The Multilingual Named Entity Recognition Framework

Thierry Poibeau and the INaLCO Named Entity Group[1]
INaLCO/CRIM
2 rue de Lille
75007 Paris

Abstract

This paper presents a multilingual system designed to recognize named entities in a wide variety of languages (currently more than 12 languages are concerned). The system includes original strategies to deal with a wide variety of encoding character sets, analysis strategies and algorithms to process these languages.

1 Introduction

Since the MUC conferences about Information Extraction, named entity recognition (NERC) is a well-established task in the NLP community (MUC-6, 1995). Examples of named entities are person names, location and company names, date and time indications, etc. A lot of systems have been developed to perform this task, ranging from manually created rule-based systems to fully automatic learning-based systems. We will shortly present these technologies below.[1]

Even if a lot of systems have been developed for languages such as English or Japanese, a large range of languages do not have access to such a technology. We propose an open framework to develop resources and tools for named entity recognition. A team of computational linguist students develops this project[1], so that it also has pedagogic purposes. But, even so, the project seems to be sufficiently attractive to interest industrial partners.

We describe the different approaches for named entity recognition. We then present the project and the different analysis techniques used. We will conclude with some considerations on evaluation and future work.

2 State of the art NERC systems

In this section, we examine the different approaches to named entity recognition. We then examine previous experiments to compare systems and techniques. Sekine and Eriguchi (2000) present an interesting classification of named entity recognition systems.

- **Manually created rule-based systems.** In this kind of system, developers initially elaborate a set of patterns that will be applied on the text to accurately recognize and tag named entities. Nearly all classical MUC systems were using this approach until the mid'90s, and most of them are still using this kind of technique (MUC-6, 1995).
- **Fully automatic learning-based systems.** These systems are using Machine Learning (ML) techniques to learn a model in order to accurately tag the texts. The result of the learning task can be a set of rules, a decision tree or a set of numeric data. Note that a human cannot always revise the result if the learning algorithm used does not provide a readable output. These systems are now very popular in the IE community (Bikel *et al.*, 1997) (Collins and Singer, 1999), even if they were initially rather dedicated to audio corpora.

[1] The members of the INaLCO Named Entity Group are: A. Acoulon, C. Avaux, L. Beroff-Bénéat-, A. Cadeau, M. Calberg, A. Delale, L. De Temmerman, A.-L. Guenet, D. Huis, M. Jamalpour, A. Krul, A. Marcus, F. Picoli and C. Plancq.

- **Mixed approach.** In this kind of systems, a set of rules is automatically learned and revised by an expert. An alternative can be the dynamic extension of an existing set of core rules previously defined by the expert, so that the system obtains a better coverage of the data. Cucchiarelli and Velardi (2000), among others, have applied this approach to NERC systems.

3 Multilingual named entity recognition

We are currently developing resources and tools for the following languages: Arabic, Chinese, English, French, German, Japanese, Finnish, Malagasy, Persian, Polish, Russian, Spanish and Swedish.

3.1 Multilingualism issues

These languages vary a lot in their characteristics, in their writing systems as much as in their grammar. Moreover, language technology is not much developed for most of them. This has a big consequence for named entity recognition: for certain languages like most of the European languages, we benefit from already existing lexical resources. For other languages, a lot of work still needs to be done. For example, there is no dictionary available for Malagasy and even electronic resources and corpora are rare.

All the texts and resources are encoded using the Unicode standard (Unicode Little-Endian). This strategy allows most of the encoding problems to be solved, even if some bugs still remain from time to time for a given language (for example, writing direction problems in Arabic, when characters appears from the left to the right, while it should be the contrary, etc.).

3.2 Overall system architecture

In spite of differences in their implementation, each system shares approximately the same architecture. The text is firstly analyzed by a classical rule-based system. This analysis is then completed by dynamic acquisition mechanisms (theory learning) and revision capabilities (see Figure 1).

