A Rule-based Named Entity Extraction Method and Syntactico-Semantic Annotation for Arabic Language

Lhioui Chahira
LaTICE Laboratory
ISI, Sousse University
Medenine, Tunisia
chahira.lhioui@ieee.org

Zouaghi Anis
LaTICE Laboratory
ISSAT, Sousse University
Sousse, Tunisia
anis.zouaghi@gmail.com

Zrigui Mounir
LaTICE Laboratory
FSM, Monastir University
Monastir, Tunisia
mounir.zrigui@fsm.rnu.tn

Abstract— **There is a widely held belief in the natural language processing (NLP) and computational linguistics communities that knowledge recognition such us Named Entities (NE) recognition is a significant step toward improving important applications, e.g., question answering and natural language understanding (NLU). In this paper, we present an NE recognition system for Modern Standard Arabic using the NooJ platform. This system exploits many aspects of the rich morphological features of the language. The experiments on the pilot Arabic Propbank data show that our system based on linguistic rules produces a global NE recognition F-measure score of 87%, which improves the current state of the art in Arabic NE recognition.**

Keywords- Named Entity Extraction; Semantic annotation; NooJ platform.

I. INTRODUCTION

The extraction and automatic recognition of named entities (NE) is a part of a syntactico-semantic analysis, which is a step that follows the morpho-lexical analysis during the automatic processing of a text or a corpus. This extraction consists in exhibiting certain grammatical concepts or syntactic structures, checking their validity and attesting their belonging to particular grammatical classes such as ”proper names”, ”temporal expressions”, ”numerical expressions”, ”abbreviations” , etc.

From the beginning, the implementation of the lexicographical solution, subsisting of electronic dictionaries enumerating all the named entities, has proven to be impossible. In particular, this is due to the problems of multiple writing and the lack of standard writing or transcription of NE, especially those of foreign origin, to the target language. Indeed, it is impossible to enumerate all the proper names in lists, as well as to collect and to maintain them. It is also impossible to treat all spelling variants and to resolve the resulting ambiguity.

Three fundamental approaches have been used for the extraction of NE issue in literature. These approaches are: rule-based approach, learning-based approach and hybrid approach. However, the most commonly used methods for NE recognition are often machine learning-based methods. In the last two decades, rule-based methods for NER (Named Entity Recognition) have progressively been abandoned. Nevertheless, these methods are robust and their results are accurate. They are generally based on non-contextual grammars. Thus, our major concern in this study is to examine a rule-based NE extraction and syntactico-semantic annotation of such important knowledge. For this purpose, we use the non-contextual grammars offered in the NooJ language development platform [10] where they are called local grammars that are used to locate in a very precise way local phenomena very precisely in texts, such as dates, numerical determinants, proper names, names of places and organisms, etc. These grammars are lexicalized graphs [10], which use dictionaries of simple and compound words. They are equivalent to recursive networks of transitions (RTN) or even networks of increased transitions (ATNs). In practical, local grammars are graphs that can call independent sub-graphs. Among the advantages of such a structure are the effectiveness of its direct application to texts, the recognition of complex linguistic concepts as well as transformational analysis and annotations production.

The choice of NooJ platform is guided by the fact that NooJ is a freely available linguistic development environment for many languages [1]. It allows developers to construct, test and maintain large coverage lexical resources as well as to apply morphological morpho-syntactic tools for Arabic processing [1]. NooJ can recognize rules written in finite-state form or context-free grammar form, facilitating the development of rule-based NER systems. Nooj provides a disambiguation technique based on grammars to resolve duplicate annotations [1]. Arabic is one of the languages that are supported by NooJ; there are free Arabic resources for use within the NooJ environment on the NooJ official Web site [1]. Mesfar [5] and Lhioui [3][15] have also used NooJ in their Arabic NER research..

In this paper, we suggest a Named Entities extraction system for Modern Standard Arabic (MSA) that exploits many aspects of the rich morphological features of the language. It is based on a linguistic approach that uses NooJ technology for the detection of such knowledge. Given the lack of a reliable electronic Arabic dictionaries, and thanks to their coverage, our strategy uses the EL-DicAr dictionary [2] developed by the NooJ platform and its extension developed in [3] for the step of morphological analysis.

