Automatic Arabic Named Entity Extraction and Classification for Information Retrieval

Omar Asbayou

AUTOMATIC ARABIC NAMED ENTITY EXTRACTION AND CLASSIFICATION FOR INFORMATION RETRIEVAL

Omar ASBAYOU

Department of LEA, Lumière Lyon 2 University, Lyon, France

ABSTRACT

This article tries to explain our rule-based Arabic Named Entity recognition (NER) and classification system. It is based on lists of classified proper names (PN) and particularly on syntactico-semantic patterns resulting in fine classification of Arabic NE. These patterns use syntactico-semantic combination of morpho-syntactic and syntactic entities. It also uses lexical classification of trigger words and NE extensions. These linguistic data are essential not only to name entity extraction but also to the taxonomic classification and to determining the NE frontiers. Our method is also based on the contextualisation and on the notion of NE class attributes and values. Inspired from X-bar theory and immediate constituents, we built a rule-based NER system composed of five levels of syntactico-semantic combination. We also show how the fine NE annotations in our system output (XML database) is exploited in information retrieval and information extraction.

KEYWORDS

Morphosynyaxique analysis, syntactico-semantic patterns, rule-based system, fine annotation and classification, information retrieval, information extraction.

1. INTRODUCTION

NE extraction is a widely studied subtask of information extraction. They are essential for many natural language processing applications such as information retrieval, information extraction, machine translation, strategic foresight, question-answering systems etc. Relatively little work has been done on Arabic NE extraction and classification (especially rule-based systems).

Three main approches are used in NE extraction process: statistical, rule-based and hybrid approaches. We want to show in this article how a linguistic rule-based system can provide very good results for the Arabic language taking into consideration the linguistic dimension and particularities of this latter (no capital letter, agglutination, no voweling). To try to solve these problems, we use morpho-syntactic analysis to describe words, contextualisation to determine the right context, and syntactico-semantic approach to specify the relations between NE constituents (trigger words and extensions).

The paper is structured into four parts. First, we present our general background of Arabic NE extraction and classification. Second, we describe the linguistic basis of Arabic NE analysis and the Arabic NE structures and constituents; we also show how these latter are combined into single and complex NE. Third, we give a description of our five-level rule-based system. Finally, we expose the evaluation of our NE extraction and classification system based on the treatment of two journalistic corpora, and show examples of the contribution of our system in the field of information retrieval and extraction.

DOI: 10.5121/ijnlc.2020.9601

2. GENERAL BACKGROUND

Three main approaches are used in NE extraction: statistical, based on large training corpus, rule-based, and hybrid, which combines both approaches. We assume that a rule-based approach is interesting in that it lies on a linguistic data.

NE are noun phrases with some linguistic properties. The linguistic approach, on which our rule system is based, is essential for an efficient NE extraction system because it relies on linguistic information: morpho-syntactic, lexical, syntactic, semantic etc. Linguistic information and classification allow recognizing NE properties and characteristics such as:

- NE belongs to a predefined class (PERSON, ORGANISATION, LOCATION, EVENT etc.). For example, **باراك أوباما** الرئيس الأمريكي al-raʾīs al-ʾamryky bārāk ʾūbāma => PERSON.
- NE is a determined noun phrase : For example, determination, by definite article or by PN, is an essential characteristic of NE. M. Ehrmann [2008] evokes referential unicity to define NE. This ‘‘referential unicity’’ is constructed by the definite article (morpho-syntactic information in Arabic) or by PN (lexical information).
- NE is caracterised by ‘‘constrained subordination’’ : some syntactic extensions, such as relative clause, are not compatible with NE structure.
- NE belongs to different lexical and lexico-syntactic types (simple/complex ; descriptive/PN/hybride). For example :

الرئيس الأمريكي باراك أوباما al-raʾīs al-ʾamryky bārāk ʾūbāma (The American President Barack Obama) => PERSON NE_Politics.

The example, illustrates these types of constituent :

Descriptive NE : الرئيس الأمريكي al-raʾīs al-ʾamryky (The American President)

PN → باراكأوباما bārāk ʾūbāma (Barack Obama)

PN → أوباما ʾūbāma (Obama)

Our rule-based system takes into account these properties in NE extraction and classification. Therefore, we used a set of lexical, syntactic, and semantic classifications to build correct syntactico-semantic rules of our system. In other words, our method of NE extraction and classification is based on the linguistic analysis and on the application of the linguistic information describing different linguistic entities:

- *Mopho-syntactic entities:* we use Techlimed morpho-syntactic analyser, which is based on the Arabic lexical resource DIINAR to determine word syntactic categories.
- *Syntactic entities* : our rules distinguish between differnent syntactic structures. The rules are based on binary combinations inspired from X-bar theory (Chomsky, 1967)and from the notion ‘‘immediate constituents’’ (L. Bloomfield, 1933).
- *Semantic entities* : we classified entities in different semantic classes according to semantic relations. We have two types of classification :
  - Simple entities : distinct listes of manually classified words (trigger words, words of semantic fields, PNs). For example, الجمعية al-ğamʿiyya (association) is in the liste of ORGANISATION ‘‘generic trigger words’’ ;الاقتصادية al-iqtisādiyyaẗ (ecnomic) is in the liste of semantic field of economy etc. To build lists of PNs, we used two ways : either exploit availale lists in the web (names of countries, cties, first/last person names, names of geonatural locations etc.) or a semi-automatique method based on Wikipedia dumps definition extraction ; for

example : Y is an German scientist. To illustrate this, the figure below presents the results of scientific person PN extraction from wikipedia.

