A New Representation Model for the Automatic Recognition and Translation of Arabic Named Entities with NooJ

Héla Fehri
Laboratory MIRACL, University of Sfax
hela.fehri@fss.rnu.tn

Kais Haddar
Laboratory MIRACL, University of Sfax
kais.haddar@fss.rnu.tn

Abdelmajid Ben hamadou
Laboratory MIRACL, University of Sfax
abdelmajid.benhamadou@isimsf.rnu.tn

Abstract

Recognition and translation of named entities (NEs) are two current research topics with regard to the proliferation of electronic documents exchanged through the Internet. The need to assimilate these documents through NLP tools has become necessary and interesting. Moreover, the formal or semi-formal modeling of these NEs may intervene in both processes of recognition and translation. Indeed, the modeling makes more reliable the constitution of linguistic resources, limits the impact of linguistic specificities and facilitates transformations from one representation to another. In this context, we propose an approach of recognition and translation based on a representation model of Arabic NEs and a set of transducers resolving morphological and syntactical phenomena.

1 Introduction

The formal or semi-formal modeling of NEs can be involved in recognition and translation process. It enables to constitute more reliable linguistic resources. Indeed, such a modeling can represent all the constituents of a NE in a standard manner and limit the impact of linguistic specificities. In fact, a formal representation of Arabic NEs can help, firstly, in the identification of dictionaries and grammars required for a given application and, secondly, in the use of advanced linguistic methods of translation (i.e., transfer or pivot method). This abstraction level favors the reuse of certain linguistic resources. The elaboration of a formal and generic representation of an NE is not an easy task because, on the one hand, we have to find a representation that takes into consideration the concept of recursion and length of NE. In fact, a NE can be formed by other NEs. So, its length is not known in advance. On the other hand, the representation to be proposed should also contain a sufficient number of features that can represent any NE independently of the domain and grammatical category.

It is in this context that the present work is situated. In fact, the main objective is to propose an approach of recognition and translation of Arabic NEs based on a representation model, a set of bilingual dictionaries and a set of transducers resolving morphological and syntactical phenomena related to the Arabic NEs and implemented with the linguistic platform NooJ (Silberztein, 2005).

In this paper, we present, firstly, a brief overview of the state-of the art. Next, we describe the hierarchy type of Arabic NEs and the identified problems in recognition and translation processes. Then, we detail our proposed representation model. After that, we give a general idea of our resources construction and their implementation in the linguistic platform NooJ. Finally, the paper concludes with some perspectives.

2 Related work

Research on NEs revolves around two complementary axes: the first involves the typing of NEs while the second concerns the identification and translation of NEs. As for the identification, the tagging and the translation of NEs, they have been implemented for multiple languages based on different approaches: linguistic (Coates-Stephens, 1993), statistic (Borthwick et al., 1998) and hybrid (Mikheev et

Proceedings of Recent Advances in Natural Language Processing, pages 545–550,
Hissar, Bulgaria, 12-14 September 2011.

al., 1998) approaches. In what follows, we focus on the linguistic approach.

Regarding the recognition of NEs, based on the linguistic approach, we cite the work presented in (Friburger, 2002). This work allows the extraction of proper names in French. The proposed method is based on multiple syntactic transformations and some priorities that are implemented with transducers. We can cite also the work described in (Mesfar, 2007). The elaborated method is applied on a biomedical domain. Other Arabic works are dealing with the recognition of elliptical expressions (Hasni et al., 2009) and most important categories in Arabic script (shaalan et al., 2009).

Other works have been dedicated to the translation of different structure (e.g., NE) from one language to another. We can cite the work presented in (Barreiro, 2008) dealing with the translation of simple sentences from English to Portuguese. Additionally, the work of (Wu, 2008) provides a noun translation of French into Chinese.

The literature review shows that the already proposed translation approaches are not well specified (e.g., lack of abstraction and genre). Each one addresses a particular phenomenon without taking into account other phenomena. We should also mention that there are few works that proposed a modeling of NEs for explicitly representing the effects of meaning within the NE and explaining phenomena like synecdoche and the metonymy (Poibeau, 2005). However, these works don't treat the concept of embedded NEs which is very important and can help to implement the recognition and the translation process of NEs. Furthermore, all translations using NooJ platform adopt a semi-direct approach of translation, in which the recognition task is combined with that of translation. Thus, the reuse of such work has become limited, which does not promote multilingualism.