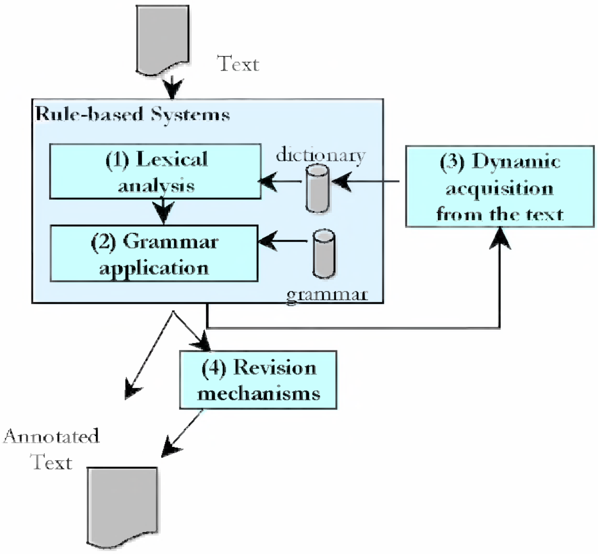

Figure 1: Architecture of the system

We detail below these 4 main knowledge sources:

- **Gazetteers**. Their role is disputed since the appearance of ML techniques allowing previously unknown named entities to be acquired from tagged corpora. However, it is simply, most of the time, not realistic to tag large amount of corpus (Appelt and Israel, 1999). Moreover, tagging great amounts of data can be compared to the elaboration of dictionaries[2].

- **Grammar.** Its aim is to group together elements pertaining to the same entity. A grammar rule is generally made of a trigger word, some tagged words and occasionally unknown words. These words can be accurately tagged given an appropriate context (especially if a trigger word disambiguates the sequence).

- **Learning capabilities.** We include, in this section, ML algorithms used to tag unknown named entities. Most ML techniques have been

[2] If one analyzes a text to tag person names, it is then easy to write a simple program that will automatically extract the sequences previously tagged to generate a dictionary. In this sense, tagging is not that different from elaborating a dictionary!

used including maximal entropy, inductive logic programming, decision tree learning, hidden Markov models and others (Bechet *et al.*, 2000) (Bikel *et al.*, 1997) (Collins and Singer, 1999) (Mikheev *et al.*, 1999). We use a kind of theory learning to extend the set of expressions identified by the rule-based system: the lexicon and the grammar is exploited as a domain theory to dynamically find new entities (Mooney, 1993).

- **Revision capabilities.** We implemented revision capabilities in the system so that it can revise tags in a certain context. For example, in an English text, isolated occurrences of *Washington* can be considered as location names. If one finds a context that potentially suggests another category for the named entity (for example, *Mrs. Washington*) the system will revise the initial tag and put the new category on the concerned word (isolated occurrences of *Washington* will be tagged as person names).

3.3 Implementation

Rule-based systems have been developed for English and French using the INTEX/UNITEX finite state toolbox (Silberztein, 1993). The resulting system has been described in (Poibeau and Kosseim, 2001). Resources are currently being defined and adapted to other languages like Russian (Cyrillic alphabet) or Arabic and Persian (Arabic writing system).

For Asian languages, like Japanese, which makes use of 4 different writing systems (hiragana, katakana, kanji and romanji), the INTEX/UNITEX was not efficient. Thus, Japanese is processed at first by the CHASEN morphological analyser (Asahara and Matsumoto, 2000). Perl scripts are then applied on top of the CHASEN analysis to produce a tagged text with highlighted named entities. Even if the CHASEN analyser uses the JIS format, the final output is encoded using the Unicode standard.

Once the system is adapted, the same strategy is adapted to the different languages. A set of trigger words is defined, along with a proper names dictionary and a named entity grammar for the concerned language. The dynamic named acquisition mechanisms implemented are classical and have been described with details in (Poibeau and Kosseim, 2001).

3.4 Resource sharing

While developing the system for different Indo-European language, we saw that resources could be shared by different languages. For example, proper name dictionaries for French and English are very similar. One has just to remove entries from the English dictionary that would be too ambiguous in French. A large part of the grammar can also be re-used provided that the grammar rules are carefully cheked and appropriate modifications are made (list of trigger words, etc.). Of course, these resources must be completed to properly cover the new language and/or the new domain.

