Arabic Named Entity Recognition Process using Transducer Cascade and Arabic Wikipedia

Fatma Ben Mesmia
University of Sfax,
Laboratory MIRACL, Multimedia, InfoRmation Systems and Advanced Computing Laboratory,
fatmabm@ymail.com

Nathalie Friburger
University François Rabelais of Tours,
LI, Computer laboratory,
nathalie.friburger@univ-tours.fr

Kais Haddar
University of Sfax,
Laboratory MIRACL, Multimedia, InfoRmation Systems and Advanced Computing Laboratory,
Kais.Haddar@fss.rnu.tn

Denis Maurel
University François Rabelais of Tours,
LI, Computer laboratory,
denis.maurel@univ-tours.fr

Abstract

Transducers namely transducer cascades are used in several NLP-applications such as Arabic named entity recognition (ANER). To experiment and evaluate an ANER process, a weight coverage corpus is necessary. In this paper, we propose an ANER method based on transducer cascade. The proposed transducer cascade is generated with the CasSys tool integrated in Unitex linguistic platform. The experimentation of our method is done on a Wikipedia corpus. The Wikipedia text format is obtained with Kiwix tool. The experiment results are satisfactory based on calculated measures.

Keywords: Cascade of transducers, Wikipedia, Arabic named entities, Unitex, CasSys

1. Introduction

Transducers can play an important role in the Information Extraction (IE) namely in the Named Entity Recognition (NER). At the same time, transducers can extract and classify the Arabic Named Entity (ANE). Generally, the use of transducers is realized in well defined succession that is called cascade (Friburger and Maurel, 2004).

In fact, the identification of necessary transducers is not an easy task because several linguistic phenomenacan interact (Shaalan, 2014; Ben Mesmia and al, 2015).

The free resource Wikipedia is an important information source. Indeed, several text processing applications based on transducer cascade can benefit from Wikipedia articles. Therefore, names of people, which are part of proper nouns, appear frequently in the Arabic Wikipedia. More efforts by NLP-researchers are concentrated on this type. Person names are considered as the most challenging task for Arabic.

In this context, our objective is to propose, using the rule-based approach, a transducer cascade for the recognition of personality's names. In this approach, we benefit from the robustness of transducers and exploit the free resource, Wikipedia. The recognition requires the identification of dictionaries, a list of trigger words and extraction rules allowing the development of a set of transducers acting on the corpus with a certain logic.

The present paper is composed of six sections. The second section presents previous work describing the developed systems for the recognizing of the personality names. The third section is dedicated to describing the categorization of person's names. The fourth section devoted to detail the proposed method that is implemented by using CasSys system. The experiment is presented and evaluated in section five. Finally, we give a conclusion and some perspectives.

2. Previous Work

There are several work treating the ANER based on several approaches among which we cite the work of (Shaalan and Raza, 2007). In this work, the authors proposed an ANER system based on the rule-based approach. This system called

Proceedings of Recent Advances in Natural Language Processing, pages 48–54,
Hissar, Bulgaria, Sep 7–9 2015.

PERA is composed of three components: gazetteers, local grammars and a filtration mechanism. PERA is applied to the ACE and ATB datasets.

In (Mesfar, 2007), the author developed a system identifying ANE of many types such as person names. This system consists of a tokenizer, a morphological analyzer and a NE finder. The system is evaluated by using the of news corpus extracted from le journal “Le monde diplomatique”.

In (Elsebai et al, 2009), the authors proposed a rule-based system that integrates pattern matching with morphological analysis to extract Arabic person names. This system is evaluated by using news articles extracted from Aljazeera website.

In (Fehri and al., 2011), authors developed a rule-based system to recognize ANE for sport’s domain such as place names and player’s names. This system is composed of a set of dictionaries, syntactic patterns and transducers implemented with the linguistic platform NooJ.

In (Aboaoga and Aziz, 2013), the authors introduced a rule-based system that extracts Arabic person names. The system is composed of three steps: the preprocessing (tokenization, data cleaning and sentence splitting), the automatic ANE tagging and the application of rules to the Arabic texts in order to extract ANEs that do not exist in the built dictionaries. The domains covered by this system are sports, politics and economics.

In (Elsebai, 2008), the author developed a system adopting statistical approach for ANER. This system allows the recognition of Arabic proper names using heuristics. Heuristics based on a set of key-words rather than complex grammars and statistical techniques. The system is evaluated by using news articles extracted from the Aljazeera television website.

