A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia

Antonio Toral
University of Alicante
Carretera San Vicente S/N
Alicante 03690, Spain
atoral@dlsi.ua.es

Rafael Muñoz
University of Alicante
Carretera San Vicente S/N
Alicante 03690, Spain
rafael@dlsi.ua.es

Abstract

This paper describes a method to automatically create and maintain gazetteers for Named Entity Recognition (NER). This method extracts the necessary information from linguistic resources. Our approach is based on the analysis of an on-line encyclopedia entries by using a noun hierarchy and optionally a PoS tagger. An important motivation is to reach a high level of language independence. This restricts the techniques that can be used but makes the method useful for languages with few resources. The evaluation carried out proves that this approach can be successfully used to build NER gazetteers for location (F 78%) and person (F 68%) categories.

1 Introduction

Named Entity Recognition (NER) was defined at the MUC conferences (Chinchor, 1998) as the task consisting of detecting and classifying strings of text which are considered to belong to different classes (e.g. person, location, organization, date, time). Named Entities are theoretically identified and classified by using evidence. Two kinds of evidence have been defined (McDonald, 1996). These are internal and external evidence. Internal evidence is the one provided from within the sequence of words that constitute the entity. In contrast, external evidence is the criteria that can be obtained by the context in which entities appear.

Since the time NER was introduced, mainly two approaches have been adopted to deal with this task. One is referred as knowledge-based and uses explicit resources like rules and gazetteers, which commonly are hand-crafted. The other follows the learning paradigm and usually uses as a resource a tagged corpus which is used to train a supervised learning algorithm.

In the knowledge-based approach two kind of gazetteers can be distinguished. On one hand there are trigger gazetteers, which contain key words that indicate the possible presence of an entity of a given type. These words usually are common nouns. E.g. ms. indicates that the entity after it is a person entity. On the other hand there are entity gazetteers which contain entities themselves, which usually are proper nouns. E.g. Portugal could be an instance in a location gazetteer.

Initially, and specially for the MUC conferences, most of the NER systems developed did belong to the knowledge-based approach. This approach proved to be able to obtain high scores. In fact, the highest score obtained by a knowledge-based system in MUC-7 reached F 93.39 % (Mikheev et al., 1998). However, this approach has an important problem: gazetteers and rules are difficult and tedious to develop and to maintain. If the system is to be used for an open domain, linguistic experts are needed to build the rules, and besides, it takes too much time to tune these resources in order to obtain satisfactory results. Because of this, lately most of the research falls into the learning-based paradigm.

Regarding the creation and maintenance of gazetteers, several problems have been identified, these are mainly:

- Creation and maintenance effort
- Overlaps between gazetteers

The first problem identified assumes that the gazetteers are manually created and maintained. However, this is not always the case. Gazetteers

could be automatically created and maintained by extracting the necessary information from available linguistic resources, which we think is a promising line of future research.

Several research works have been carried out in this direction. An example of this is a NER system which uses trigger gazetteers automatically extracted from WordNet (Magnini et al., 2002) by using wordnet predicates. The advantage in this case is that the resource used is multilingual and thus, porting it to another language is almost straightforward (Negri and Magnini, 2004).

There is also a work that deals with automatically building location gazetteers from internet texts by applying text mining procedures (Ourioupina, 2002), (Uryupina, 2003). However, this work uses linguistic patterns, and thus is language dependent. The author claims that the approach may successfully be used to create gazetteers for NER.

We agree with (Magnini et al., 2002) that in order to automatically create and maintain trigger gazetteers, using a hierarchy of common nouns is a good approach. Therefore, we want to focus on the automatically creation and maintenance of entity gazetteers. Another reason for this is that the class of common nouns (the ones being triggers) is much more stable than the class of proper names (the ones in entity gazetteers). Because of this, the maintenance of the latter is important as new entities to be taken into account appear. For example, if we refer to presidents, the trigger word used might be 'president' and it is uncommon that the trigger used to refer to them changes over time. On the other hand, the entities being presidents change as new presidents appear and current presidents will disappear.