3 Hierarchy of Arabic NEs and identified problems

3.1 Hierarchy of Arabic NEs

The hierarchy of Arabic NEs that we propose is inspired from MUC conferences (Grishman, 1995). This hierarchy does not differ from other typologies of other languages. In fact, categories that make up the proposed hierarchy are common to almost all domains. Indeed, our contribution focuses on the refinement done in different categories in various levels. In order to do this refinement, we must choose a domain. In our work, we chose the sport domain. Therefore, all our examples are related to this domain and especially to the category of place names belonging to the category of proper names but we should mention that our work is also applied to place names regardless of the domain. Figure 1 illustrates the suggested hierarchy.

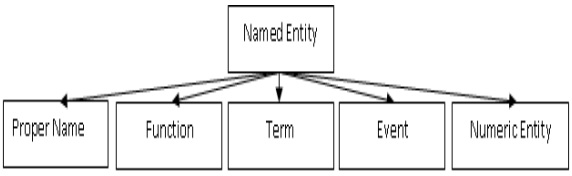

Figure 1: NE Hierarchy of the sport domain

Let's note that the proposed typological model comes as a result of the study of various forms of denomination of sports names (e.g., stadium, swimming pools, teams names) on corpora and lists of official names of the sport domain available on the Internet for Arabic countries

This proposed hierarchy allows the typing of the main constituents of NEs from a set of predefined categories. In fact a NE can be composed of others NEs. It is obvious that if a NE contains several NEs, it can cause different problems such as polysemy as mentioned in (Poibeau, 2005). However, it proves also the concept of embedded NEs. So, a modeling by a set of features may be an appropriate solution to explicitly represent this notion. In fact, it can help the process of recognition and translation of NEs. Later, we detail our proposed model allowing the implementation of the process of recognition and translation of NEs.

3.2 Identified problems in recognition and translation of Arabic NEs

Problems in Arabic NE' recognition: Arabic NE' recognition needs to solve some problems. For example, we can cite:

Proper name problem. In Arabic, there is a big challenge for finding those proper names in the text because they do neither start with capital letter as in many other languages, nor do they have special sign to identify and distinguish between them and other words in the text.

Syntactic problem. Arabic NE grammar is rich and variant. Indeed, the length of NE (number of constituents) is not known in advance.

Problems in Arabic NE' translation: In our work, NE' translation is done from Arabic to

French. The study of this process shows that there exist many problems. For example, we cite:

- Gender feature correspondence. Gender feature value is not always the same for Arabic word and its equivalent in French. For example, the word مســـبح *swimming pool* is masculine but its translation to French *piscine* is feminine.
- Ambiguity between capital name and city name. For example, the toponym تـــونس *Tunisia* can be translated to *Tunisie* or *Tunis* in French.
- Arabic adjective position is different in French. For example, ملعب عبد العزيز**الأولمبي** *malaab Abdelaziz* ***el oulimpi*** *Abdelaziz* ***Olympic*** *stadium* is translated to *Stade* ***olympique*** *Abdelaziz*.

4 Proposed model for representing Arabic NEs

The model that we propose is used to formalize and to identify Arabic NEs. This model is inspired by formalisms based on structural features like Head-driven Phrase Structure Grammar (Pollard et al., 1994). Its features are inspired from the concepts "Head and Expansion" introduced by (Bourigault, 2002).

The essential characteristics of the feature structure of the proposed model are: an element of the structure can be atomic or complex and an internal structure of an element is defined by its attributes and values.

4.1 Structure and features of the proposed model

Each NE has a type and is composed of two parts: one is essential and the other is extensional. The essential part is also a NE and has itself essential and extensional parts. This proves the recursion for an NE. The type of a NE ″Type_EN″ is usually indicated by a trigger word. The essential part is represented by the feature ″Tête_EN″ (head of NE) and the trigger word is represented by the feature ″Mot_declencheur″. The extensional part represents the final form that composes the NE. It does not admit a type because it is preceded by a lexical item ″Element_EN″ (e.g., preposition, special character). Then, it can not be considered as a NE but it can contain a NE. Its existence or non-existence doesn't affect the well-formation of the NE. This part is represented by the feature ″Fin_EN″.