In (Shaalan and Oudah, 2014), the authors proposed a system based on hybrid approach. This system, which is capable of recognizing 11 types of Arabic named entities such as person names, is applied to ANERcorp standard dataset. According the study made by Shaalan (2014), systems which are developed for the ANER, are essentially based on restraint domains.

Namely in the NER, the use of transducer cascade is very frequent. A cascade is defined as a succession of transducers applied to text in a specific order to convert or extract patterns. Each transducer of the cascade uses the results of the previous transducer (Maurel and al., 2009).

Several systems based on cascades were developed in NLP that touch essentially the following domains: parsing, information extraction and translation. Among the systems constracted for the IE task, we cite the following work.

In EU project FACILE, (Ciravegna and Lavelli, 1999) implemented a module based on three transducers cascades. These cascades contain transducers representing respectively empirical, regular and default rules.

CasEN, the system developed by (Maurel and al., 2011) uses lexical resources and transducers acting together on texts by insertions, deletions or substitutions.

For Arabic, (Ben Mesmia and al, 2015) developed a transducer cascade allowing the recognition of ANE more precisely the dates. This cascade is generated by the CasSys that is module available under the Unitex platform.

3. Typology of Arabic Person’s Names

The Arabic names may have variations related to origin of country, religion, culture, level of formality and even personal preference. In this section, we present firstly our study corpus. Secondly, we give the categorization of person names. We explain also phenomena that are related to their recognition.

3.1 Corpus of Study

The corpus of study was collected from Arabic Wikipedia through Arabic kiwix[1] tool. It regroups a number of texts from 19 Arabic countries and contains text files for a cumulative 79 659 tokens. This corpus allows us to identify the forms that will be transformed into extraction rules and transformed later in transducers.

3.2 Categorization of Person’s Names

In general, an Arabic name can contain five parts, which follow no particular order: the ism, kunya, nasab, laqab, and nisba (Shaalan, 2014).

The ism is the first name. These are the names given to children at their birth. Male isms are such names as “*`bd allah*[2] / Abdullah”, “*`aadl* / Adel”, “*Hsyn* / Hussein”. Men’s isms are sometimes preceded by one of the attributes of Allah such as “*’aaHmd* / Ahmed”, “*mHmwd* / Mahmoud” but this practice is declining, especially in areas influenced by Western practices, such as Lebanon,

[1] http://www.kiwix.org/wiki/Main_Page/fr

[2] Transliteration system : Al-Qalam : http://www.cs.bu.edu/ftp/amass/Arabic/qalam.txt

Morocco, and other North African countries. Female isms include "*`aa'sht* / Ayisha" and "*smyrt* / Samira". The "*t*" sound is a feminine ending.

The kunya is an honorific name. It is not part of a person's formal name. The kunya is used as an informal form of address and respect, much as we use "aunt" and "uncle". It indicates that the man or woman is the father or mother of a particular person, the birth of a child being considered praiseworthy and deserving of recognition. For example, "*'aam klthwm* / Oum Kultthum" means "mother of Kulthum", and "*'aabw klthwm* / Abu kulthum" means "father of Kulthum".

The nasab is the patronymic and starts with "*bn* /bin" or "*aabn*/ ibn", which means "son of", or "*bnt* / bint", which means "daughter of". It acknowledges the father of the child. The nasab often follows the ism, so that you have, for example, "*fHd bn `bd aal`zyz* / Fahad ibn Abdul Aziz", which means "Fahad, son of Abdul Aziz". A daughter would be "*mrym bnt `bd alla`zyz* / Maryam bint Abdul Aziz". If someone wishes to acknowledge the grandfather and great-grandfather as well, these names may be added. So one could have "khaald bn fySl bn `bd aal`zyz / Khalid ibn Faisal ibn Abdul Aziz". The use of bin and ibn varies greatly.

The laqab is defined as an epithet, usually a religious or descriptive one. For example, "*aalrshyd* / Al-Rashid" means "the rightly guided" and "*aalfZl* / Al-fadl" means "the prominent".

The nisba is similar to what people in the West call the surname. Again, the use of this term varies in Egypt and Lebanon, such as nisba is not used at all. Instead, laqab incorporates its meaning. The nisba is often used as the last name, although its use has decreased in some areas.

3.3 Difficulties of Extraction

In Arabic, several causes make the NER difficult. In the following, we mention some of them.

Absence of capitalization. In Arabic, capitalization does not exist.