Our aim is to find a method which allow us to automatically create and maintain entity gazetteers by extracting the necessary information from linguistic resources. An important restriction though, is that we want our method to be as independent of language as possible.

The rest of this paper is structured as follows. In the next section we discuss about our proposal. Section three presents the results we have obtained and some comments about them. Finally, in section four we outline our conclusions and future work.

2 Approach

In this section we present our approach to automatically build and maintain dictionaries of proper nouns. In a nutshell, we analyse the entries of an encyclopedia with the aid of a noun hierarchy. Our motivation is that proper nouns that form entities can be obtained from the entries in an encyclopedia and that some features of their definitions in the encyclopedia can help to classify them into their correct entity category.

The encyclopedia used has been Wikipedia[1]. According to the English version of Wikipedia [2], Wikipedia is a multi-lingual web-based, free-content encyclopedia which is updated continuously in a collaborative way. The reasons why we have chosen this encyclopedia are the following:

- It is a big source of information. By December 2005, it has over 2,500,000 definitions. The English version alone has more than 850,000 entries.
- Its content has a free license, meaning that it will always be available for research without restrictions and without needing to acquire any license.
- It is a general knowledge resource. Thus, it can be used to extract information for open domain systems.
- Its data has some degree of formality and structure (e.g. categories) which helps to process it.
- It is a multilingual resource. Thus, if we are able to develop a language independent system, it can be used to create gazetteers for any language for which Wikipedia is available.
- It is continuously updated. This is a very important fact for the maintenance of the gazetteers.

The noun hierarchy used has been the noun hierarchy from WordNet (Miller, 1995). This is a widely used resource for NLP tasks. Although initially being a monolingual resource for the English language, a later project called EuroWordNet (Vossen, 1998), provided wordnet-like hierarchies

[1] http://www.wikipedia.org

[2] http://en.wikipedia.org/wiki/Main_Page

for a set of languages of the European Union. Besides, EuroWordNet defines a language independent index called Inter-Lingual-Index (ILI) which allows to establish relations between words in wordnets of different languages. The ILI facilitates also the development of wordnets for other languages.

From this noun hierarchy we consider the nodes (called synsets in WordNet) which in our opinion represent more accurately the different kind of entities we are working with (location, organization and person). For example, we consider the synset 6026 as the corresponding to the entity class Person. This is the information contained in synset number 6026:

```
person, individual, someone,
somebody, mortal,
human, soul -- (a human being;
"there was too much for one person
to do")
```

Given an entry from Wikipedia, a PoS-tagger (Carreras et al., 2004) is applied to the first sentence of its definition. As an example, the first sentence of the entry Portugal in the Simple English Wikipedia [3] is presented here:

```
Portugal portugal NN
is be VBZ
a a DT
country country NN
in in IN
the the DT
south-west south-west NN
of of IN
Europe Europe NP
. . Fp
```

For every noun in a definition we obtain the synset of WordNet that contains its first sense[4]. We follow the hyperonymy branch of this synset until we arrive to a synset we have considered belonging to an entity class or we arrive to the root of the hierarchy. If we arrive to a considered synset, then we consider that noun as belonging to the entity class of the considered synset. The following example may clarify this explanation:

```
portugal --> LOCATION
country --> LOCATION
south-west --> NONE
europe --> LOCATION
```

As it has been said in the abstract, the application of a PoS tagger is optional. The algorithm will perform considerably faster with it as with the PoS data we only need to process the nouns. If a PoS tagger is not available for a language, the algorithm can still be applied. The only drawback is that it will perform slower as it needs to process all the words. However, through our experimentation we can conclude that the results do not significantly change.

Finally, we apply a weighting algorithm which takes into account the amount of nouns in the definition identified as belonging to the different entity types considered and decides to which entity type the entry belongs. This algorithm has a constant Kappa which allows to increase or decrease the distance required within categories in order to assign an entry to a given class. The value of Kappa is the minimum difference of number of occurrences between the first and second most frequent categories in an entry in order to assign the entry to the first category. In our example, for any value of Kappa lower than 4, the algorithm would say that the entry Portugal belongs to the location entity type.