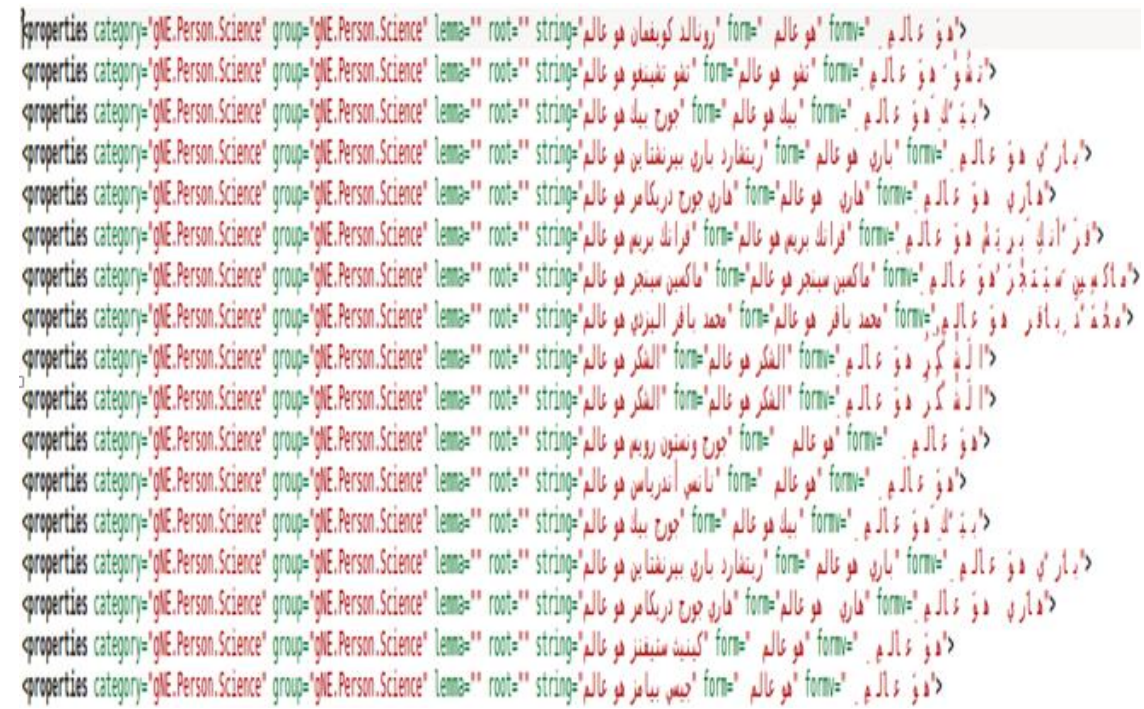


Figure 1. An example of the output of our rules designed for PN lists

3. LINGUISTIC INFORMATION

Our study deals with NE linguistic specificities and the elements involved in their syntactic and semantic structure. These information highlight several levels of analysis.

3.1. Lexical information

Lexical resources and information are the basic element in our NE recognition and classification system. It provides morpho-syntactic and semantic information (i.e. classification) necessary to our rule construction.

3.1.1. Morpho-syntactic analysis

We use an efficient morpho-syntactic analysis system based on DIINAR: a rich arabic lexical database. For example, the word مدينة is analysed as follows:

```
<pos start="0" finish="0" content="مدينة" group="word">
      <interpretation>
        <morphology category="C_NOUN" group="C_NOUN" lemma="مُدَيْن" root="دين" string="مدينة" form="مدينة" formv="مُدَيْنَة">
          <traitNoun gender="Female" number="Singular" mode="Indetermined" case="Nominative"/>
        </morphology>
      </interpretation>
      <interpretation>
        <morphology category="C_NOUN" group="C_NOUN" lemma="مَدِينَة" root="مدن" string="مدينة" form="مدينة" formv="مَدِينَةٌ">
          <traitNoun gender="Female" number="Singular" mode="Indetermined" case="Nominative"/>
        </morphology>
      </interpretation>
      <interpretation>
        <morphology category="C_NOUN" group="C_NOUN" lemma="مُدَيْن" root="دين" string="مدينة" form="مدينة" formv="مُدَيْنَةِ">
          <traitNoun gender="Female" number="Singular" mode="Annexion" case="Genetive"/>
        </morphology>
      </interpretation>
      <interpretation>
        <morphology category="C_NOUN" group="C_NOUN" lemma="مَدِين" root="دين" string="مدينة" form="مدينة" formv="مَدِينَةِ">
          <traitNoun gender="Female" number="Singular" mode="Annexion" case="Genetive"/>
        </morphology>
      </interpretation>
      <interpretation>
        <morphology category="C_NOUN" group="C_NOUN" lemma="مُدِين" root="دين" string="مدينة" form="مدينة" formv="مُدِينَةِ">
          <traitNoun gender="Female" number="Singular" mode="Annexion" case="Genetive"/>
        </morphology>
      </interpretation>
  </pos>
```

Figure 2. Output of our morphosyntactic analysis of "مدينة" (city)

This analysis constitute the base of our syntactic rules.