 ISBN: 978-1-61208-552-4

In this article, we begin by presenting some of the existing work on Arabic NE extraction. Then, we describe the difficulties inherent in the recognition of NE. After that, we explain with more details our preconized approach. Finally, we check and evaluate our proposed approach.

This paper is laid out as follows: Section 2 presents the definition of named entity concept and its categorizations; Section 3 outlines different approaches that treat this problematic and some related works; Section 4 reveals some difficulties that inhibit the extraction of NE in texts written in Arabic language; Section 5 describes and argue the approach and system adopted for this work; Section 6 gives the experimental setup, results and discussion. Finally, Section 7 draws our conclusions.

II. THE NAMED ENTITY CONCEPT DEFINITION AND CATEGORIZATIONS

The extraction of NE is one of the most popular areas in recent years. According to MUC (Message Understanding Conference) [4], we distinguish at least three types of entities to be recognized and classified by category [2][3]:

- ENAMEX: This class groups the proper names. Indeed, proper names are very common in electronic texts, especially journalistic articles. However, in spite of the frequency of their appearances and the importance of the information they encapsulate in particular for the semantic interpretation of the texts, the proper names remain inadequately illustrated in the electronic lexical resources and their automatic extraction is just only a relatively young field. This class contains at least three subcategories:
  - Person: Names of persons such as names of politicians, poets, athletes, etc.
  - Organization: refers to the names of companies, banks, associations, universities, research centers, pharmacies, clinics, etc.
  - Event: such as sporting events, political events, war and crime event, etc.
- NUMEX: This class groups numeric expressions of percentages, size, currency expressions, etc.
- TIMEX: This class refers to temporal expressions of date or duration.

III. RELATED WORK

Numerous studies have been conducted on the Latin languages as well as the Arabic language to automatically extract knowledge. Looking over the state of the art, we have found that there are three main types of extraction systems of named entities. These systems are based on three types of approaches, which are, respectively:

- Rule-based approach: Most systems use this approach. Typical rule-based systems use both internal and external evidence, as well as word-trigger dictionaries for locating help. The rules are manually built by an expert linguist. The advantages of such approach are principally the accuracy, the robustness and the coverage of the obtained results. In brief, this kind of approach is has been well appreciated so far in literature [3][5]-[7].
- Learning approach: Systems based on this approach use stochastic techniques and learn specific knowledge on a large learning corpus where target NEs are labeled. Learning algorithms are then applied automatically to develop a NE base using several statistical models (such as Hidden Markov Model (HMM), Support Vector Machine (SVM), Conditional Random Fields (CRF), etc.) [8][9]. Nevertheless, this approach requires a huge amount of learning data for its learning algorithm, which is quasi-absent for some scientific research neglected languages, such as Arabic [1].
- Hybrid approach: This approach combines the two above-mentioned approaches for their complementarity. This approach leads to systems based on the use of both manually-written and rules that are constructed automatically using syntactic and contextual information derived from training data to learning algorithms and decision trees [11][12].

The adequacy of rule-based systems was recognized at the MUC conference. It is this same technique that we advocate for the development of a recognition component of named entities. This component is based on rules written by hand and represented in the form of local grammars that are constructed using the syntactic module of NooJ. These rules were based on internal and external evidence in order to identify and categorize named entities where:

- Internal evidence: is provided by the constituents of the named entity. The constituents can be contained in lists of triggering words or proper names called gazetteers.
- External evidence: is provided by the context in which, a named entity appears. They are based on the syntactic relations within a sentence to assign the category of such an entity. This categorization uses the morpho-syntactic information provided by the previous morphological analysis stage.

The use of this evidence is indispensable because of the absence of obvious indications to detect the presence of a proper name, such as the presence of capital letters at the beginning of such names in the Romance languages. This imposes a rather thorough understanding of the morphological nature of each form of the text, particularly its grammatical categories and semantic information (e.g., + Person, + Country, + Housing, + Money, etc.).

