Rule-Based Named Entity Extraction For Ontology Population

Aurore de Amaral
LIASD, EA 4383, Université Paris 8
aurore.de-amaral@etud.univ-paris8.fr

Abstract

Currently, Text analysis techniques such as named entity recognition rely mainly on ontologies which represent the semantics of an application domain. To build such an ontology from specialized texts, this article presents a tool which detects proper names, locations and dates from texts by using manually written linguistic rules. The most challenging task is to extract not only entities but also interpret the information and adapt in a specific corpus in French.

Keywords

named entity extraction, information retrieval, ontology population

1 Introduction

Information extraction is fundamental since a wide variety of texts were digitized and created through Web. In this area, ontology learning is a good option to provide such information and efficiently share conceptualizations with experts and researchers. Due to this environment, it is crucial to do an efficient extraction in the texts. People and their relationships as well as locations, dates and domain terms must be discovered to create (Aussenac-Gilles et al., 2000) or complete an ontology (Magnini et al., 2006).

Because of the quantity of textual data to analyze and the continuous evolution of information (Reymonet, 2008), the extraction step should be automatically processed as much as possible. Extraction of named entities (NE) is one of the first task in ontology learning because they represent persons, names, locations and are unambiguous (Mondary, 2011). They are related to noun names like primarly defined in the MUC[1] conferences (Chinchor, 1997) and are an important part of the information retrieval domain.

This paper describes a mining method of named entities for improving the search in annotated corpora. It uses linguistic rules and lexicons. It is a qualitative method, for which the use of quantitative elements may optimize the number of results. This is the first part of an ontology learning architecture which transforms raw text data in a semantic network. From the network, a final ontology will be built, extended or populated, which will not be explained in this paper. We focus on information extraction, named entity recognition.

In section 2, the corpus that we used is described. In section 3, we present a state of art in named entity extraction. The proposed approach is exposed in section 4. In section 5, we evaluate our method of extraction and discuss it. Finally, we conclude and suggest some perspectives.

2 Domain Based Corpus

The corpus used is a digitized french dictionary, *Le dictionnaire de la Spiritualité* (the Dictionary of Spirituality), published by Éditions Beauchesne (Paris). It is an encyclopedia used by researchers in religious studies and Divinity. With more than ten thousand articles spread over a dozen volumes, it studies all the actors of Christianity. Historical events are widely represented and are a huge source of knowledge. That is why it is a reference work for all students interested in religious history of Christianity and more broadly for all historians.

The encyclopedia contains a set of entries related to other books via a number of bibliographic references that can be found at the end of each entry. Each reference contains names, places and dates.

[1] MUC: Message Understanding Conference

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 58–62,
Hissar, Bulgaria, 9-11 September 2013.

3 Named Entity Extraction

Currently, The systems evaluated in MUC (Poibeau, 2011) or ESTER 2 (Galliano et al., 2009) campaigns produce good results in named entity extraction, especially in newspaper articles. But the ease of use of these systems are rarely evaluated (Marrero et al., 2009), although it is important to use them at the beginning of an information extraction system.

3.1 Different Approaches

The challenges in NE recognition are found in the issue of the definition of the named entities. With the first MUC evaluation campaigns, the point was to detect persons, organization, locations, dates and numbers (ENAMEX, TIMEX and NUMEX (Chinchor, 1997)). Later, the definition of a named entity has included other categories (e.g, business concepts, also called "specfic interest entity" (Dutrey et al., 2012)) : issues involved recognition and categorisation of entities, with disambiguisation of homonymy and metonymy.

Two main approaches exist in NE extraction : linguistic approach (also called symbolic approach) (Ben Hamadou et al., 2010; Poibeau, 2003) and a statistical approach (Favre et al., 2005). The two approaches ensure satisfying results, the second one particularly on speech systems (Poibeau, 2011). The results tend to improve the precision without changing the recall of the first algorithms.

3.2 Lexicons

Our choice is to add lexical entries to expand the global lexicon. This lexicon is created with ontology concept names and their synonyms found in a dictionary on the Web[2]. Thus, the detection of people roles and locations is improved by applying lexico-syntactic rules. The method is really relevant and domain-dependent. However, the learning process admits the creation of new concept names during the searching step.

3.3 NLP Tools

In order to help this term extraction step, a natural language processing platform may be used. In (Poibeau, 2003), the author uses SYNTEX to create the grammar rules. We have chosen NooJ, which proposes syntactic parser to process and represents all types of linguistic units (Silberztein, 2009). This system is also able to show transformational analysis and export them. Finally, the ease of use with a graphical user interface tend to help the evolution of the system. In the next section, all the steps of the NE recognition method will be detailed.

