.g.* "Professor Christopher Hansteen" in the Table 2 examples) (Wacholder et al., 1997). In a kind of discourse anaphora, other references to the entity take the form of shorter, more ambiguous variants (*e.g.* "Hansteen wrote an atlas of magnetic strength"). Co-reference resolution aims at capturing all these variants exploiting the hierarchy of tag categories described above. If a proper name has been tagged with the general category NAMEX, the system checks if the same name, or a combination of words containing it, has been found anywhere in the text and has already been tagged with a more specific category. If so, the proper name is also tagged with that more specific category.

PATTERN	NE1 NE2
NE1	[start = n] [end = m] [TAG = A]
NE2	[start = n$\leq$o<m] [end = o<p$\leq$m] [TAG = B$\dashv$A]
OUTPUT NE1	 [start = n] [end = m] [TAG = A]

Table 6: A composition rule for handling inclusions

PATTERN	NE1 NE* NE2
NE1	[entity = α] [TAG = A]
NE2	[entity = $\beta\subseteq\alpha$] [TAG = "NAMEX"]
OUTPUT NE2	 [entity = β] [TAG = A]

Table 7: A composition rule for co-reference resolution

Table 7 shows an example of composition rule for co-reference resolution (the symbol $\subseteq$ is used to indicate that the content of a recognized entitiy is equal to or is part of the

content of another recognized entity).

4 Results and discussion

In order to evaluate the performance of our system, an experiment was carried out using the test corpora and the scoring software provided in the framework of the DARPA/NIST HUB4 evaluation exercise (Chinchor et al., 1998). Scores (i.e. F-measure, Precision and Recall) are computed by comparing a reference tagged corpus with an automatically tagged corpus according to *type*, *content* and *extension* of each NE. Results achieved over a 365Kb test corpus of newswire texts are shown in Table 8.

	Recall	*Precision*	*F-Measure*
PERS.	73.19	73.59	73.39
LOC.	91.90	85.78	88.74
ORG.	90.15	74.84	81.79
MEAS.	92.59	57.47	70.92
MONEY	96.80	94.29	95.54
DUR.	97.92	94.00	95.92
DATE	94.29	99.00	96.59
TIME	96.80	88.98	92.89
CARD.	86.87	87.76	87.31
PERC.	97.58	91.63	94.61
ALL	87.56	82.32	84.86

Table 8: Overall Precision, Recall and F-Measure scores

The testing phase revealed an acceptable overall system's performance: as can be seen, the F-Measure score for all the tags is around 85%. This result confirms the initial working hypothesis that a WordNet-based approach to NE recognition avoids the difficulties related to the creation and maintainance of reliable gazetteers, without a great loss in terms of performance.

We believe that the gap between our system and the results achieved by rule-based systems exploiting extensive gazetteer lists can be filled following two main directions.

First, by improving the basic rules set adding new rules for some NE categories. As an example, Table 8 shows an unforeseen low result for the category Measure. In fact, even though numerical expressions are in general easier to handle, the number of constructions commonly used to express quantities was dramatically understimated. For instance, no rules were created for dealing with expressions of measure like "a three-judge panel", "a three-turnover performance", "a three-and-a-half-game", "a half-speed grounder", "a two-run double", etc. As we are currently filling this gap we expect the next evaluation of the system to give much better results.

Further improvements to the system can also be obtained by moving to WordNet 1.7, which contains a significantly larger number of NE with respect to WordNet 1.6. As an example, if we consider the concepts "philosopher" and "port", which belong to the categories Person and Location respectively, we see that WordNet 1.7 contains 81 Word_Instances for the first concept and 157 for the second, while the 1.6 version contains only 24 and 4 of such instances. It is important to notice that the performances of systems based on the use of gazetteer lists crucially depend on the availability and reliability of these resources. As an example, (Mikheev et al., 1999) reports that, for the category Location, the Precision and Recall scores of their system decrease from more than 90% to 50% without the availability of gazetteers. Even though our approach is not only focused on capturing internal evidence in the input text, the system will doubtless benefit from the availability of the large number of new Word_Instances present in WordNet 1.7.