 ISBN: 978-1-61208-552-4

IV. ISSUES IN NAMED ENTITY RECOGNITION

According to the state of the art overfished by [3][5] [13], the recognition of the TIMEX and NUMEX in Arabic poses no problem, this can be challenging in the case of the ENAMEX. This can be explained by the lack of structural or contextual indices. In fact, all temporal and numerical expressions are identifiable by a list of lexical markers (day names, month names, currencies, units of measure, etc.). On the other hand, ENAMEX suffers from a lack of structural or contextual clues to be recognized.

Moreover, in addition to the absence of capital letters as a naïve index for recognition in Latin languages, Arabic ENAMEX requires linguistic information to be dynamically generated by a prior semantic annotation step.

Other problems specific to the recognition of the NE in Arabic arise also in the identification and delimitation as in the semantic annotation of these NE. In what follows, we depict the repercussion of the absence of voyellation and the problem of delimitation of the Arabic NE.

A. *The absence of vowels*

The absence of diacritical marks may affect the recognition systems of named entities. This is mainly due to the semantic ambiguity that arises from the set of potential vocalizations that can be attributed to any partial vowel or unvoiced form. Indeed, vocalizations accepted for any form of text can lead to the absence of diacritical signs and it can affect the recognition systems of named entities. This is mainly traceable to the semantic ambiguity that arises from the set of potential vocalizations that can be attributable to any partial vowel or unvoiced form. Indeed, vocalizations accepted for some form of text can lead to different triggers of NE. For example, the unbounded form معلّم (m‘llm) can accept, among other things, the two following vocalizations: in different senses the triggers of the NE.

- مُعَلِّمٌ (mu‘allimu) : Word trigger for a teacher
- مَعْلَمٌ (ma.‘alamu) : Word trigger for a museum of monuments

This example illustrates the implications of the absence of vowels in the text words on the annotation step of the named entities.

B. *Morphological complexity*

Arabic is a highly-inflected language. It uses an agglutinative strategy to form a word. If NE appears in its agglutinative form, then this poses a difficulty for the identification and hence the recognition of this entity [3]. For example, if we take the simple word بَلْدَتُنَا <baldatunA>, which means ”our town”, this Arabic word is composed from two sub-words: the lemma بلد <balda> ”town” and the suffix تنا <tunA> ”our”. Hence, it would be difficult and ambiguous in Arabic processing to treat agglutinative words. Many works focus on this phenomenon. However, NooJ gives the possibility to treat agglutination problem by the use of flexional and derivational rules [10]. Hence, the choice of NooJ linguistic tool, in our work, is justified.

V. OUR RULE-BASED METHOD FOR NE RECOGNITION AND SEMANTIC ANNOTATION

To remedy all these problems, we construct a system of recognition and extraction of Arabic entities. According to Figure 1, we proceed:

- A morphological analysis: to collect the maximum information for all the words of the text. This is done with a consultation of the electronic dictionary El-DicAr of [2] and the Arabic touristic dictionary developed by [3].
- Subsequently, this information will be used in local syntactico-semantic grammars in order to locate the relevant sequences.

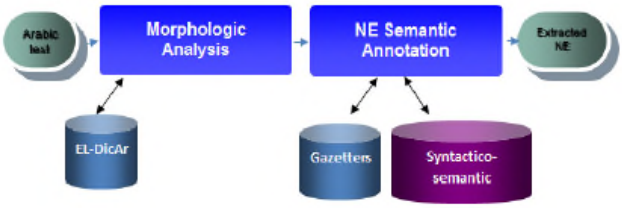

Figure 1. General architecture of the recognition of the Arab NE

A. *Morphological Analysis*

Given the agglutinating structure of the majority of Arabic words, our morphological analyzer makes it possible to separate and identify the morphemes of the input forms and to associate them with the set of information necessary for the current processing. These forms are decomposed to recognize the affixes (conjunctions, prepositions, personal pronouns, etc.) attached to them. These morphological possibilities in NooJ facilitate the identification of triggering words, names of persons or localities even when they are agglutinated.