Once we have this basic approach we apply different heuristics which we think may improve the results obtained and which effect will be analysed in the section about results.

The first heuristic, called is_instance, tries to determine whether the entries from Wikipedia are instances (e.g. Portugal) or word classes (e.g. country). This is done because of the fact that named entities only consider instances. Therefore, we are not interested in word classes. We consider that an entry from Wikipedia is an instance when it has an associated entry in WordNet and it is an instance. The procedure to determine if an entry from WordNet is an instance or a word class is similar to the one used in (Magnini et al., 2002).

The second heuristic is called is_in_wordnet. It simply determines if the entries from Wikipedia have an associated entry in WordNet. If so, we may use the information from WordNet to determine its category.

[3] http://simple.wikipedia.org/wiki/Portugal

[4] We have also carried out experiments taking into account all the senses provided by WordNet. However, the performance obtained is not substantially better while the processing time increases notably.

3 Experiments and results

We have tested our approach by applying it to 3517 entries of the Simple English Wikipedia which were randomly selected. Thus, these entries have been manually tagged with the expected entity category[5]. The distribution by entity classes can be seen in table 1:

As it can be seen in table 1, the amount of entities of the categories Person and Location are balanced but this is not the case for the type Organization. There are very few instances of this type. This is understandable as in an encyclopedia locations and people are defined but this is not the usual case for organizations.

According to what was said in section 2, we considered the heuristics explained there by carrying out two experiments. In the first one we applied the is_instance heuristic. The second experiment considers the two heuristics explained in section 2 (is_instance and is_in_wordnet). We do not present results without the first heuristic as through our experimentation it proved to increase both recall and precision for every entity category.

For each experiment we considered two values of a constant Kappa which is used in our algorithm. The values are 0 and 2 as through experimentation we found these are the values which provide the highest recall and the highest precision, respectively. Results for the first experiment can be seen in table 2 and results for the second experiment in table 3.

As it can be seen in these tables, the best recall for all classes is obtained in experiment 2 with Kappa 0 (table 3) while the best precision is obtained in experiment 1 with Kappa 2 (table 2).

The results both for location and person categories are in our opinion good enough to the purpose of building and maintaining good quality gazetteers after a manual supervision. However, the results obtained for the organization class are very low. This is mainly due to the fact of the high interaction between this category and location combined with the practically absence of traditional entities of the organization type such as companies. This interaction can be seen in the in-depth results which presentation follows.

In order to clarify these results, we present more in-depth data in tables 4 and 5. These tables present an error analysis, showing the false positives, false negatives, true positives and true negatives among all the categories for the configuration that provides the highest recall (experiment 2 with Kappa 0) and for the one that provides the highest precision (experiment 1 with Kappa 2).

In tables 4 and 5 we can see that the interactions within classes (occurrences tagged as belonging to one class but NONE and guessed as belonging to other different class but NONE) is low. The only case in which it is significant is between location and organization. In table 5 we can see that 12 entities tagged as organization are classified as LOC while 20 tagged as organization are guessed with the correct type. Following with these, 5 entities tagged as location where classified as organization. This is due to the fact that countries and related entities such as "European Union" can be considered both as organizations or locations depending on their role in a text.

4 Conclusions

We have presented a method to automatically create and maintain entity gazetteers using as resources an encyclopedia, a noun hierarchy and, optionally, a PoS tagger. The method proves to be helpful for these tasks as it facilitates the creation and maintenance of this kind of resources.

In our opinion, the principal drawback of our system is that it has a low precision for the configuration for which it obtains an acceptable value of recall. Therefore, the automatically created gazetteers need to pass a step of manual supervision in order to have a good quality.

On the positive side, we can conclude that our method is helpful as it takes less time to automatically create gazetteers with our method and after that to supervise them than to create that dictionaries from scratch. Moreover, the updating of the gazetteers is straightforward; just by executing the procedure, the new entries in Wikipedia (the entries that did not exist at the time the procedure was performed the last time) would be analysed and from these set, the ones detected as entities would be added to the corresponding gazetteers.