The value of the feature ″Tête_EN″ can be atomic or structured. If it is structured, then it is composed by the features ″Mot_déclencheur″, ″Tête_EN″, ″Fin_EN″ and ″Type_EN″. The ″Mot_déclencheur″ value is simple or composed. Indeed, the trigger word can be formed by a word or a sequence of words. It can also be empty. The ″Fin_EN″ value can be atomic or structured. If it is structured, then it is composed by the features ″Element_EN″, ″Tête_EN″ and ″Fin_EN″. It can also be empty. The feature ″Type_EN″ value is always simple or composed but not empty. In fact, it represents one of the categories identified in the NE hierarchy. The ″Element_EN″ value is always simple. The structure can be equipped with a set of principles allowing the construction and evaluation of NE-representation.

4.2 Principles of the proposed model

Saturation principle: A structure is called saturated if it can be considered as a well-formed NE. That means, it consists of a NE head (″Tête_EN″) whose value is not empty. Figure 2 describes an example of a formal representation that satisfies a saturation principle.

El *Riadh* الرياض

Tête_EN	الرياض
Fin_EN	ϕ
Type_EN	Ville

Figure 2: Representation of the word الرياض *el Riadh*

Non-saturation principle: A structure is called unsaturated if it isn't a NE and can be completed to become a NE. That means, it is formed only by a NE end (*″Fin_EN″*) or if the value of the feature ″Tête_EN″ is empty. For example, in the word بالرياض *bi Riadh*, the value of the feature ″Tête_EN″ is empty because this word doesn't have a type. Thus, this word cannot be considered as a NE. It doesn't satisfy the saturation principle. However, it should be noted that this word can contain a NE. The two mentioned principles allow us to avoid ambiguity between a NE-word (or set of words) and a non NE-word.

4.3 Literal translation representation

Word-to-word translation consists to translate each feature value composing a NE structure

representation. This translation is done with bilingual dictionaries without any risk of information loss. For instance, in the NE ملعب الملك عبد العزيزالدولي Malaab el malik Abd el Aziz el doali bil Riadh, the word ملعب *malaab stadium* is translated to *stade*, the word الملك *el malik king* to *roi*, the adjective الدولي *el doali international* to *international and* the preposition ب *bi in* to *de*.

Let's note that the representation of a word-to-word translation is not sufficient to generate a well formed NE in the target language. Therefore, readjustment rules are necessary and should be associated in translation process.

5 NooJ implementation of the set of transducers

The NooJ implementation of our system requires two phases process: recognition of Arabic NEs phase and translation phase in which the transliteration process is integrated.

5.1 Phase of recognition

The proposed representation model helps us to identify the necessary resources for the recognition and translation of NEs. In fact, each structured feature ″Tête_EN″ containing not empty features, other than the feature ″Type_EN″, is transformed into a grammar. Whereas, each elementary NE (value of ″Tête_EN″ feature is atomic) will be transformed into a dictionary.

From the NE representation in the considered model, we have created the following transducer:

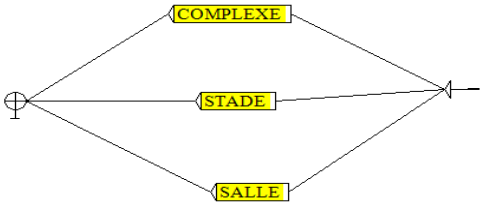

Figure 3. Main transducer of NE' recognition

The transducer of Figure 3 allows recognition of NEs belonging in the sport place name category. Each path of each sub-graph represents a rule extracted in the study corpus.

In the recognition phase, we have solved the problems related to the Arabic language (eg, agglutination) establishing morphological grammars built into the platform NooJ. This phase contains 19 graphs respecting the production rules identified in the study corpus.

5.2 Phase of translation

Word-to-word translation: To implement the process of word-to-word translation in the platform NooJ, we built a syntactic grammar allowing the translation of each word composing a NE with the exception of words not found in dictionaries, or can not be translated (number, special character, etc..). This grammar takes as input the NE list extracted by the transducer of Figure 3 allowing the recognition and it is described by the transducer of Figure 4.

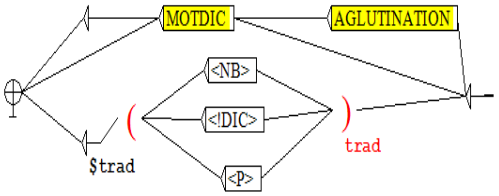

Figure 4: Transducer of word-to-word translation

The sub-graph MOTDIC treats the words existing in dictionaries which require a specific treatment.

Translation with readjustments: Several readjustment rules must be applied to improve the word-to-word translation step. These rules have essentially a relationship with the order of the words composing a NE and with the agglutination. For instance, on the one hand, if a NE in the source language contains an adjective then we have to know whether this adjective belongs to the trigger word or to the noun that comes just before. On the other hand, if a NE in the source language contains a noun then some rules are applied to solve the problem of contracted forms in Arabic.