Each of these forms is associated, by morphological analysis, with a set of linguistic information useful for the next step: lemma, grammatical label, gender and number, syntactic information (+ Transitive), semantic information (+ Person), etc.

Consequently, instead of enumerating all the inflected forms (singular, dual, plural, masculine, feminine) of the occupational names considered as lexical markers of person names (e.g., مهندس <engineer>), we use the syntax of the regular expressions of NooJ where the grammatical symbol مهندس (<mhnds>, <engineer>) refers to all vocalized, partially vocalized, and unvoiced bent forms attached to this lemma. Our morphological analysis is based on two Arabic dictionaries described in table III:

TABLE I. RESOURCES USED IN MORPHOLOGICAL ANALYSIS.

	EL-DicAr [2]	**’Touristic Arabic DICtionary [3]**
Nouns	19504	8789
Verbs	10162	345
NE	3686 localizations +11860 Proper names	622 500 (Organisations +Localizations+Events)

The two dictionaries are also used and detailed in [16].

B. NE Semantic Annotation

The information provided by morphological analysis is directly used by our recognition system of named entities. In addition to its morpho-syntactic information gathered, this system is based on the use of two types of linguistic resources:

- Gazetteers: these are lexical markers previously recognized as potential members of properly named and properly classified entities. Among these, we perceive:
  - Names of persons
  - Names of places: countries, cities, regions, states, names of roads, seas, oceans, mountains, rivers, etc.
  - Names of organizations: associations (regional, national and international), universities, televisions, banks, etc.
  - Currency expressions: cost, money, etc.
  - Temporal expressions: the names of the days of the week, months, etc.
- Local Grammars: These are represented in the form of Augmented Transition Network-ATNs. They are used to represent sequences of words. These sequences are described by manually written rules and consequently produce certain linguistic information such as the type of the identified named entity (person name, organization, location, etc.).

Figure 2 shows the main graph of NE represented with NooJ linguistic platform. The same graph contains embedded sub-graphs.

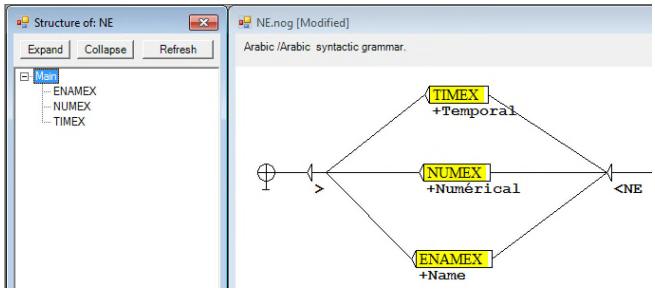

Figure 2. Main graph embedded the three types of NE: TIMEX, NUMEX and ENAMEX

1) Local grammars for the extraction and annotation of NUMEX: The problem of automatic recognition of numerical determinants in a text is part of the more or less complex linguistic phenomena. Generally, they can not be processed at the level of lexical analysis. They require very redundant descriptions that would be very tedious, if not impossible, to describe them manually in electronic dictionaries compiled in the form of finite automata.

We have classified numerical expressions into four categories: percentage expressions, weight expressions, measurement expressions, and monetary expressions (see Figure 3).

Then, we focused on the extraction of numerical values.

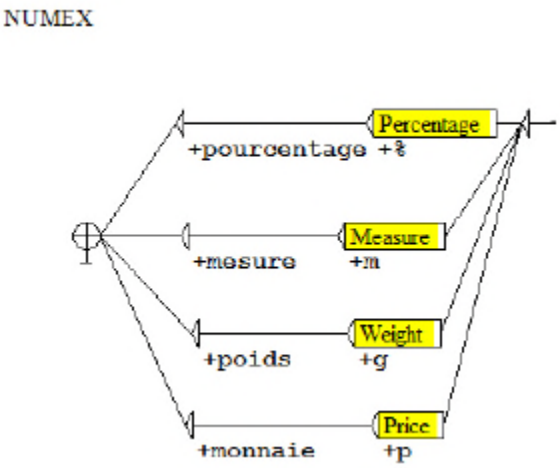

Figure 3. Main sub-graphs of the different types of NE.

