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

To cite this version:

HAL Id: hal-00547940
https://hal.science/hal-00547940
Submitted on 17 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform

Abdelmajid Ben Hamadou (1) Odile Piton (2) and Héla Fehri (3)

(1) MIRACL-University of Sfax, Tunisia. 
abdelmajid.benhamadou@isimsf.rnu.tn
(2) SAMM-University of Paris I Pantheon-Sorbonne, Paris, France.
(2) MIRACL-University of Franche-Comte and University of Sfax, Tunisia
hela.fehri @ fss.rnu.tn

Abstract

The extraction of relation between Named Entities (NE) has become the last few years an interesting research domain. It is very useful for many applications such as Web mining, Information extraction and retrieval, Business intelligence, Automatic databases filling with Entities & types, Questions answering task and document Summarization.

Several works has been performed for relation discovery in texts written in Latin languages and as far as we know, very few works has been done for Arabic language. In this paper, we focus on functional relations between ENAMEX and ORG Arabic Named Entities. The extraction approach is rule based and the implementation is performed using NooJ Platform.

Keywords: Relations between Named Entities, Extraction Process, NooJ transducers, Relation translation

1. Introduction

The extraction of semantic relations between Named Entities (NE) has become the last few years an interesting research domain. It is very useful for many applications such as Web mining, Information extraction and retrieval, Automatic databases filling with Entities & types, Questions answering task and document Summarization.

Relations between Named Entities can be binary (involving two entities) or more complex up to notion of event. There are also several types of relations based on the types of the involved Named Entities.

Several works related to semantic relations discovery has been performed for European languages and especially for English. As far as we know, very few works has been done for Arabic language.

In this paper we propose a rule based relation extraction system for Arabic language. We focus on functional relations between ENAMEX and ORG Named Entities (director, responsible, president…). Functional Relation can be explicit or implicit. Explicit relation is indicated by a special word or sequence of words in the text. Implicit relation is a relation that can be mined from the text using the context.

The extraction process is performed in three steps: Identification of the Named Entities (ENAMED and ORG), Detection of relations between identified Named Entities, Generation of the predicate form representing the relation in Arabic and French. The translation of the relation allows cross lingual information retrieval.

The rest of the paper is organized as follows. We begin by giving an idea about the proposed approaches for extraction of semantic relations between NEs. Then, we address the major challenges posed by extraction of semantic relations between NEs for
Arabic language. After that, we detail our approach and its implementation using NooJ linguistic platform [Silberstein 2005]. Finally, we present the evaluation results obtained from a test corpus.

2. Approaches for NE relation Discovery

Discovering Relations between Named Entities can be done using a linguistic or a numerical or hybrid approach.

The first approach is rule-based, which tried to use syntactic and semantic patterns to capture the corresponding relations by means of manually written linguistic rules. This approach is very interesting for restricted domain and has a good quality of analysis. The major drawback of this approach is the poor adaptability and the poor robustness in handling large-scale or new domain data. This is due to two reasons: rules have to be rewritten for different tasks or when the application is enlarged to different domains and generating rules manually is quite laborious and time consuming (Santos and al., 2010).

The statistical approach (Jun and al., 2009), (Culotta and al., 2006) is based on a learning process from an annotated corpus which can be supervised when the corpus is large, or weakly supervised when the corpus is reduced, or unsupervised if the corpus is not needed. The supervised approach (Miller S. and al., 2000), (Zelenko and al., 2003), (Culotta and al., 2004) and (Kambhatla and al., 2004) costs time and efforts to annotate the corpus, and its performance depends on the size of the corpus. So, to decrease the corpus annotation requirement, some researchers turned to weakly supervised learning approaches, which rely on a small set of initial seed (bootstrap) instead of a large annotated corpus (Agichtein and al., 2000), (Stevenson, 2004). However, the major problem lies in selecting the initial corpus and deciding its “optimal” size. The unsupervised approach works effectively on high-frequent entity pairs. However, this approach has a limited quality in analysis (Hasegawa and al., 2004).

3. Arabic Named Entities Relations discovery challenges

Arabic Named Entities Relations discovery is not an easy task. In fact, besides the problems related to the Arabic Named Entities recognition (Ben Hamadou and al., 2010), (Shaalan and al. 2009), the extraction of relations between Named Entities poses some specific challenges:

- Discontinuity of the multiple relations concerning the same NE (person).

  *Prof. Dr. Imed Abou-Roub the secretary general of organization of computational faculties – Dean of Science Faculty*

  The Person Name: *Prof. Dr. Imed Abou-Roub* is concerned by the relation *Secretary-General* and the relation *Dean-of* in the same time. Between the Person Name and the second relation (*Dean-of*) there is a gap.

