Multilingual Databases for Named Entity Recognition

Augustin Delale, Delphine Huis, Cyrille Robert,
Susan Schneider and Thierry Poibeau
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.

The paper focuses more particularly on the Japanese and French parts of the system, which include original strategies for named entity analysis. We detail the overall architecture and specific developments for Japanese. We give in annex lists of trigger words for both languages.

1 Introduction

Since the Message Understanding Conferences (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.

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, so that it also has pedagogic purposes. But, even so, the project seems to be sufficiently attractive to interest industrial partners.

1 The members of the INaLCO Named Entity Group are: A Accoulon, 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. Plançois.
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 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.

2.1 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 appear from the left to the right, while it should be the contrary, etc.).

2.2 Overall system architecture

The system is organized around a multilingual database giving access to resource for named entities in a large variety of languages. This database can easily be extended to other languages if the new resources respect the database structure and XML tags used to encode linguistic information. Linguistic resources for a given language are automatically derived from the general database.

Figure 1 gives an overview of the overall architecture.
Resources are mainly made of:

- Gazetteers. Their role is disputed since the appearance of Machine Learning 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 texts (Appelt and Israel, 1999). Moreover, tagging large amounts of data can be compared to the elaboration of dictionaries.
- 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).

2.3 Implementation

Rule-based systems have been developed for English and French using the Intex/Unitex finite state toolbox (Silberstein, 1993). The resulting system has been described in (Poièbeau 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 romaji), 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.

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!
The following schema presents the architecture for the Japanese system.

![Diagram of the Japanese system architecture](image)

**Figure 2:** System architecture for Japanese

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).

### 2.4 Resource sharing

While developing the system for different Indo-European languages, 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 checked 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.

3 Evaluation

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

3.1 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 (F is the harmonic combined value of precision and recall).

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBN</td>
<td>.98</td>
<td>.98</td>
<td>.98</td>
</tr>
<tr>
<td>SRA</td>
<td>.97</td>
<td>.99</td>
<td>.98</td>
</tr>
<tr>
<td>NYU</td>
<td>.94</td>
<td>.99</td>
<td>.96</td>
</tr>
<tr>
<td>U. Sheffield</td>
<td>.84</td>
<td>.96</td>
<td>.90</td>
</tr>
<tr>
<td>Our system(^3)</td>
<td>.86</td>
<td>.95</td>
<td>.90</td>
</tr>
</tbody>
</table>

Figure 3: 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 Koseim, 2001).

3.2 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.

\(^3\) There were about 15 systems at MUC-6. However, our results are not strictly comparable since we benefited from resources developed for this evaluation that are available on the web.
4 Conclusion and perspectives for multilingual NLP: the case of the Papillon project

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 co-identify identification in parallel corpora using different character sets and writing systems.

The Papillon database is intended to produce resources for a large variety of languages, especially Asian ones. However, one knows that for most industrial applications (for example, information filtering and extraction), named entities are crucial but rarely taken into account. Thus this project could be an important contribution to complete the general linguistic knowledge contained in the Papillon databases.

The project is described at http://crimlangueso.asso.fr/realisations/entites/frameEN.html. These resources are freely available for research purposes.

5 References


ANNEX: Resources for named entity recognition in Japanese

This annex presents a part of the database elaborated for Japanese (together with the Hepburn transcription and a tentative French translation).

**Person names**

長 [chō] chief, and its derivatives:

- 社長 [shachō] *PDG*
- 部長 [buchō] *directeur général*
- 課長 [kachō] *chef de service*
- 学長 [gakuchō] *proviseur*
- 議長 [gichō] *président (d'un comité)*
- 総長 [sōchō] *secrétaire-général*

師 [shi] professor, and its derivatives:

- 教師 [kyōshi] *professeur*
- 技師 [gishī] *ingénieur*
- 牧師 [bokushī] *pasteur*

手 [shū] expert, and its derivatives:

- 選手 [senshu] *sportif*
- 歌手 [kashū] *chanteur*

Diverse:

- 大統領 [daitōrynō] *président (vie politique)*
- 総理大臣 [sōri daijin] *premier ministre*
<table>
<thead>
<tr>
<th>Term</th>
<th>Pronunciation</th>
<th>French Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>首相</td>
<td>[shushô]</td>
<td>premier ministre</td>
</tr>
<tr>
<td>大臣</td>
<td>[daizin]</td>
<td>ministre</td>
</tr>
<tr>
<td>サン ou さん</td>
<td>[san]</td>
<td>Monsieur ou Madame</td>
</tr>
<tr>
<td>サマ ou 様</td>
<td>[sama]</td>
<td>Monsieur ou Madame</td>
</tr>
<tr>
<td>夫人</td>
<td>[fujin]</td>
<td>Madame</td>
</tr>
<tr>
<td>氏</td>
<td>[shi]</td>
<td>Monsieur</td>
</tr>
<tr>
<td>委員</td>
<td>[i.in]</td>
<td>membre d’un groupe</td>
</tr>
</tbody>
</table>