Another important fact is that the method has a high degree of language independence; in order to apply this approach to a new language, we need a version of Wikipedia and WordNet for that language, but the algorithm and the process does not change. Therefore, we think that our method can be useful for the creation of gazetteers for lan-

[5]This data is available for research at `http://www.dlsi.ua.es/~atoral/index.html\#resources`

Entity type	Number of instances	Percentage
NONE	2822	
LOC	404	58
ORG	55	8
PER	236	34

Table 1: Distribution by entity classes

k	LOC			ORG			PER		
	prec	**rec**	$\mathbf{F}_{\beta=1}$	**prec**	**rec**	$\mathbf{F}_{\beta=1}$	**prec**	**rec**	$\mathbf{F}_{\beta=1}$
0	66.90	94.55	78.35	28.57	18.18	22.22	61.07	77.11	68.16
2	86.74	56.68	68.56	66.66	3.63	6.89	86.74	30.50	45.14

Table 2: Experiment 1. Results applying is_instance heuristic

k	LOC			ORG			PER		
	prec	**rec**	$\mathbf{F}_{\beta=1}$	**prec**	**rec**	$\mathbf{F}_{\beta=1}$	**prec**	**rec**	$\mathbf{F}_{\beta=1}$
0	62.88	96.03	76.00	16.17	20.00	17.88	43.19	84.74	57.22
2	77.68	89.60	83.21	13.95	10.90	12.24	46.10	62.71	53.14

Table 3: Experiment 2. Results applying is_instance and is_in_wordnet heuristics

Tagged	Guessed			
	NONE	**LOC**	**ORG**	**PER**
NONE	2777	33	1	11
LOC	175	229	0	0
ORG	52	1	2	0
PER	163	1	0	72

Table 4: Results fn-fp (results 1 k=2)

Tagged	Guessed			
	NONE	**LOC**	**ORG**	**PER**
NONE	2220	196	163	243
LOC	8	387	5	4
ORG	20	12	20	3
PER	30	9	2	195

Table 5: Results fn-fp (results 2 k=0)

guages in which NER gazetteers are not available but have Wikipedia and WordNet resources.

During the development of this research, several future works possibilities have appeared. Regarding the task we have developed, we consider to carry out new experiments incorporating features that Wikipedia provides such as links between pairs of entries. Following with this, we consider to test more complex weighting techniques for our algorithm.

Besides, we think that the resulting gazetteers for the configurations that provide high precision and low recall, although not being appropriate for building gazetteers for NER systems, can be interesting for other tasks. As an example, we consider to use them to extract verb frequencies for the entity categories considered which can be later used as features for a learning based Named Entity Recogniser.

Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the Valencia Government under project number GV04B-268.

We also would like to specially thank Borja Navarro for his valuable help on WordNet.

References

X. Carreras, I. Chao, L. Padró, and M. Padró. 2004. Freeling: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th LREC Conference*.

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

B. Magnini, M. Negri, R. Preete, and H. Tanev. 2002. A wordnet-based approach to named entities recognition. In *Proceedings of SemaNet '02: Building and Using Semantic Networks*, pages 38–44.

D. McDonald. 1996. Internal and external evidence in the identification and semantic categorization of proper names. *Corpus Processing for Lexical Aquisition*, pages 21–39, chapter 2.

A. Mikheev, C. Grover, and M. Moens. 1998. Description of the LTG system used for MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference held in Fairfax, Virginia, 29 April-1 May*.

G. A. Miller. 1995. Wordnet: A lexical database for english. *Communications of ACM*, (11):39–41.

M. Negri and B. Magnini. 2004. Using wordnet predicates for multilingual named entity recognition. In *Proceedings of The Second Global Wordnet Conference*, pages 169–174.

O. Ourioupina. 2002. Extracting geographical knowledge from the internet. In *Proceedings of the ICDM-AM International Workshop on Active Mining*.

O. Uryupina. 2003. Semi-supervised learning of geographical gazetteers from the internet. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 18–25.

P. Vossen. 1998. Introduction to eurowordnet. *Computers and the Humanities*, 32:73–89.