Nature of proper nouns. Proper noun can belong to the adjective category or to the temporal expression. For example, "*jmyla*" can be a girl name or an adjective and "*jm`t*" can be a day (Friday) or a boy name.

Agglutination. An Arabic word can be a whole sentence. In fact, several particles can be attached to a root such as prepositions. For example, "*لكتابته*" means in English for writing it

Typographic variants. The drop of Hamza sign. For example, the proper name "*Aahmd*" can be written with or without the Hamza sign.

Nested ANE. To find the limit of ANE is not easy. A personality name can be a part of an event NE. For example, "*frHaat Hshaad*" is a personality name which it a part of the event "*laastshhaad aalmnaaDl frHaat Hshaad*". This event is also a part in "*Dhkrae stt w styn laastshhaad aalmnaaDl frHaat Hshaad*".

3.4 Relationship between Personality's Names with other ANE

The relationship between ANE can be binary (involving two entities) or more complex to be an imbrication of ANE. The ANE describing events and place names can have a compositional relationship with ANE of the type names of personality. In (1) and (2), "*aalTyb aalmhyry* / Al-Taieb al-Mhiri" and "*mHmd aalkhaams* / Mohamed Al-Khames" are two names of personality integrated in two ANE of the type name places preceded respectively by "*ml`b*" and "*shaar`*".

(1) ملعب الطيب المهيري بصفاقس
ml`b aalTyb aalmhyry b Sfaaqs

(2) شارع محمد الخامس
Shaar` mHmd aalkhaams

The organization name can contain famous name of personality such as in (3), "*aal`nwd*" is a first name of a princess.

(3) مؤسسة العنود الخيرية
M'wsst aal`nwd aalkhyryt

Arabic names of personalities can appear also in Events such as in (4), "*mraasm tnSyb*" are the two trigger words recognizing this type and the rest of the entity is the name of personality.

(4) مراسم تنصيب الملك عبد الله بن عبد العزيز آل سعود
mraasm tnSyb aalmlk `bd alllh bn `bd aal`zyz aal s`wd

4. Proposed Method Recognizing Personality Names

The proposed method is based on three steps: the construction of necessary dictionaries, the identification of extraction rules to recognize ANE and the establishment of the corresponding transducers. In the following, we detail these steps.

4.1 Construction of Dictionaries

For our method, we construct two dictionaries with several features. One contains the first names. The second dictionary contains the last names. Therefore, these dictionaries treat different variations of Arabic person's names

4.2 Identification of Extraction Rules

According to our study, we identify 14 extractions rules. Each rule describes an alternative form of personality name. These extraction rules are detected through trigger words. We identified 180 trigger words that are classified in eight classes. They are distributed as in Table 1.

Class names	Number of trigger words
Artistic function	47
Civilities	21
Military function	7
Nobiliare function	22
Political function	27
Profession	14
Religious	17
Sportive function	25

Table 1. Distribution of the trigger words by class

In the following, we give trigger word grammar for the identified classes.

Trigger Word → *Artistic function* | *Civilities* | *Military function* | *Nobiliare function* | *Political function* | *Profession* | *Religious* | *Sportive function*

Artistic function → *aalma'lf* | *aalmw'lft* | *aalmbd`*| *aalmbd`t* | *aalktb* | *aalktbt* | ...

Civility → *aalsydt* | *aalsyd* | *aalaa'nst* | ...

Military function → *aaljysh* | *aalraaae'd* | *aalz`ym* | *aalmqdm* | *aal`qyd* | ...

Nobiliare function → *aalaa'myr* | *aalaa'myrt* | *aalslTan* | *aalslTant* | ...

Political function → *rae'ys aaljmhwryt* | *aalwzyr* | *wzyr aaldwlt* | *rae'ys aaldwlt* | ...

Profession → *aalm`lm* | *aalmdyr* | *aalaa'staadh* | *aalm`lmt* | ...

Religious → *aalaa'maam* | *aalmw'dhn* | ...

Sportive function → *aallaa`b* | *aallaa`bt* | ...

Concerning the established extraction rules, we propose a classification based on three classes. The first class contains recognition paths depending on trigger words; the second class describes the recognition of independent paths. The third class concerns rules that appear in exceptional cases encountered during the study. Table 2 shows an example of extraction rules.