Figure 4 shows the main recognition graph of these values. This is restricted to call to sub-graphs relating to the identification of numerals representing units, tens, hundreds and thousands. As outputs, we attribute the grammatical category ”DET” (a determinant), the semantic information ”+NUM” (numerical) as well as the arithmetic value that it represents ”+Val”. Thus, each recognized numeral occurs with its equivalent written in numbers.

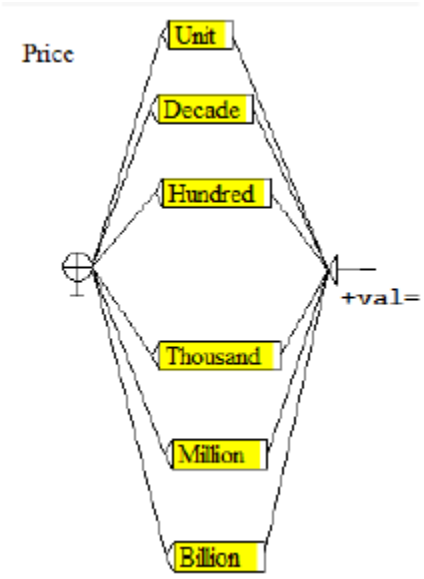

Figure 4. Main sub-graph describing the local grammar responsible for the extraction of numerical values

2) Local grammars for extraction and annotation of TIMEX: Temporal expressions, TIMEX, are as important as numerical expressions in syntactic or information extraction systems. Indeed, a user can query our system to get information about an event. Usually, any event is linked to a date or time represented as a time expression. As a result, according to Figure 5, our rule-based system allows the extraction of ages, hours, dates and periods.

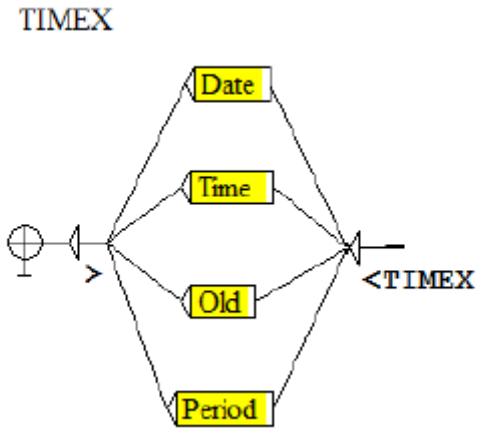

Figure 5. Main sub-graph describing the local grammar responsible for the extraction of TIMEX

3) Local grammars for extraction and annotation of ENAMEX: In our work, the ENAMEX extraction means the extraction of proper names , localizations and organizations. Figure 6 shows the NooJ local grammar [10], which is responsible for the syntaxic-semantic annotation of different ENAMEX type.

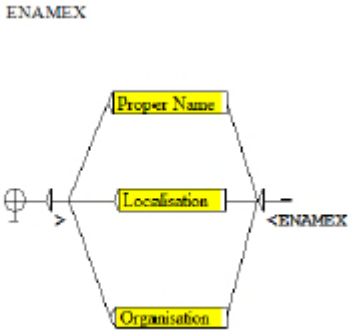

Figure 6. Main graph describing the local grammar responsible for extracting expressions associated with ENAMEX

For the names of places, we began by developing a grammar of internal proofs associated to the cities name, regions, hotels, itineraries, avenues, rivers, seas, oceans, etc. Thus, we identified the triggering words as مدينة (<mdiynT>, <city>), جبل (<jbl>, < mountain >), جزيرة (<jzIrT>, <island>),دولة (<dwlT>, <country>), نهج (<nhj>, <avenue>).

These lexical (triggers) markers are used to describe recognition rules in local grammars. Figure 7 shows the main graph of local grammar responsible for the extraction of localization expressions. In the same manner, this grammar is implemented with linguistic NooJ platform.