[2] http://www.crisco.unicaen.fr/des/

4 The Proposed Approach

4.1 Lexicon Data

First, a lexicon of french cities and european countries is created. Then, a lexicon of religion domain is created. This lexicon is based on an ontology, which represents religion and other concepts validated by an expert. Classes' leaves and individuals are used to create entries. The parent's classes are used to add a semantic annotation to them. Then, morphological structures like inflectional paradigms may be manually written, for instance french plurals.

The concept names create a general lexicon. The search for synonyms of the same grammatical category automatically adds new entries to the lexicon. Without these new entries, the lexicon contains 63 entries. NooJ adds plural ones and the total is 110. There is an exemple of the NooJ dictionary showed below of french inflexional plural, when suffixes "al" become "aux" in plural :

```
cardinal,cardinal,
N+Hierarchical
+FLX=Cheval+m+s
cardinaux,cardinal,
N+Hierarchical
+FLX=Cheval+m+p
```

4.2 Syntactic Label Rules

The second step consists in manually creating global rules to delimit the main NE in the text: proper names, hierarchical names, dates, places of worship and cities. With this basic information, it will be easier to understand the relationships between actors and events. The transformational analysis shows what and where it is more suitable to annotate. Then, it shows all exceptions of predetermined rules. Tokens frequences and concordances are some of the examples of the tools the NooJ platform can perform.

The corpus contains 9466 different tokens. There is 50 entries (some of them are blank). After a syntactic parsing, named entity rules are applied. Since NooJ cannot disambiguate frequent words, a

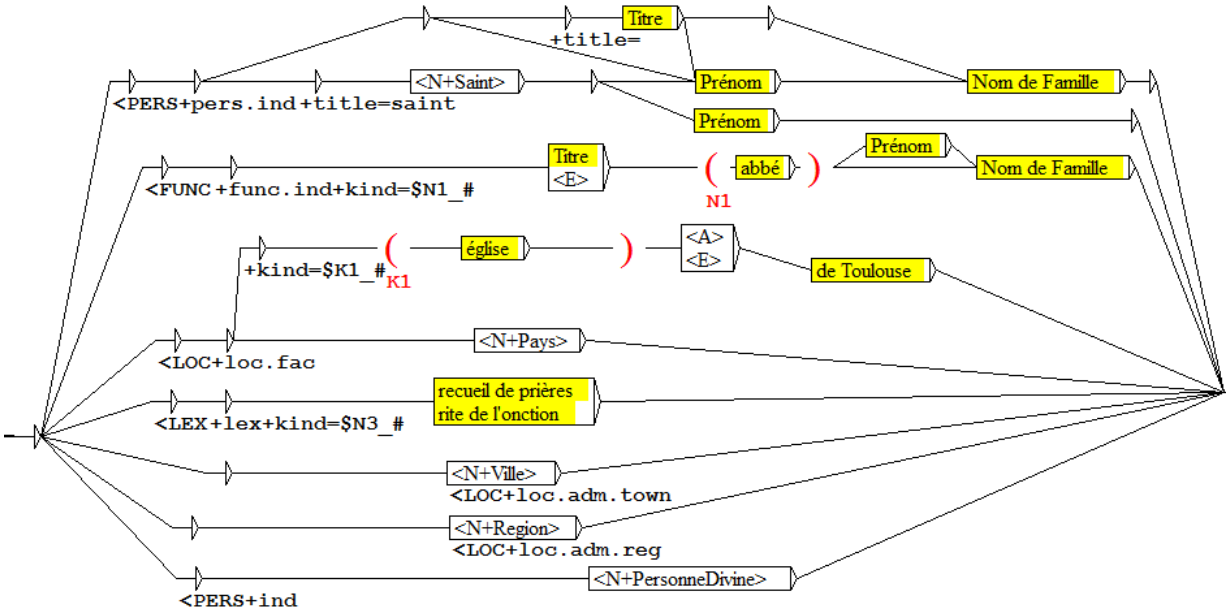

Figure 1: Main graph in NooJ

small grammar is used to identify the french grammatical more used words. There is a grammar for dates, proper names (PN) and places. The main graph in NooJ is shown in Figure 1.

The gender of the names can be identified with forenames or roles. Locations or patronyms can induce cities or place names. The nominal groups, which contain lexicon words, are also annotated. The number of results of the method is presented in Table 1.

PN	Locations	Dates
1016	985	1198

Table 1: Number of annotated text

There are different kinds of annotated proper names. The first ones and the most representative are general patronyms. A general rule distinguishes patronyms by a unique forename, which already exists in NooJ's lexicon forename or a list of uppercase words with dash followed by an other general rule for french surnames.

One of the surname's rules detects an uppercase word which may contain *de* followed by another uppercase word. *de* (of) is a french preposition which is often found in proper nouns, and can also describe functions and roles. 48 different patronyms contain a french abbreviation title (*M.* for mister, *Mme* for madam, *Melle* for unmarried womens and *Mgr* for an honorific), but other titles could be added. The number of results is shown in Table 2. Names preceded by the word *Saint* or words like priest point out a name and a religious function or a job. The compound names designates persons by their roles and not their names.