3.1.2. Semantic classification

The task consists in going beyond the morpho-syntactic analysis (on the word level) to a semantic classification of simple lexical entities resulting in semantic meta-information. It is about a classification based on conceptual and contextual analysis. That is to say, we classified the linguistic entities involved in NE structure, in our linguistic rule system, according to different semantic and conceptual relations. We put these trigger word lexicon into different categories according to common semantic and conceptual criteria. For example, مجلس (council), جمعية (association), هيئة (committee) etc. belong to the same class ORGANISATION (a kind of synset in WordNet lexical classification). This goes for the rest of classes: PERSON, LOCATION, EVENT, ARTEFACT, SUBSTANCE TIME, NUMBER etc.

We subdivided these groups into morpho-syntactic sub-categories to distinguish “determined” and “undetermined” trigger words so that we can use them to build correct syntactic entities. As far as number is concerned, all the trigger words are singular.

Our semantic classification of these simple lexical entities is also based on their syntactico-semantic properties. For example, مجلس majlis (council) and وزير wazīr (minister), have respectively different NE extensions from قنصلية qunsulya (consulat) and ملك malik (king). Therefore, مجلس majlis (council) and قنصلية qunsulya (consulat) belong to different ORGANISATION sub-classes since they participate in different rules.

3.1.3. Syntactico-semantic information

Concerning the NE syntactic complexity, we have taken into account in our syntactico-semantic rule construction, the elements mentioned above: the morpho-syntactic information and the trigger words classification.

In addition, we have take into consideration:
- different syntactic structures and different phrases.
- immediate constituent analysis X-bar theory.

Our study starts from the principles that:
- Trigger words are the heads of the phrase followed by different extensions.
- Each NE class has attributes (a clearly defined lists of entities); and the category ORGANISATION is in the center of NE extraction in that it has "*function"* (classified NP or PP), "*person"*, "*location"*, "*law"* etc (O. ASBAYOU, 2017).
- Complex NE structures are composed of many linguistic entities combined in a defined order.
- NE class attributes are put in a well-determined distribution in NE extraction patterns.

In the syntactico-semantic level, we will also shed light on two important aspects : NE structure constiuents and NE class attributes.

3.2. NE structure constiuents

NE constituents are crucial in knowing the linguistic entities that should be combined in our extraction rules. The structure of NE is divided into two parts: trigger words and extensions.

3.2.1. Trigger word

Trigger words are the contextual entities in the head initial position of NE syntactico-semantic structure. They do not only provide the context but also the first and basic classification properties. We have taken into account different perspectives in constructing the typology of trigger words:

A. Lexical approach

This approach aims to distinguish between:

- *Internal evidence* : they are inherent parts of NE and cannot be detached from them (E.g. المجلس الوطني للتنمية الإقتصادية al-mağlis al-waṭaniy littanmyyaẗ al-iqtisādiyyaẗ (the National Council for Economic Development)
- *External evidence* : they are not essential to establishing the autonomy of the NE in which they are trigger words ; they are optional in the NE structure (E.g. مدينة نيويورك madīnat nyū yūrk (The city of New York).

B. Syntactic approach

As far as complexity of NE structure is concerned, the syntactic approach allows us to distinguish between two types of trigger words:

- *Simple trigger words*: they are simple words functioning as heads of the NE.
- *Complex trigger words* : this class is the output of rules made up of :

- Simple trigger words, which are heads of noun phrases, modified by nationalities or numbers. This pattern concerns organisations (المجلس المغربي al-mağlis al-maġriby) (the Moroccan Council), events (الدورة الخامسة عشرة al-ddawraẗ al-ẖāmisa ʿašraẗ; المؤتمر الخليجي al-muʾtamar al-ẖalyğy) (The fifteenth session; Golf meeting), persons (النائب الأول al-nāʾibu alawwalu) (The first representative).
- Simple trigger words (head) modified by a PN (e.g. مجلس فرنسا mağlis faransā France Council; دورة برلين dawraẗ birlyn Berlin Session).

These elements simplify the combination of the trigger words and the appropriate NE extensions. However, these trigger words are annotated as autonomous NE in case they do not have any extension in context (e.g. المجلس الخليجي al-mağlis al-ẖalyğy Golf council; مجلس فرنسا mağlis faransā France Council;دورة برلين dawraẗ birlyn Berlin Session).