5 Conclusion and future work

Information Retrieval and Information Extraction related tasks doubtlessly are among the most natural applications of semantic networks. An increasing number of NLP applications actually take great advantage of the highly structured WordNet taxonomy for a wide range of activities. Among these, NE recognition could benefit from the availability of WordNet-like resources for two main reasons.

First, WordNet is a free, well known, standard resource which provides a powerful alternative to the use of hand-crafted gazetteer lists. In this paper we showed how the highly structured WordNet hierarchy can be exploited to cope with the difficulty of building and maintaining comprehensive lists of proper nouns and trigger words. The approach presented is based

on the assumption that the huge number of possible trigger words that can be mined from WordNet compensates for the relatively limited availability of proper nouns. Experimental results compare well with results achieved by other rule-based systems and confirm the validity of our methodology.

Second, the recent spread of multilingual semantic networks aligned with the English version of WordNet makes the construction of multilingual NE recognition systems quite an easier challenge. In fact, even though the basic rules are language-dependent (i.e. any extension to languages other than English would require the development of a set of language-specific rules), the predicates for the extraction of proper nouns and trigger words from the WordNet hierarchy are language-independent and reusable.

In order to achieve multilinguality, we are planning to combine our English rule set with a set of Italian rules. Trigger words and entity words will be extracted from MultiWordNet (Pianta et al., 2002), a multilingual lexical database including information about English and Italian words. MultiWordNet has been developed keeping as much as possible of the semantic relations available in the English WordNet: Italian synsets have been created in correspondence with English synsets, importing semantic relations from the corresponding English synsets. The Italian part of MultiWordNet currently covers about 40,000 lemmas, strictly aligned with WordNet 1.6.

References

Alonge, A., Calzolari, N., Vossen, P., Bloksma, L., Castellon, I., Marti, M. A., and Peters, W. (1998). The Linguistic Design of the EuroWordNet Database. *Computers and the Humanities*, pages 91–115.

Chinchor, N. (1998). Overview of MUC-7. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

Chinchor, N., Robinson, P., and Brown, E. (1998). Hub-4 Named Entity Task Definition (version 4.8). Technical report, SAIC. *http ://www.nist.gov/speech/hub4_98*.

Coates-Stephens, S. (1992). *The Analysis and Acquisiton of Proper Names for Robust Text Understanding*. PhD thesis, Department of Computer Science, City University, London.

Cucchiarelli, A., Luzi, D., and Velardi, P. (1998). Automatic Semantic Tagging of Unknown Proper Names. In *Proceedings of COLING-ACL 1998*, Montreal, Canada.

Fellbaum, C. (1998). *WordNet, An Electronic Lexical Database*. The MIT Press.

Gangemi, A., Guarino, N., Oltramari, A., and Borgo, S. (2002). Restructuring WordNet's Top-Level: The OntoClean based Approach. In *Proceedings of the LREC 2002 Workshop on Ontologies and Lexical Knowledge Bases (OntoLex 2002)*, Canary Islands, Spain.

McDonald, D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. In Boguraev, I. and Pustejovsky, J., editors, *Corpus Processing for Lexical Acquisition*, chapter 2, pages 21–39. The MIT Press, Cambridge, MA.

Mikheev, A., Moens, M., and Grover, C. (1999). Named Entity Recognition without Gazetteers. In *Proceedings of the Ninth International Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, Bergen, Norway.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing an Aligned Multilingual Database. In *Proceedings of the 1st International Global WordNet Conference*, Mysore, India.

Wacholder, N., Ravin, Y., and Choi, M. (1997). Disambiguation of Proper Names in Text. In *Proceedings of Fifth Conference on Applied Natural Language Processing*, pages 202–208.

Wakao, T., Gaizauskas, R., and Wilks, Y. (1996). Evaluation of an Algorithm for the Recognition and Classification of Proper Names. In *Proceedings of the 16th Conference on Computational Linguistics (COLING'96)*, pages 418–423, Copenhagen.