Extraction rules
<Trigger Word> < first name>$^+$ <last name>
(< first name> ben <first name>)$^+$ <last name>
(< first name> ben <first name>)$^+$ <Nisba>
<Trigger Word> <last name>
<Trigger Word> < Country name> <first name>
< first name> <Kunya> (ben <first name>)$^+$

Table 2. A set of extraction rules extracted from the study corpus

4.3 Establishment of Transducers

The extraction rules are translated in transducers. Each transducer regroups similar forms. Most of them are based on trigger words, which facilitate the recognition process. Even the trigger words are grouped into sub-transducers because they will be called by other graphs.

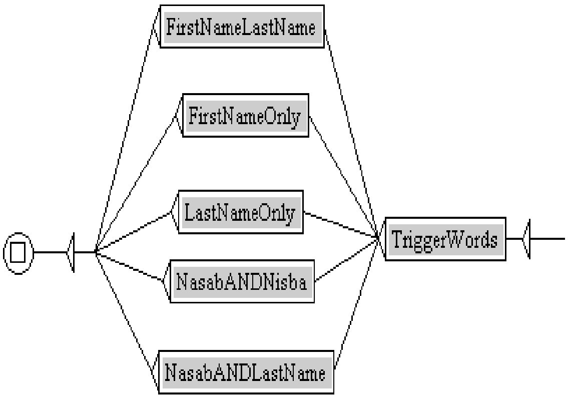


Figure 1. A transducer that call sub-transducers using the trigger words

Figure 1 shows the implementation of many extraction rules, which use triggers words, allowing the personality's names recognition. The subgraph entitled "NasabANDNisba" describes the path allowing the recognition of an Nsab followed by a Nisba. This sub-graph is described in the following figure.

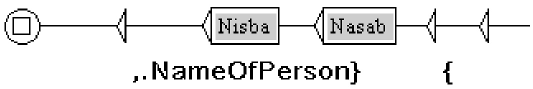


Figure 2. The path allowing the recognition of a Nasab followed by a Nisba

In Figure 2, there is two sub-graphs, which are respectively "Nasab" and "Nisba". These subsgraphs are surrounded by two-box containing the annotation that will appear in the corpus on which the transducer will be passed. The graphs "Nasab" and "Nisba" are presented in Figure 4 and 5.

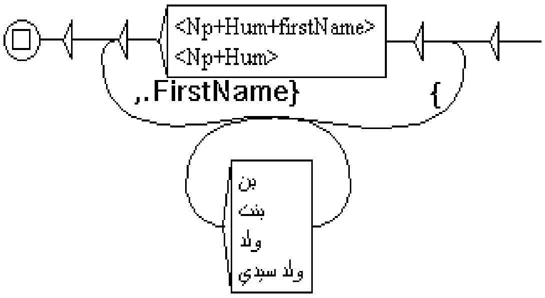


Figure 4. Transducer recognizing the Nasab

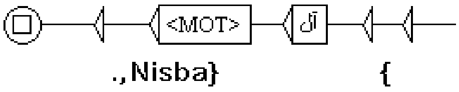


Figure 5. Transducer recognizing the Nisba

The sub-graph "Nasab" can also be called in another transducer that recognize a new form of appearance of personality's name.

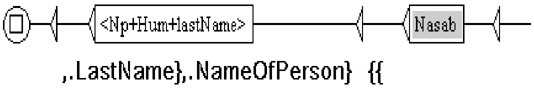


Figure 6. Transducer recognizing the Nasab followed by a last name

Figure 6 shows that the Nasab can be followed by a last name.

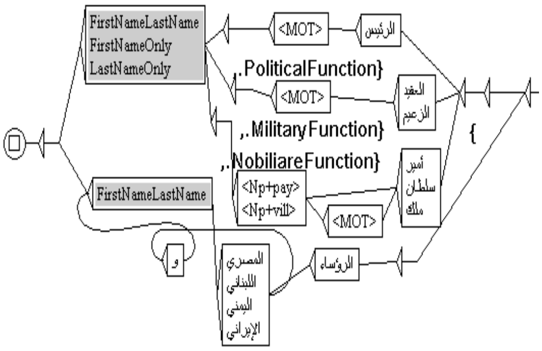


Figure 7. Transducer recognizing exceptional cases

In Figure 7, the transducer treat exceptional cases in the corpus of study. Knowing that those cases are dependent on trigger words.

4.4 Construction of Transducer Cascade

The constructed transducer cascade is based on the following principle: the passage of the main transducers is done in a specific order; labels in output files would enrich the recognized ANE with markup defined into the transducers.

5. Experimentation and Evaluation

As discussed, our prototype is based on the transducer cascade that we have proposed. The general architecture prototype is illustrated in Figure 8.