C. Semantic approach

In addition to the classification of trigger words according to referential classes (PERSON, ORGANISATON, LOCATION, EVENT etc.), the semantic approach makes a conceptual distinction between two categories of trigger words:

- *Specific trigger words* : their sense is rather specific and allows a fine classification (class + field). In other words, their head position is enough for a specific classification and the extension is no longer used to refine the entity's annotation. For example :

b) **كلية** الآداب والعلوم الإنسانية kulliyyaẗ al-ʾādāb wa al-ʿulūm al-ʿinsāniyyaẗ (the Faculty of letters and human sciences) => (ORGANISATION NE_Science)
d) **وزير** الطاقة والمعادن wazyr al-ṭṭāqa wa al-maʿādin (the minister of energy and mineral resources) => (PERSON NE_Politics)

- *Generic trigger words* **:** their sense is larger giving only the general class. For example :

a) **المجلس** القومي al-mağlis al-qawmy (The National Council)
b) **المنظمة** العالمية للصحة al-munaẓẓamaẗ al-ʿālamiyyaẗ lissiḥaẗ (the World Health Organisation)
c) **رئيس** الوكالة الدولية للطاقة raʾīs al-wakālaẗ al-ddawliyyaẗ liṭṭāqa (The President of Word Energy Organisation)

The generic trigger words in these examples produce the first level of classification (PERSON, ORGANISATION, LOCATION etc.). NE annotations are specified by semantic field denoted by the entities in the extension. For this reason, to refine our classification we use the extensions that contain semantically classified entities (morpho-syntactic or syntactic). For example, in b) above, the entity للصحة (of health) determines the NP/PP field *health, which* is not denoted by the generic trigger word المنظمة (the organisation) but by the NE extension.

3.2.2. NE extensions

The NE extensions concern the morpho-syntactic or syntactic entities occurring after trigger words, they, mark the frontiers of the extracted NE and specify their sub-classes. Studying the extension is very useful in solving the problems of NE frontiers and classification.

- *Morpho-syntactic extension* :

a) **الأمن** مجلس mağlis alʾamn (the Security Council)
b) **الروسي** الرئيسal-raʾīs al-rrūsy (The Russian President)

These morpho-syntactic extensions are either annexion (complement) a) or identification (modifier) b); they have to be determined by the definite article.

The extension can also be semantically classified as in a) denoting the semantic field of security and allowing for the NE sub-classification (ORGANISATION NE_ Security).

- *Syntactic extension* :

a) Noun phrase (NP) : e.g. مجلس **التنمية الإقتصادية** mağlis al-ttanmyyaï al-iqtisādiyyaï (The Organisation of Economic Development)
b) Prepositional phrase (PP): e.g. المجلس الوطني **لحقوق الإنسان** raʾīs al-mağlis al-waṭaniy liḥuqūq al-insān (the National Council of Human Rights).

- *Lexical extension : PN*

a) الرئيس (الأمريكي) **باراك أوباما** al-raʾīs (al-ʾamryky) bārāk ʾūbāma (The American President Barack Obama)

b) شارع **شارل دوكول** šāriʿ šārl dūgul (Charles De Gaulle Street)

4. LEVELS OF SYNTACTICO-SEMANTIC RULE CONSTRUCTION

Our system is made of many levels of analysis. The output of the preceding level is the input of its following one. The diagram below shows this structure:

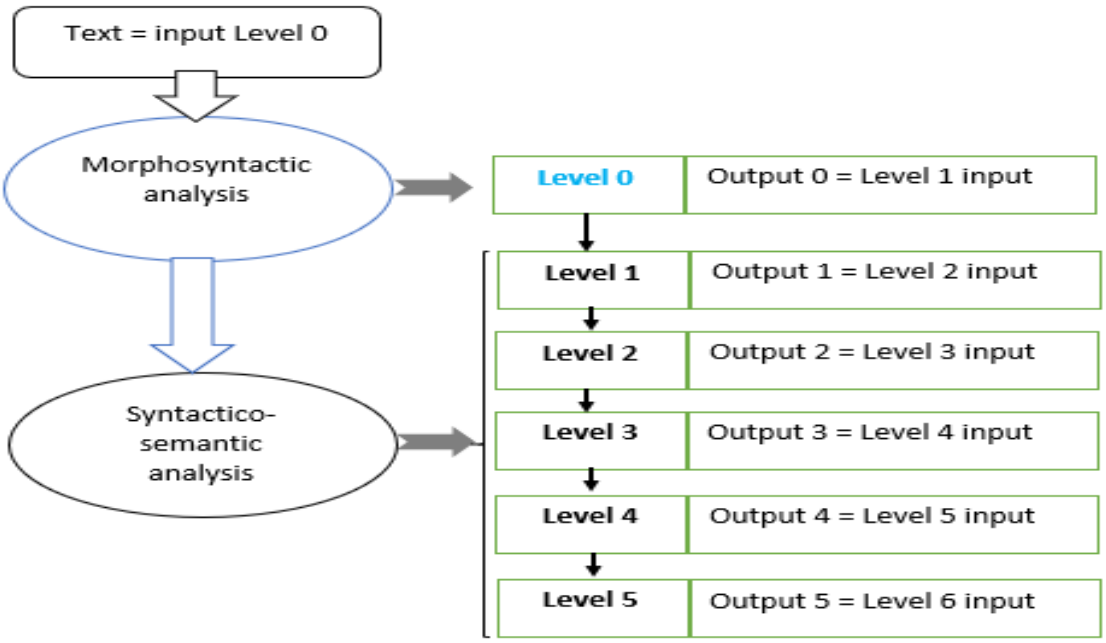


Figure 3. the levels of our NE extraction and annotation

Level 0 applies the morpho-syntactic analysis which we mentioned above and which is the first and basic step in our system. Then follows, as we have already alluded to in the pages above, our syntactico-semantic rule construction composed of 5 levels. The posterior levels exploit the output of the anterior ones.