Transliteration process: The transliteration is done after having executed all the transducers allowing the NE' recognition and translation. In fact, it consists in transliterating all the non-translated words whish are written in the source language (Arabic characters) using the appropriate resources. In this process, we consider the rules respecting the chosen transliteration system El Qalam and also the transformation rules. These rules are implemented with NooJ morphological transducers. The transliteration is preceded by a voweling phase to avoid some problems. However, the connection between a vowel transducer and transliteration transducer can not be done in NooJ; that is why, we resort to use noojapply. noojapply is a command-line program which can be called either directly from a "shell" script, or from more sophisticated programs

written in PERL, C++, JAVA, etc. In our work, we use C#.

6 Experimentation and evaluation

The experimentation of our resources is done with the linguistic platform NooJ. As mentioned above, this platform uses (syntactic and morphological) grammars already built and dictionaries. To the resources of NooJ, we added these dictionaries: Team Names (5785 entries), Sport Names (337 entries), Capital and country Names (610 entries), Personality Names (300 entries), Trigger words (20 entries) and Functions Names (100 entries).

In addition to the mentioned dictionaries, we use other dictionaries existing in NooJ like dictionary of adjectives, nouns and First Names. To these dictionaries, we add some entries related to the sport domain. We also add French translations of all entries in all mentioned dictionaries. Let's note that the First Name dictionary remains monolingual because its entries can be transliterated. To experiment and evaluate our work, we have applied our resources to two types of corpus: sport and education.

6.1 Experimentation of recognition phase

To evaluate a recognition phase, we have applied our resources to a corpus formed by 4000 texts (94,5 Mo) of sport domain (different of the study corpus). It contains 180000 NEs belonging to different categories of sport domain (e.g., player name, name of sport, sports term). In these NEs, there are 40000 NEs belonging to the category place name. These NEs are manually identified using NooJ queries.

Let's note that NE is detected if it satisfies one of the paths described by the transducer of Figure 3. Indeed, a transducer is characterized by an initial node and one or many end nodes. If multiple paths are verified, we maintain the longest one.

The obtained results give 98% of precision, 90% of recall and 94% of F-measure. This measures show that there are problems that are not yet resolved. Some problems are related to the lack of standards for writing proper names (e.g., el hamza) and the absence of some words in the dictionaries. This causes a silence. Other problems are related to specific concepts in the Arabic language as metaphor.

We have also applied our resources to the education domain. We have collected a corpus composed of 300 texts (14.5 Mo) containing university 3000 institution names. The performance measure of the obtained results gives 98% of precision, 70% of recall and 82% of F-measure. We deduce that silence is increased. This is caused by the incompleteness of specific dictionaries to this domain and lack of some paths in the developed transducers. So our resources are applicable regardless of the domain, provided that we use the same features adopted in dictionaries we have built. It is evident that for reasons specific to the field, we should sometimes add other paths and other sub-graphs, but we do not have to redo everything.

6.2 Experimentation of translation phase

The translation phase is applied to the extracted Arabic NEs during the recognition phase. Note that erroneous results are inherited. Therefore, heuristics filtering are necessary before the translation process. The obtained results of the translation phase are illustrated in Figure 5.

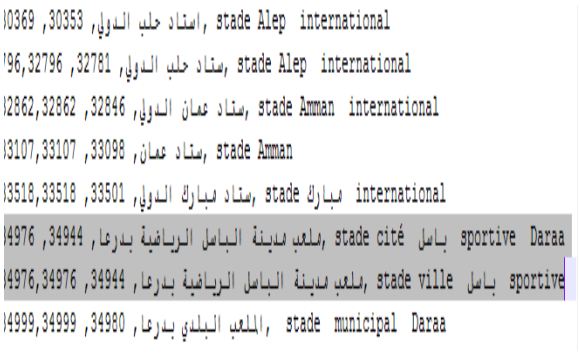

Figure 5: Extract of results of word-to-word translation

As shown in Figure 5, the proper problems of this phase involve multiple translations that can be assigned to a word. For example, the selected lines in Figure 5 represent the NE' translation ملعب مدينة الباسل الرياضية بدرعا *malaab madinat el bacel el riadhiya bi deraa stadium of city Bacel sportive in Deraa*. In this NE, the word مدينة *madina* can be translated to the word "cité" *city* or "ville" *country*. NooJ displays all possibilities. In this case, the adjective can resolve this ambiguity. In fact, the adjective الرياضية *el riadhiya sportive* is generally related to the city and not to the country. Let's note that the word "باسل" *Bacel* remains in the source language because it is a first name, so it will be transliterated later.