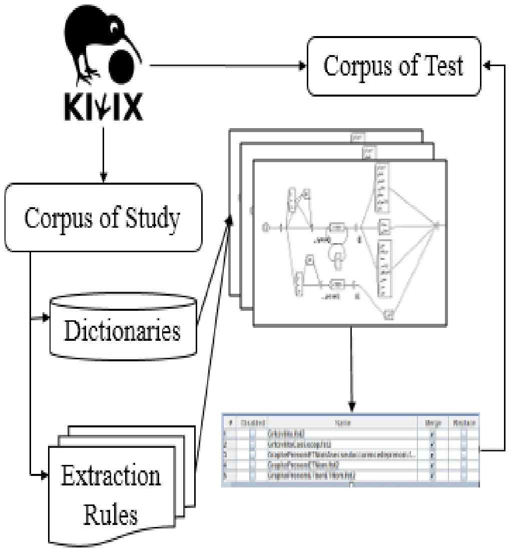
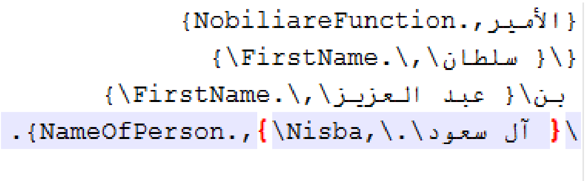


Figure 8. System architecture

Figure 8 shows the system architecture, which describes the steps of our proposed method for the recognition of Arabic personality names. The transducer cascade is applied on the test corpus. The collection of the test corpus is made in the same way as the study corpus presented in Section 3.1. It regroups a number of texts from 19 Arabic countries and contains text files for a cumulative 454 959 tokens.

As an output, we get an annotated corpus. Figure 9 illustrates an Arabic personality name that contains a trigger word. This entity will be annotated as follow: this entity contain a nobiliare trigger word, a Nasab; two first name related by the word "ben" and a Nisba.



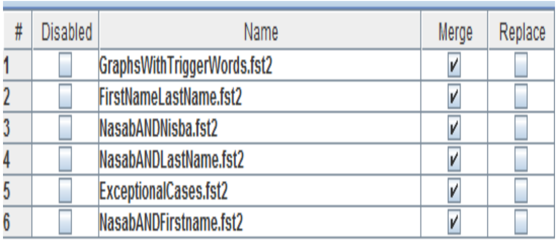
```
{NobiliareFunction.,الأمير}
{\FirstName.\,\سلطان }\}
{\FirstName.\,\العزيز عبد }\بن
.{NameOfPerson.,{\Nisba,\.\سعود آل }\
```

Figure 9. Annotation of an Arabic personality's name

In addition to our dictionaries, we use the dictionary of proper names elaborated by (Doumi et al., 2013) available under Unitex platform. Table 3 shows the coverage of dictionaries exploited in our recognition process.

Dictionary	Coverage
Proper names	8 353
First names	1 152
Last names	895

Table 3. Coverage of dictionaries



#	Disabled	Name	Merge	Replace
1	☐	GraphsWithTriggerWords.fst2	☑	☐
2	☐	FirstNameLastName.fst2	☑	☐
3	☐	NasabANDNisba.fst2	☑	☐
4	☐	NasabANDLastName.fst2	☑	☐
5	☐	ExceptionalCases.fst2	☑	☐
6	☐	NasabANDFirstname.fst2	☑	☐

Figure 10. Transducer cascade recognizing names of personalities

Figure 10 shows the form of this cascade. The cascade call the six transducers with certain logic. It is generated through the CasSys tool that is integrated in Unitex the free linguistic platform. Moreover, the choice of passing the transducers is not random. First, the cascade must recognize personality's names having trigger words to add certain certitude (transducer 2). Then, we move to the recognition of personality names which contains first name and last name with one occurrence of the first name (transducer 2) and the recognition of Nasab followed by Nisba (transducer 3) or Last name (transducer 4). Afterward, exceptional cases must be recognized (transducer 5). Finally, we finish the recognition process by the recognition Nasab followed by a first name when the word "ben" is omitted (transducer number 6).

Every graph adds annotations to the text using the mode "Merge". This mode provides, as output, a recognized NE surrounded by a tag defined in defined in the boxes output in the transducer.

Recall	Precision	F-measure
0.98	0.94	0.95

Table 4. Table summarizing the measure values

We manually evaluated the quality of our work on the Wikipedia corpus. This evaluation is performed by evaluation metrics that are the precision, recall and F-measure. These measures are illustrated in Table 4.