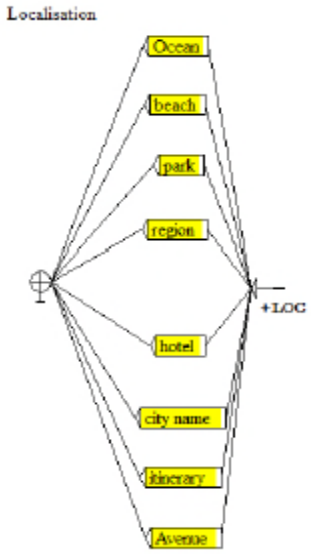

Figure 7. Main sub-graph describing the local grammar responsible for extracting location expressions

Besides the places name, we made the recognition grammar of people names. Figure 8 below shows a graphical implementation of proper names grammar in NooJ platform.

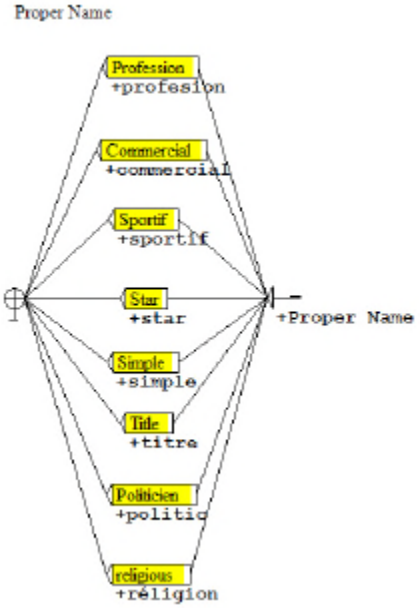

Figure 8. Main sub-graph describing the local grammar responsible for extracting proper names expressions

The identification of organizations names began with the elaboration of a dictionary, which contains 959 organizations names such as: يونسكو (<yUnskw>, Unesco) recognized by the mean of (N + Org) syntactico-semantic annotation or what we call lexical markers. We note that the majority of entries are compound and abbreviated words. Then, we have a list of 626 trigger words like مؤسسة (<m'wssasT>, company), جمعية (<jm'yT>, association), and so on. These lexical markers are used to describe recognition rules in local grammars. In total, we have ten sub-graphs that implement recognition global grammar of organizations name (cf. Figure 9)

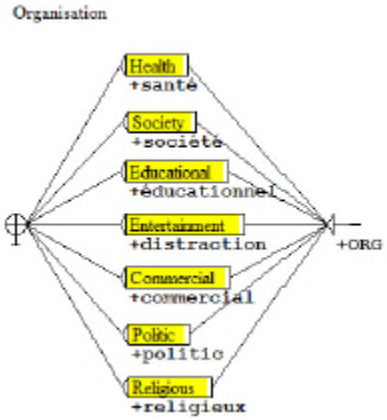

Figure 9. Main sub-graph describing the local grammar responsible to extracting names and abbreviations for organizations

VI. EVALUATION OF THE NE EXTRACTION SYSTEM

After collecting the corpus, we had to go to experimentation. This step seems to be the most important one because it measures the reliability of the work.

The experimentation of our resources was done with NooJ concordance [10]. As mentioned before, this platform uses (syntactical, morphological and semantic) local grammars already built.

Traditionally, the evaluation of any information retrieval system relies on the computation of a set of metrics. These calculations make it possible to evaluate the proportion of the errors displayed by the system relative to the ideal result.

The metrics usually used are: Recall (R), Accuracy (P), F-Measure (F).

We evaluate our recognition system on 70% of the Arabic PropBank [14] and a 70% of our own corpus described in [3] (see Table II). The rest of these corpora is used for the test.

An evaluation carried out on these corpora gives the results presented in Table III.

Our syntactico semantic recognizer yields F-scores included in the interval of [76%-96%] which are satisfying measures compared to [8], [12] and [14].