The same approach seems to be valid for other romance languages (Italian, Spanish). For Germanic and Slavic languages, dictionaries must be modified to take into account inflectional forms. A large amount of work is then needed to modify and adapt dictionaries firstly developed for English (add an inflectional code on each word; This code is language-dependent). The approach has not been investigated for non Indo-European languages.

4 Evaluation

The system is under implementation. A complete evaluation is then impossible but we present in this section some first results.

4.1 Overall performances

For the moment, only the English and the French systems have been intensively tested. Their performance is comparable to systems having participated to MUC conferences (P&R is the combined value of precision and recall).

	Recall	Precision	P&R
BBN	.98	.98	.98
SRA	.97	.99	.98
NYU	.94	.99	.96
U. Sheffield	.84	.96	.90
Our system	.86	.95	.90

Figure 2: Performances on the MUC-6 corpus

Their performance has also been tested on different corpora and it appears that these hybrid systems are less sensitive to corpus or domain changes than classical rule-based systems (Poibeau and Kosseim, 2001).

4.2 Other experiments

The developed systems are systematically tested on the *Monde Diplomatique* corpus (when available!), a multilingual international journal published in 10 languages on the web. We hope to achieve for most of the other languages under implementation better or similar results to the ones obtained for French and English. This multilingual named entity recogniser is already used in a wider project concerning corpus alignment. The idea is to use cognates and named entities as cues for sentence alignment.

5 Conclusion

This paper presented a multilingual framework for named entity recognition. More than 12 languages are currently under development with very encouraging results. This project will produce stand-alone applications as well as modules for sentence alignment and cognate identification in parallel corpora using different character sets and writing systems.

6 References

Appelt D. and Israel D. (1999). Introduction to Information Extraction Technology. (IJCAI-99) Tutorial, Stockholm, Sweden (available at: http://www.ai.sri. com/~appelt/ie-tutorial/)

Asahara M., Matsumoto M. (2000) Extended Models and Tools for High-performance Part-of-Speech Tagger". In *Proceedings of Coling'2000*, Saarbrücken, Germany, pp. 21—27.

Bechet F., Nasr A., Genet F. (2000) Tagging Unknown Proper Names Using Decision Trees. In *Proceedings of the 38th ACL Conference*, Hong-Kong, pp. 77—84

Bikel D., Miller S., Schwartz R. and Weischedel R. (1997) Nymble: a high performance learning name-finder. In *Proceeding of the 5th ANLP Conference*, Washington, USA.

Borthwick A. (1999) *A maximum entropy approach for named entity recognition.* PhD Thesis, New York University.

Collins M. and Singer Y. (1999) Unsupervised models for named entity classification. In *Proceedings of EMNLP/WVLC*, 1999, MA, pp. 189—196.

Cucchiarelli A. and Velardi P. (1999) Adaptability of linguistic resources to new domains: an experiment with proper noun dictionaries. In *Proceedings of the Vextal Conference*, Venice, Italy, pp. 25—30.

Mikheev A., Moens M. and Grover C. (1999) Named Entity recognition without gazetteers. In *Proceedings of the Annual Meeting of the European Association for Computational Linguistics* EACL'99, Bergen, Norway, pp. 1—8.

Mooney R. (1993) Induction over the unexplained: using overly general domain theories to aid concept learning, *Machine Learning*, 10:79.

MUC-6 (1995) *Proceedings of the Sixth Message Understanding Conference* (DARPA), Morgan Kaufmann Publishers, San Francisco.

Poibeau T and Kosseim L. (2001) Proper-name Extraction from Non-Journalistic Texts. *Proceeding of the 11th Conference Computational Linguistics in the Netherlands*, Tilburg. Netherlands, Rodopi.

Sekine S., Eriguchi Y. (2000) Japanese Named Entity Extraction Evaluation - Analysis of Results. In *Proceedings of Coling'2000*, Saarbrücken, Germany, pp. 1106—1110.

Silberztein M. (1993) *Dictionnaires électroniques.* Masson, Paris.

Yarowsky D. (1995) Unsupervised Word Sense Disambiguation rivaling Supervised Methods. In *Proceedings of the 33rd ACL Conference*, Cambridge, USA.