4.1. Level 1

Level 1 contains a set XML documents each of which deals with particular type of entity.
Note: the syntactic entities in Level 1 are NPs with modification by adjective(s) (see the examples below).

- Document 1 *: Simple (morpho-syntactic) and complex semantically classified NPs and PPs with modification by the adjective(s).* For example :

E.g.الصناعة التقليدية(the handcraft)=> NP IDENIFICATION_Economy

⇨ Noun field (economy)_Determined+ Adjective_Determined = NP_Economy_Determined

E.g.للصناعة التقليدية(the handcraft)=> PP IDENIFICATION_Economy

⇨ Noun field with pcl (economy)_Determined + Adjective_Determined = PP_Economy_Determined

المحليةالصناعة التقليدية (the local handcraft) =>NP IDENIFICATION_Economy

⇨ NP_Economy_Determined + Adjective_Determined = NP_Economy_Determined

E.g. محليةللصناعة التقليدية (the local handcraft) =>PP IDENIFICATION_Economy

⇨ PP_Economy_Determined + Adjective_Determined = PP_Economy_Determined

- Document 2 : *Simple and complex syntactically classified NPs and PPs with modification by the adjectif(s).*

E.g. التنمية المستدامة(the sustainable development) => NP IDENIFICATION

⇨ Noun_Determined + Adjective_Determined = NP_Determined

E.g.للتنمية المستدامة(the sustainable development) => PP IDENIFICATION

⇨ Noun pcl_Determined + Adjective_Determined = PP_Determined

- Document 3 : *PNs*

E.g.فلاديمير بوتين (Vladimir Poutine), ابراهيم سعيد (Ibrahim Saïd)

⇨ Person_PN/unkown word + Person_PN/unkown word = PERSON NE_PN

- Document 4 : *Definite Description*

E.g. العاهل المغربي(le Moroccan king)

⇨ Trigger word_Person_Politics_Determined + Nationality_Determined = PERSON NE_Politics

- Document 5 : Groups of people (Populations)

E.g. الشيوعيون (the communists); الأفارقة (the africans) ; المجتمع الأمريكي (la société américain)
المجتمع الأمريكي (la société américain) => Triger word_Society + Nationality = PERSON GROUP NE

- Document 6 : *ORGANISATION NE*

E.g. الهيئة العليا المستقلة(The Independent Higher Committee)

⇨ Trigger word_Organisation_Determined + Adjective_Determined = ORGANISATION NE

⇨ ORGANISATION NE + Adjectif_Determined = ORGANISATION NE

- Document 7 : *LOCATION NE* :

- *Listes of the countries/cities/rivers/mountains*

a) المغرب (Morocco), فرنسا (France) => LOCATION NE_Country

b) باريس(Paris), الرباط (Rabat) etc. => LOCATION NE_City

- *Syntactic compositinal entities :*

a) الحدود الجنوبية الشرقية (The south-eastern frontiers) => LOCATION NE

b)الملعب الاولمبي الياباني(The Japanese Olympic Stadium) => LOCATION NE_Sport

c) مدينة تالسينت (the city of Talsint) => LOCATION NE_City

- Document 8 : *EVENT NE*

a) المؤتمر الوطني (the National Congres) => EVENT NE
b) الألعاب الاولمبية(the Olympic Games) => EVENT NE_Sport
c) المحادثات الأمنية العراقية الايرانية (Iraki-iranian Security Negotiations) EVENT NE
d) المؤتمر البيئي الاردني (theJordanien Environnemental Congrès) =>EVENT NE_Environment
e) تسونامي (Tsunami) => EVENT NE

- Document 9 : *NUMBER NE*

a) خمسة (cinq), 145 ,ستة وعشرون (vingt six) => NUMBER NE
b) 25% ; خمسون في المائة (fifty percent) => NUMBER NE_Percentage
c) ثلاثة ملايين دولار (three million dollars) => NUMBER NE_Money

- Document 10 : DATE NE (E.g. 1988 ماي 15 (May 15th 1988)
- Document 11 : HOUR NE(E.g. الساعة الثانية وخمسة دقائق (Five past eight)
- Document 12 : ARTEFACT NE

a) فولكسفاكن(Volkswagen) =>ARTEFACT NE_Car
b) القانون الدستوري (the Constitutional Law) => ARTEFACT NE_Law

- Document 13 : SUBSTANCE NE (E.g. مادة البوتاسيوم7 (the potassium substance))
- Document 14 : LIVING BEING NE (E.g. بكتيريا السالمونيلا (Salmonella))

We note that, in Level 1, we extract and classify two types of entities:

1) Entities with identification (base (trigger word)_Determined + modifier(s))_Determined
2) Pure PN : to extract pure PNs we use lists, that we have constructed using web databases, and trigger words.