The precision is the number of correct ANE for personality names recognized on the total of recognized ANE for personality names. Applying this formula, we get the value 0.94.

The recall is the total correct ANE for personality names recognized on the total ANE for personality names. Applying the formula, we get the value 0.98.

The F-measure is a combination of Precision and Recall for penalizing the large inequalities between these two measures. It is 2*P*R/(P+R). Applying this formula, we get the value 0.95. Therefore, we find that the results for the proposed method are motivating.

	Our system	(Shaalan and Raza, 2007)	(Elsebai and al., 2009)
Precision	94 %	85 %	93 %
Recall	98 %	89 %	86 %
F-measure	95 %	87.5 %	89 %

Table 5. Evaluation between Systems recognizing the type name of person

Table 5 shows an evaluation between our system and those developed by (Shaalan and Raza, 2007) and (Elsebai and al., 2009). We can remark that the results obtained by our system are efficient measures as those of the other two systems.

6. Conclusion and Perspectives

In this paper, we presented a method for recognizing ANE based on transducer cascade. We established a set of dictionaries, a list of extraction rules depending essentially on trigger words and a set of transducers allowing the recognition of several ANE categories. We gave also an experimentation on Wikipedia test corpus fitted with kiwix tool. The obtained results are satisfactory because the calculated measure values are encouraged.

As perspectives, we will improve our dictionaries by adding other features. Then, we will experiment the generated cascade on other types of ENA having relationship with personality's name. Finally, we are going to take advantage of our annotated corpus to develop an enrichment process to establish links to free resources such as Wikipedia and Geonames and to disambiguate them if needed.

References

Aboaoga M. and Aziz MJA. 2013. Arabic person names recognition by using a rule based approach. Journal of Computer Science, 922–927.

Ciravegna F. and Lavelli A. 1999. « Full text parsing using cascades of rules: An information extraction perspective », In Proceedings of EACL'99, Bergen, Norway, 102-109.

Elsebai A., Meziane F. and BelKredim FZ. 2009. A rule based Persons names Arabic extraction system. Communications of the IBIMA, 11: 53–59.

Elsebai A. 2008. Arabic Proper Names Recognition Using Heuristics. Proceeding of the 9th Annual Post Graduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNET), ISBN: 978-1-902560-19-9.

Ben Mesmia F., Friburger N., Haddar K. and Maurel D. 2015. Transducer cascade for an automatic recognition of Arabic Named Entities in order to establish links to free resources. Will appear in IEEE-proceedings issue from CICLING’15.

Ben Mesmia F., Friburger N., Haddar K. and Maurel D. 2015. Construction d’une cascade de transducteurs pour la reconnaissance des dates à partir d’un corpus Wikipédia. Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications, 8-11.

Fehri H., Haddar K., Hamadou A. B. 2011. Recognition and Translation of Arabic Named Entities with NooJ Using a New Representation Model, in M. Constant, A. Maletti, A. Savary (eds), FSMNLP, 9th International Workshop, ACL, Blois, France, 134-142.

Friburger N., Maurel D. 2004. Finite-state transducer cascades to extract named entities in texts, Theoretical Computer Science, volume 313, 94-104.

Doumi, N., Lehireche, A., Maurel, D., and Ali Cherif, M. (2013a). La conception d'un jeu de ressources libres pour le TAL arabe sous Unitex.Paper presented at the TRADETAL2013, Colloque international en Traductologie et TAL, Oran - Algeria, 5-6 may.

Maurel D., Friburger N., Eshkol I. 2009. « Who are you, you who speak? Transducer cascades for information retrieval ». In Proceedings of 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznan, Poland, 220-223.

Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I. and Nouvel D. 2011. Cascades de transducteurs autour de la reconnaissance des entités nommées. Traitement automatique des langues, 52(1) : 69-961.

Mesfar S. 2007. Named Entity Recognition for Arabic Using Syntactic Grammars. Proceedings of the 12th International Conference on Application of Natural Language to Information Systems. Berlin, Heidelberg, 305-316.

Shaalan K. and Raza H. 2007. Person name entity recognition for Arabic. In: Proceedings of the 5th workshop on important unresolved matters, 17-24.

Shaalan K. and Oudah M. 2014. A hybrid approach to Arabic named entity recognition. Journal of Information Science, 40(1): 67–87.

Shaalan K. 2014. A Survey of Arabic Named Entity Recognition and Classification. Computational Linguistics, 40 (2) 469-510.