TABLE II. RTRH [3] TOURISTIQUE CORPUS

Corpus dialogue number	4000
Cities and towns	3120
Restaurants	9130
Itineraries appellations	3100
Locations	6125
Organizations	9125
Persons names	4125
Entertainments	6150
Localizations	8125
Transport fields	6120
Specialties	1130
Hotel and restaurant categories	1125
Contacts	6125

TABLE III. EVALUATION OF NE SYSTEM

		Precision	Record	F-Measure
TIMEX		97%	95%	96%
NUMEX		97%	94%	95.5%
ENAMEX	**Proper Names**	92%	79%	85%
	Organisation Names	90%	78%	84%
	Location Names	82%	71%	76%

VII. CONCLUSION

We have described a system for extracting proper nouns, temporal and numerical expressions through a combination of morphological analyzer and a rule-based recognition system using local NooJ grammars. This permitted to achieve performance by providing lexical coverage in more than 87%. Despite the above-described problems, the recommended method seems to be adequate and exhibits very encouraging extraction rates.

REFERENCES

[1] K. Shaalan, "A Survey of Arabic Named Entity Recognition and Classification," Computational Linguistics, 40 (2), pp. 496-510 , 2014.

[2] S. Mesfar, "Analyse Morpho-syntaxique automatique et reconnaissance des entits nommes en arabe standard," A Doctorat thesis, vol. Franche- Compt University, 2008.

[3] L. Chahira, Z. Anis, and Z. Mounir, "Knowledge Extraction with NooJ Using a syntactico-Semantic Approach for the Arabic Utterances Understanding," CICLing, Konya, 2016.

 ISBN: 978-1-61208-552-4

[4] "MUC," 2014, URL: http://en.wikipedia.org/wiki/Message Understanding Conference [retrived: 10, 2016].

[5] S. Mesfar, "Named entity recognition for Arabic using syntactic grammars," In Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems pages pp. 305-316, Berlin, 2007.

[6] B. Siham, T. Meryem, and A. Driss, "Named Entity Recognition using a A rule based approach," Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on, pp. 478-484, DOI 10.1109/AICCSA.2014.7073237, 2014.

[7] A. Sherief, K. Shaalan, and M. Shoaib, "Integrating rule-based system with classification for Arabic named entity recognition," In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 7181 of Lecture Notes in Computer Science, vol. Springer, Berlin Heidelberg, 2012, pp. 311-322.

[8] B. Mohit, S. Nathan, B. Rishav, K. Oflazer, and S. Noah, "Recalloriented learning of named entities in Arabic wikipedia," In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2012.

[9] R. P. Valetta-Malta. Y. Benajiba, M. Diab, "Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition," The International Arab Journal of Information Technology, Vol. 6, No. 5, November 2009, pp. 464-472, 2009.

[10] M. Silberztein, Ed., Formalizing Natural Languages: The NooJ Approach. Wiley-ISTE, Jan. 2016, ISBN: 978-1-84821-902-1.

[11] Z. Ins, H. Souha, Mezghani, and B. Lamia, Hadrich, " The contribution of a hybrid approach to the recognition of Arabic-language entities," TALN Montral, 19-23 juillet, 2010.

[12] S. Abuleil, "Hybrid System for Extracting and Classifying Arabic Proper Names," Proceedings of the WSEAS Int.Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid-Spain, pp. 205-210, 2006.

[13] H. Fehri, K. Haddar, and A. Ben Hamadou, "Recognition and translation of Arabic named entities with NooJ using a new representation model, In Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing (pp. 134-142). Association for Computational Linguistics. July, 2011.

[14] Palmer, M., Babko-Malaya, O., Bies, A., Diab, M. T., Maamouri, M., Mansouri, A., & Zaghouani, W. (2008). A Pilot Arabic Propbank. In *LREC*.

[15] L. Chahira, " Development of an automatic spoken language understanding system of spontaneous Arabic speech based on an hybrid approach, linguistic and stochastic approach," A Doctorat thesis, vol. Faculty of Economics and Management of Sfax University, LaTICE Laboratory, Tunis 2017.

[16] L. Chahira, Z. Anis, and Z. Mounir, "Knowledge Extraction with NooJ Using a syntactico-Semantic Approach for the Arabic Utterances Understanding," CICLing, Konya, 2016.

 ISBN: 978-1-61208-552-4