To illustrate the results of Level 1 we suggest the following two sentences, in which different colours mark different entity classes and sub-classes recognised in this level by our NE recognition and classification system:

Sentence 1:

التقى الرئيس الروسي فلاديمير بوتين رئيس الاستخبارات العامة السعودية الأمير بندر بن سلطان أمس الثلاثاء في موسكو

Sentence 2:

أكد رئيس المجلس الوطني لتنظيم القطاع الخصوصي وتشجيع المبادرات الشيخ سلمان بن علي بهذا الخصوص على أن تعمل على زيادة مساهمة الطاقة المتجددة في خليط الطاقة الكلية

Note : we obtain this representation, in which different entities are highlighted with different colours, by applying a XSL style sheet to the XML results (the user selects and ticks the classes and sub-classes that he wants to be highlighted). We do this for two reasons : for best visibility, and for concision and clarity (XML document of the two examples above makes 6 pages).

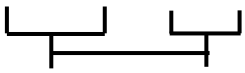
4.2. Level 2

In Level 2, we apply the second step of entity combinations using Level 1 results to extract two types of phases: NP with annexion and PP composed of the prepositional proclitic ل /li/ and NP with annexion.

4.2.1. NP with annexion

A) Syntactically classified :

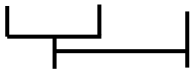
E.g.تشجيع وتدبير المبادرات الجديدة (the support and management of original initiatives)



=> NP ANNEXION

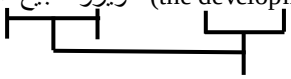
B) Semanticaly classified: E.g.

E.g.دعم الاستثماراتالخارجية (the support of foreign investment)



=> NP ANNEXION_Economy

E.g.تطويروتشجيع المبادرات الاقتصادية(the development and support of economic initiatives)



=> NP ANNEXION Economy

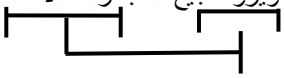
4.2.2. PP composed of the prepositional proclitic ل/li/ and NP with annexion

A) Syntactically classified :

E.g. لتشجيع وتدبير المبادرات الجديدة (of support and management of original initiatives)

B) Semanticaly classified

E.g. لتطويروتشجيع المبادرات الاقتصادية (for the development and support of economic initiatives)



These types of phrases, which can be ORGANISATION/EVENT EN EXTENSIONS, are extracted in Level 2 because they can have in extension NP with identification obtained in Level

1. We can illustrate the results level with the output of our system using the same example above:

أكد رئيسالمجلس الوطنيلتنظيم القطاع الخصوصيوتشجيع المبادراتالشيخ سلمان بن علي بهذا الخصوص على أن تعمل على زيادة مساهمة الطاقة المتجددة في خليط الطاقة الكلية

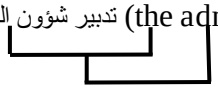
4.3. Level 3

Level 3 uses the results of the precedent levels to extend NP and PP with annexion, which can have NP with annexion in extension obtained in Level 2. For example:

A) Syntactically classified phrases

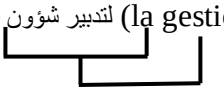
a) *NP with annexion* :

تدبير شؤون الثورة (الشعبية) (the administration of the populaire revolution affaires)



b) *PP composed of the prepositional proclitic ل/li/ and NP with annexion* :

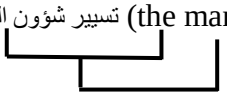
لتدبير شؤون الثورة (الشعبية) (la gestion des affaires de la révolution populaire)



B) Semantically classified

a) *NP with annexion* :

تسيير شؤون السياحة المحلية (the management of local tourisme affaires)



b) *PP composed of the prepositional proclitic ل/li/ and NP with annexion* :

c)

لتسيير شؤون السياحة المحلية (for the management of local tourisme affaires)

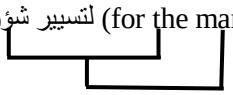


Illustration Level 3 :

أكدرئيسالمجلس الوطنيلتنظيم القطاع الخصوصي وتشجيع المبادراتالشيخ سلمان بن علي بهذا الخصوص على أن 1تعمل على زيادة مساهمة الطاقة المتجددة في خليط الطاقة الكلية.

This example illustrates two types of extension fulfilled in Level 3:

➔ Extension by subordination : زيادة مساهمة الطاقة المتجددة

➔ Extension by coordination : لتنظيم القطاع الخصوصي وتشجيع المبادرات

NP and PP with identification (Level 1) and with annexion (Level 2 and 3) correspondent to attributes after ORGANISATION and EVENT NE trigger word discussed above.