**Trigger word for location names:**

<table>
<thead>
<tr>
<th>Term</th>
<th>Pronunciation</th>
<th>French Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>内</td>
<td>[nai]</td>
<td>à l’intérieur de, dans</td>
</tr>
<tr>
<td>周辺</td>
<td>[shūhen]</td>
<td>aux alentours de</td>
</tr>
<tr>
<td>地域</td>
<td>[chi.iki]</td>
<td>région</td>
</tr>
<tr>
<td>地方</td>
<td>[chihô]</td>
<td>localité</td>
</tr>
<tr>
<td>圏</td>
<td>[ken]</td>
<td>étendue</td>
</tr>
<tr>
<td>諸国</td>
<td>[shokoku]</td>
<td>groupement de pays</td>
</tr>
<tr>
<td>出身</td>
<td>[shusshin]</td>
<td>né à</td>
</tr>
<tr>
<td>県</td>
<td>[ken]</td>
<td>préfecture</td>
</tr>
<tr>
<td>都</td>
<td>[to]</td>
<td>métropole ; capitale</td>
</tr>
<tr>
<td>府</td>
<td>[fū]</td>
<td>agglomération</td>
</tr>
<tr>
<td>市</td>
<td>[shi]</td>
<td>ville</td>
</tr>
<tr>
<td>区</td>
<td>[ku]</td>
<td>arrondissement</td>
</tr>
<tr>
<td>郊外</td>
<td>[kōgai]</td>
<td>banlieue</td>
</tr>
<tr>
<td>町</td>
<td>[chô, machi]</td>
<td>quartier ; ville ; arrondissement</td>
</tr>
<tr>
<td>郡</td>
<td>[gun]</td>
<td>canton</td>
</tr>
<tr>
<td>通り ou 通り</td>
<td>[doori]</td>
<td>avenue</td>
</tr>
<tr>
<td>国道</td>
<td>[kokudô]</td>
<td>route nationale</td>
</tr>
<tr>
<td>道路</td>
<td>[dôro]</td>
<td>rue, route</td>
</tr>
<tr>
<td>島</td>
<td>[shima, tô]</td>
<td>île</td>
</tr>
<tr>
<td>海</td>
<td>[kai]</td>
<td>mer</td>
</tr>
<tr>
<td>洋</td>
<td>[yō]</td>
<td>océan</td>
</tr>
<tr>
<td>川</td>
<td>[kawa, gawa]</td>
<td>rivièrre</td>
</tr>
</tbody>
</table>
河 [ka, ga] rivièr
湖 [ko] lac
海峡 [kaikyō] détroit
海岸 [kaigan] côte, littoral
湾 [wan] baie, golfe
駅 [eki] gare
山 [san] montagne
山脈 [sanmyaku] chaîne de montagnes

Trigger word for company names:
(株) [kabu] abréviation de 株式会社 [kabushikigaisha]
株式会社 [kabushiki gaisha] société par actions (S.A.R.L.)
会社 [kaisha] société, ainsi que ses composants
航空会社 [kōkū gaisha] compagnie aérienne
省 [shō] ministère
会 [kai] cercle
教会 [kyōkai] église
学校 [gakkō] école, ainsi que ses composants
小学校 [shōgakkō] école primaire
中学校 [chūgakkō] collège
高等学校 [kōtōgakkō] lycée
大学 [daigaku] université, ainsi que ses composants
大学院 [daigakuin] troisième cycle
研究所 [kenkyūjo] laboratoire
工場 [kōjō] usine
大使館 [taishikan] ambassade
党 [tō] parti politique
派 [ha] aile (vie politique)
墓地 [bochi] cimetière
ホテル [hoteru] hôtel
軍 [gun] armée
[kikan] organisme, institution

**Trigger word for dates:**

- 世紀 [séki] siècle
- 年 [nen] année
- 年代 [nen.dai] « période »
- 月 [getsu] ou [gatsu] mois
- 日 [nichi] ou [ka] jour
- 曜日 [yōbi] jour de la semaine
- 昭和 [shōwa] (1926–1989) ère Shōwa
- 大正 [taishō] (1912–1926) ère Taishō
- 明治 [meiji] (1868–1912) ère Meiji