Our method provides 97% of well translated NEs while ensuring the specificities of the target language. The obtained result is promising and

shows that there are some problems not resolved. These problems are related to the multiple translations assigned to a toponym (e.g., تونس *tounis* can be translated in tunis or tunisie).

The proposed representation model facilitates the implementation and the building of the linguistic resources with the platform NooJ. It facilitates also the transformation from the semi-direct translation to transfer translation. Indeed, we have separated the NE' recognition of their translation. In addition, it helps the promotion to the reuse of the needed grammars. In fact, it is sufficient to change the inputs (i.e., dictionaries, morphological grammars) of the syntactic grammars for the desired results. Thus, for example, if we want to translate Arabic NE to another language other than French, the recognition module can be reused with some modifications if necessary (related to the specificities of the domain).

7 Conclusion and perspectives

In this paper, we have proposed an approach for recognition and translation of Arabic NEs (eventually NE from other language) based on a representation model, a set of bilingual dictionaries and a set of transducers resolving morphological and syntactical phenomena related to the Arabic NEs. Moreover, we have given an idea of the hierarchy types of Arabic NEs and of the identified problems in the recognition and translation processes. Besides, we have described the representation model structure, its features and principles that should be satisfied. We have also given an experimentation and evaluation on the sports and education domains proving that our resources can be reused independently of the domain. The experimentation and the evaluation are done in the linguistic platform NooJ. The obtained results are satisfactory.

As perspectives, we seek to improve the model by introducing other features related to the semantics. Furthermore, we are currently identifying heuristics filtering enabling finer translation.

References

Barreiro, A. 2008. *Port4NooJ: an open source, ontology-driven Portuguese linguistic system with applications in machine translation*. NooJ'08, Budapest.

Borthwick, A., Sterling, J., Agichtein, E. Grishman, R. 1998. *NYU: Description of the MENE Named Entity System as used in MUC-7*. In Proc. of the Seventh Message Understanding Conference (MUC-7).

Bourigault, D. 2002. *UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus*. TALN.

Coates-Stephens, S. 1993. *The Analysis and Acquisition of Proper Names for the Understanding of Free Text*. In Computers and the Humanities, Kluwer Academic Publishers, Vol. 26(5-6), Hingham, MA, p. 441-456.

Friburger, N. 2002. *Reconnaissance automatique des noms propres*. PhD thesis, university of François Rabelais.

Grishman, R. 1995. *Where's the Syntax? The NYU MUC-6 System*. In Acts of MUC-6, Morgan Kaufmann Publishers, San Francisco.

Hasni, E., Haddar, K., Abdelwahed, A. 2009. *Reconnaissance des expressions elliptiques arabes avec NOOJ*. In proceedings of the 3rd International Conference on Arabic Language Processing (CITALA'09) sponsored by IEEE Morocco Section, 4-5 May 2009, Rabat, Morocco, pp 83-88.

Mesfar, S. 2007. *Named Entity Recognition for Arabic Using Syntactic grammars*. NLDB 2007 Paris, 28-38.

Mikheev, A., Grover, C. et Moens, M. 1998. *Description of the LTG system used for MUC -7*. In Proc. of 7th Message Understanding Conference (MUC-7), http://www.itl.nist.gov/iad/ 894.02/ related_projects/muc/.

Poibeau, T. 2005. *Sur le statut référentiel des entités nommées*. Laboratory of data processing of Paris North – CNRS and University Paris 13.

Pollard, C., Sag., I.A. 1994. *Head-Driven Phrase Structure Grammar*. Published by the press in the University of Chicago, Edition Golgoldmittu, Chicago, LSLI.

Shaalan, K., Raza, H. 2009. NERA: *Named Entity Recognition for Arabic*. Published in Journal of the American Society for Information Science and Technology, Volume 60 Issue 8.

Silberztein, M. 2004. *NooJ : an Object-Oriented Approach*. In INTEX pour la Linguistique et le Traitement Automatique des Langues. C. Muller, J. Royauté M. Silberztein Eds, book of the MSH Ledoux. Presses University of Franche-Comte, pp. 359-369.

Wu, M. 2008. *La traduction automatique français-chinois pour les groupes nominaux avec NooJ*. Budapest.