4.4. Level 4

In Level 4, we extract and classify the following entities by combining entities attributes and trigger words:

- *PERSON NE*

PERSON trigger word + PN = PERSON NE_(Field)

a) الرئيس فرانسوا هولاند (The president François Holland) => PERSON NE_Politics

b) العالم الفيزيائي البرت اينشتاين (the physician Albert Einstein) => PERSON NE_Science

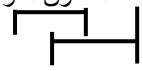
The combined entities/attributes are extracted and classified in Level 1.

- ORGANISATION NE

In this level we combine ORGANISATION trigger word (Level 1) with attributes after this later (i.e. attributes after trigger word) (Level 1, 2, 3):

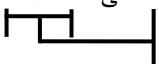
ORGANISATION trigger word +*NP* = ORGANISATION NE_(Field)

وكالة التعاون الدولي(the International coopération Agency) => ORGANISATION NE



ORGANISATION trigger word + *PN* = ORGANISATION NE_(Field)

مستشفى محمد الخامس(Hôpital Mohamed V) => ORGANISATION NE_Healh



ORGANISATION NE + *''Fonction''* attribute = ORGANISATION NE_(Field)

مؤسسة محمد الخامس للتضامن(MohamedVFoundationfor Solidarity)=> EN_ORGANISATION

This rule extends ORGNAISATION NE with pure PNs by adding ''*function''* attribute (generally realized with PP with proclitic ل /li/).

For illustration, we can see the results shown by the processed following sentences:

التقىالرئيس الروسي فلاديمير بوتين رئيس الاستخبارات العامة السعوديةالأمير بندر بن سلطانأمس الثلاثاء في موسكو
قرارمجلس الأمن الدوليرقم 1325بشأنالمرأة والأمنهوواحد من الأدوات الهامةللمساواةبين ...

- EVENT NE

EVENT has the same attributes as ORGANISATION. However, EVENT is distinguished by ''*date''and ''number''*. Thus, EVENT NE is composed of EVENT trigger word followed by EVENT attributes:

EVENT trigger word +NP = EVENT NE_(Field)

مؤتمر التنمية المستدامة(the Sustainable Development Forum) => EVENT NE

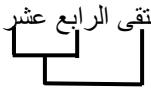
EVENT trigger word + PN = EVENT NE

اعصار **ساندي** (Hurricane Sandy)

Generally, the PN corresponds to the location of event or the PN of the event organiser.

EVENT trigger word + Number = EVENT NE_(Field)

الملتقى الرابع عشر (the fourteenth meeting)



EVENT NE_(Field) + SN_Field = EVENT NE_(Field)

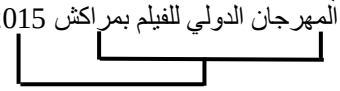
مؤتمر الرياض للتعليم العالي (Riyadh Conference on Higher Eduction)

EVENT NE_(Field) + Location= EVENT NE_(Field)

المهرجان الدولي للفيلم بمراكش(the Marrakech International Film Festival 2015)

EVENT NE_(Field) + Date = EVENT NE_(Field)

المهرجان الدولي للفيلم بمراكش 2015 (the Marrakech International Film Festival 2015)



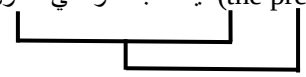
4.5. Level 5

In level 3, we extract the following NE classes by combining ''*Person*'', ''*Location*'' and ''*Law'' attributes* with the corresponding entities:

- PERSON NE

PERSON trigger word (simple/complex) + ORGANISATION NE_(Field) = PERSON NE_(Field)

a)ئيسالمجلسالوطني لحقوق الإنسان (the president of the National Council of Human Rights)



=> PERSON NE_Society

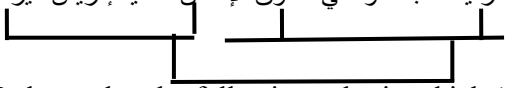
b) المدير العام للشركة العامة(the *Société Générale Managing Director*) => PERSON

=> PERSON NE_Economy

PERSON NE_ (Field) + PERSON_PN= PERSON NE_(Field)

c)السيد إدريس اليزميرئيسالمجلسالوطني لحقوق الإنسان(the president of the National Council of Human Rights Mr Idris El Yazami) => EN_PERSON_POLITICS

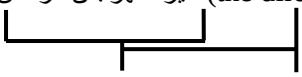
رئيسالمجلسالوطني لحقوق الإنسان السيد إدريس اليزمي



We have also the following rule, in which ‘‘*person*’’ are combined to it corresponding class of EVENT:

PERSON trigger word (simple/complex) + EVENT NE_(Field) = PERSON NE_(Field)

مديرة مهرجان مراكش الدولي للفيلم(the director of Marrakech International Film Festival)



=>EN_PERSON_ART

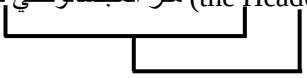
- LOCATION_BUILDING NE

Here we combine ‘‘*location’’* attribute to its corresponding ORGANISATION NE.