- Implicit Relations: they are relations that are not directly specified in the text. They are mined from the text using contextual elements.

  *Doctor Abd el Hadi Moussa, pupil’s committee of medicine Faculty*

  The relation between the two NEs: *Doctor Abd el Hadi Moussa* and *pupil’s committee of medicine Faculty* is not explicitly indicated by a specific word in the text. In this example the affected relation is *Belong-to.*
Necessity to use the previous context of the relation in order to know the missing element involved in the relation.

Professor Doctor Salah Hachem the secretary general

In this example the ORG named entity is absent, but it can be recovered from the right context.

Interference between implicit relation and discontinuity. In the text we have in the same time implicit relation and discontinuity as shown in this example.

The first implicit relation is Belong-to and there is a discontinuity between the Person Name Prof. Dr El-Kacem Ali El-Kacem and the Relation Assistant-of.

4. The proposed approach

As indicated above, the proposed approach for relation discovery is rule based. It is founded on a balance between grammar and lexical resources. The grammar indicates the composition rules of the lexical components, in order to form the different patterns of functional relations. Patterns are transformed into transducers implemented using NooJ Platform.

The lexical resources correspond to one dictionary for each entity type used by the transducers.

a. Patterns of relations

The approach of extraction functional relations between ORG and ENAMEX Named Entities is based on a notion of linguistic pattern that we transform into rules and transducers.

Patterns are considered as regular expressions integrating different elements representing relations and concerned Named Entities.

These patterns are identified semi-automatically from our journalistic learning corpus using NooJ facilities. Indeed, NooJ allows to identify regular expressions in a corpus such as all First Names with the right and the left contexts, all passengers containing a specific relation with the right and the left context.

In the following, we give a list of the main patterns identified in the learning corpus:

<Pattern 1>:=

{<Title>} {<PersName>} {<P>} <REL> <ORG>

المهندس علي التويجري مدير المجمع الكيميائي

Engineer Ali Al-Touijri Director of Chemical group

<Title> Engineer المهندس

<PersName> Ali Al-Touijri

<REL> Director-of مدير

<ORG> Chemical group المجمع الكيميائي
The figure 1 shows the main graph allowing the recognizing of the functional relation between ENAMEX and ORG Named Entities. Each path of this graph represents a different pattern.
The sub graphs:
- NomPersonne allows to recognize the Person Names
- Titre allows to recognize triggers words preceding the Person Name
- Organisation allows the recognition of the ORG Named Entity.
- Fonction allows the recognition of the relation between Person Name and ORG Named Entities.

The figure 2: details the Organisation graph. It allows the recognition of the ORG Named Entity witch has a complex structure and it includes ENAMEX (Person Name) and LOC (i.e. SN Lieu) Named entities.

Beside those graphs, we use a graph for splitting the agglutinated particles (prepositions, definite articles, conjunctions,..) in the analysed Arabic words as shown in this example:

Analysed word: 

Result: Conjonction = + Préposition = + Definite Article = + Noun =
c. NooJ Dictionaries

As mentioned above, the recognition of the relations is based on the following main modules:

- Recognition of the Person Names, including the possible Title,
- Recognition of the ORG Named Entity
- Recognition of the Relation between the indicated Named Entities.

Each module uses its own dictionaries whose entries contain lemmas with the corresponding possible flexional model to generate all derived forms and the corresponding French lemmas.

The recognition of the Person Names module uses the following dictionaries:

- For the Person Names Recognition Step, we build the following dictionaries:
  - Titles Dictionary
  - First Names Dictionary
  - Last Names Dictionary
- For the ORG Recognition module, we build the following dictionaries:
  - Geographical names Dictionary
  - Type institution Dictionary
  - Adjectives Dictionary
- For the Recognition of relations module, we build the following dictionaries:
  - Demonym adjectives Dictionary
  - Functions Dictionary