Patronyms	With functions	"saint"
832	98	86

Table 2: Number of different kinds of recognized persons

The search for locations like places of worship may identify towns, even if the lexicon of towns does not contain them. In general cases, a noun, which designates a location defined in the dictionary is followed by *de* and a first uppercase word. This uppercase word is a country or a city. A dictionary of towns and regions of France is used to disambiguate these relations.

Then, absolute dates and some kinds of relatives dates are found. There are a lot of occurrences of years.

4.3 Markup Export For The Ontology

The NooJ export file like shown in Table 3 contains several lines. This file is treated like a CSV file. The first information is the entry of the encyclopedia where the entity was found, then the entity surrounded by his left and right context. Each entity have markup tags. The markup tags used in our context take into account the general guide-

lines of Quaero (Rosset et al., 2011). These guidelines extend the first ones for named entities defined in MUC (Chinchor, 1997). Proper names have the *pers.ind* tag, people's function *func.ind*, locations *loc.adm.town* for towns and *loc.fac* for countries and general places. Then, dates have the *time.date* tag.

du bourg Verbe incarné, récemment rétabli a **Azérables**/LOC+loc.adm.town , elle est retenue à Limoges
du bourg ruction de la jeunesse. Elle expose ces faits l' **évêque de Limoge**s /FUNC+func.ind+kind=évêque +loc.adm.town=Limoges et lui communique son projet. Celui-ci approuve
du bourg Saint-Sacrement. Sa première communion, **le 24 juin 1800**/DATE+time.date.abs +year=1800+day=24+month=juin+year=1800 , lui laissera un souvenir qui

Table 3: Results with Quaero markup

5 Evaluation

For the system evaluation, a new corpus was created with three random articles to compare human and rule-based annotations. The evaluation results are shown in Table 4. We use F-measure which measures relevant results. Some improvements could be made by detecting more locations and adding more lexicon entries. There are 6 redundant results due to ambiguous surnames detected with NooJ. So, we could improve the proper names detection rules to eliminate some ambiguous answers and add roles in the lexicon.

	Persons	**Locations**	**Dates**
recall	0,64	0,53	0,95
precision	0,94	0,79	1
F-mesure	76%	63%	97%

Table 4: Evaluation results

6 Conclusion

The first step of the creation of an ontology learning architecture is information extraction. For this purpose, we choose to detect named entities because of the relative monosemic representation in text. Our tool uses rule-based methods and lexicons, partially created automatically with synonyms, applied on a domain-dependent corpus. The results are moderate with a good precision and relatively good performance for dates. Some improvements will be applied, especially with the detection of proper names without change the lexicons. Relations between all of this information and a parsing of bibliographic entries is the next step before the ontology learning process.

References

Nathalie Aussenac-Gilles, Brigitte Biébow, and Sylvie Szulman. 2000. Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In *Actes de la 9e Conférence Francophone d'Ingénierie des Connaissances IC 2000*. Université Paul Sabatier.

Abdelmajid Ben Hamadou, Odile Piton, and Héla Fehri. 2010. Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform.

Nancy Chinchor. 1997. Muc-7 named entity task definition `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html`.

Camille Dutrey, Chloé Clavel, Sophie Rosset, Ioana Vasilescu, and Martine Adda-Decker. 2012. Quel est l'apport de la détection d'entités nommées pour l'extraction d'information en domaine restreint ? In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 359–366. ATALA/AFCP.

Benoît Favre, Frédéric Béchet, and Pascal Nocéra. 2005. Robust named entity extraction from large spoken archives. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 491–498. Association for Computational Linguistics.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech 2009*.

Bernardo Magnini, Emanuele Pianta, Octavian Popescu, and Manuela Speranza. 2006. Ontology population from textual mentions: Task definition and benchmark. In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. Association for Computational Linguistics.

Monica Marrero, Sonia Sanchez-Cuadrado, Jorge Morato, and Yorgos Andreadakis. 2009. Evaluation

of Named Entity Extraction Systems. In *Research In Computer Science*, volume 41, pages 47–58. Centro de Investigación en Computación del IPN.

Thibault Mondary. 2011. *Construction d'ontologies à partir de textes. L'apport de l'analyse de concepts formels*. Ph.D. thesis, Université Paris 13 - LIPN.

Thierry Poibeau. 2003. *Extraction automatique d'information : Du texte brut au web sémantique*. Lavoisier.

Thierry Poibeau. 2011. *Traitement automatique du contenu textuel*. Lavoisier.

Axel Reymonet. 2008. *Modélisation de connaissances à partir de textes pour une recherche d'information sémantique*. Ph.D. thesis, Université Paul Sabatier.

Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. 2011. *Entités nommées structurées: guide d'annotation Quaero*. LIMSI-Centre national de la recherche scientifique.

Max Silberztein. 2009. Syntactic parsing with NooJ.