Table 1 gives extracts of those dictionaries

<table>
<thead>
<tr>
<th>Extracts of Arabic Dictionaries</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong># Title/Functions names</strong></td>
</tr>
<tr>
<td>ریدم, N+Fonction+FLX=A1+FR=directeur</td>
</tr>
<tr>
<td>كلم, N+Fonction+FLX=كلم+FR=roi</td>
</tr>
<tr>
<td>امین, N+Fonction+FLX=رئیس+FR=&quot;Secrétaire&quot;</td>
</tr>
<tr>
<td>نائیب, N+Fonction+FLX=A1+FR=&quot;vice&quot;</td>
</tr>
<tr>
<td>مهندس, N+Fonction+Titre+FLX=A1+FR=&quot;Ingénieur&quot;</td>
</tr>
<tr>
<td><strong># names of geographical categories</strong></td>
</tr>
<tr>
<td>ترکومج, N+FLX=قار+Cat_Geo+Toponyme+FR=république</td>
</tr>
<tr>
<td>غرب, N+FLX=قار+Cat_Geo+Toponyme+FR=royaume</td>
</tr>
<tr>
<td><strong># Geographical names</strong></td>
</tr>
<tr>
<td>سنوت, N+PR+s=Pays+Toponyme+FR=Tunisie</td>
</tr>
<tr>
<td>مونت, N+PR+s=Ville+Toponyme+FR=Tunis</td>
</tr>
<tr>
<td>مارا, N+PR+s=Ville+Toponyme+FR=Riyadh</td>
</tr>
<tr>
<td><strong># Personalities' names</strong></td>
</tr>
<tr>
<td>ابن, N+PR+Perso+f+s+FR=Elisabeth</td>
</tr>
<tr>
<td>بیتیج فیضیروب, N+PR+Perso+m+s+FR=&quot;Habib Bourguiba&quot;</td>
</tr>
</tbody>
</table>
d. Predicate Representation of the Relations

Recognized explicit relations are represented using First Order Logic in the form of a Predicate (i.e., Relation Name) with two arguments: Person Name as the First Argument and the ORG Named Entity as the second Argument. The Name of the relation is the Lemma extracted from the Function Dictionary.

Example:
أمين عام (أبو الرب، جمعية كليات الحاسبات والعلومات)
Secretary-General (Abu al-Rub, Association of Colleges of Computing and Information)

For the implicit relation we generate systematically the name “Belong_to” ينتمي_إلى.

Example:
يتنتمي إلى (القائم على القاسم، كلية الزراعة)
Belong_to (Al-Kacem Ali Al-Kacem, Faculty of Agriculture)

e. Translation of the relations

The translation of the recognized relations is done in the perspective of a multilingual assimilation of the analysed texts or documents. And the target language is the French but we can add other languages easily.

So we do not pay a great attention to the quality of the translation process, and we translate only lemmas of the words (i.e., nouns and adjectives) and not their derived forms as they occur in the text. For example:

جامعات Universities is translated as University,
جامعة Générale is translated as Général (masculine form).

French corresponding lemmas are added to the entries of the different Dictionaries that are used in the recognition process.
This translation, although it is not very good, allows us to question an Arabic text using French language and get answers in Arabic or even in French.

Example:
For the question: *Qui est le Président de l’Université Al-Albeit?* (see figure 4), we get the answer in Arabic: ﺍﻟﺸﻮﻗﻔﺔ ﺍﻟﺪآﺘﻮر ﺍﻟﺄﺳﺘﺎذ and in French: Professeur Docteur Al-Chaouafka.

## 5. Experimentation and Evaluation

For the system evaluation, we created a new corpus composed of the learning corpus and Wikipedia pages.

Figure 3 gives a sample execution of the proposed recognition system given by NooJ platform on the indicated corpus. It represents the recognised Relations between Person Names and ORG Named Entities with its corresponding Predicate representation.
The evaluation metrics we used for the recognition process, are Recall, Precision and F-measure ($2*P*R/(P+R)$). Let’s remember that the recall measures the quantity of relevant responses of the system compared to the ideal number of responses; Precision is the number of relevant responses of the system among all the responses he gave and the F-measure is a combination of Precision and Recall for penalizing the very large inequalities between these two measures. The values obtained in the evaluation of our work are:

<table>
<thead>
<tr>
<th>Sports venues</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>63%</td>
<td>78%</td>
<td>70%</td>
</tr>
</tbody>
</table>

The values obtained can be explained by the fact that our dictionaries are not very large, especially for First Names and Last Names, and also by the fact that the names of organizations are very long and complex.

6. Conclusion

In this paper, we have presented a rule based approach for the extraction of functional relations between ENAMEX and ORG Arabic Named Entities. We particularly highlighted the NE relation Discovery problems. Some of them are specific to the Arabic language. These problems have been largely resolved, but some merit special consideration. The approach was implemented using NooJ platform. We have also given experimentation on a journalistic test corpus. The experimentation and the evaluation results are satisfactory.

As perspectives, we are working on expanding the dictionaries, especially for First Names and Family Names with corresponding translations. Also, we are interested in new types of relations for the economic domain in order to recognize events. Finally we project to integrate our system in a question answering system as a component for factoid questions.
References


César P., Juan P., Isabel S., Paloma M. 2009. The UC3M team at the Knowledge Base Population task.