LOCATION_BUILDING trigger word + ORGANISATION NE_(Field) = LOCATION NE _(Field)

مقر المجلسالوطني لحقوق الإنسان(the Headquarters of the National Council of Human Rights) => LOCATION NE_Society

مقر المجلسالوطني لحقوق الإنسان (the Headquarters of National Council of Human Rights)



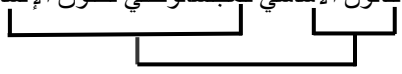
- ARTIFACT_LAW

Concerning LAW NE, we have the following rules :

LAW trigger word (simple/complex) + ORGANISATION NE_(Field) = ARTEFACT _LAW NE _(Field)

القانون الأساسي للمجلسالوطني لحقوق الإنسان (the Rules of Procedure the National Council of Human Rights) => ARTEFACT NE_LAW

القانون الأساسي للمجلسالوطني لحقوق الإنسان



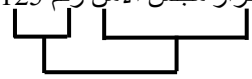
‘‘*Law*’’ attribute of ORGANISATION before trigger word is combined with ORGANISATION NE.

LAW trigger word (simple/complex) + NUMBER NE = ARTEFACT _Law NE

القانون رقم 131 (Law number 131)

ARTEFACT NE_Law + NUMBER NE = ARTEFACT _Law NE_(Field)

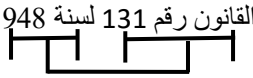
قرار مجلس الأمن رقم 125(the Security Council resolution number 1325)



In the two preceding rules, we combine ‘‘number’’ attribute of Law to LAW entities.

EN_ ARTEFACT_LAW _(Domaine)+ DATE NE = ARTEFACT_Law NE _(Field)

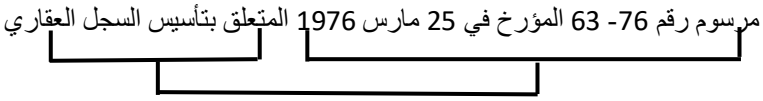
القانون رقم 131 لسنة 1948 (Law number 131 of 1948)



“*Date*” attribute is added to LAW entity.

ARTEFACT_Law NE _(Field) + NP/PP_Context = EN_ ARTEFACT_Law _(Field)

مرسوم رقم 76- 63 المؤرخ في 25 مارس 1976 المتعلق بتأسيس السجل العقاري(Decree No 63-76 dated March 25th 1976 relating to the etablishement of immovable property registry)



NP/PP_Context means an NP or PP marked by a context such as the trigger word المتعلق (relation to) in the example above.

The following example of our system results is an illustration of Level 5 output:

... هو واحد من الأدوات الهامة للمساواة بين قرار مجلس الأمن الدولي رقم 1325 بشأن المرأة والأمن

5. EVALUATION AND APPLICATION

To evaluate our system of Arabic NE recognition and classification we used two corpora:

- *ANERCorp* (available in http://users.dsic.upv.es/grupos/nle/?file=kop4.php) : this corpus is available online ; it contains 154 674 words accompanied with its manually annotated version which facilitate the comparaison with the output of our system. This manual annotation is made of four classes : « PERSON », « ORGANISATION », « LOCATION », and « MISCELLANEOUS ».

Table 1. Results of our Arabic NE recognition system in ANERCorp

class	NE in the corpus	correctly annotated NE	Recall	Precision	F-measure
PERSON	3309	3195	96,55 %	90,68 %	93,52 %
ORGANISATION	1855	1752	94,44 %	92,40 %	93,40
LIEU	4008	3709	92,53 %	96,46%	94,45

- French Press Agency (FPA) *Corpus* : a corpus of about 30 000 words which we made up of FPA articles. The results obtained by the treatment of the corpus are as follows :

Table 2. Results of our Arabic NE recognition system in French Press Agency (FPA) Corpus

Class	NE in the corpus	Correct annotated NE	Recall	Precision	F-Measure
PERSON	1267	1144	96,60 %	93,46 %	95 %
ORGANISATION	881	807	95,45 %	95,95 %	95,69 %
LOCATION	2523	2344	94,96 %	97,82 %	96,36%
EVENT	91	74	91,20 %	89,15 %	90,16 %
TIME (DATE/HOUR)	387	378	98,96 %	98,69 %	98,82 %
NUMBER	753	741	98,67 %	99,73 %	99,19 %

As we can see, the results are very good. In addition, the system is efficient in the process of NE classification and sub-classification in that we note very few errors in the output classification and the results of fine-grained annotations are excellent. These results show the importance of the linguistic approach in NE. The text below shows an example of our system's annotations. The different colours highlights the diversity of the extracted NE sub-classes.

إلتقى الرئيس الروسي فلاديمير بوتينرئيس الاستخبارات العامة السعودية الأمير بندر بن سلطانأمس الثلاثاء في موسكو ، وبحث معه الوضع في سوريا على ضوء التحضير لعقد مؤتمر جنيف2، بينما أكدت إيران أن مشاركتها في المؤتمر الدولي ستكون دون شروط مسبقة . ونقلت وسائل إعلام روسية عن ديمتري بيسكوف المتحدث باسم الرئيس الروسي قوله إن بوتين إلتقى بن سلطان ، " وجرى تبادل الآراء بشكل مفصّل بشأن الوضع في سوريا " على ضوء التحضير لمؤتمر جنيف2 الذي ينعقد يوم 22 يناير / كانون الثاني المقبل . وقال الكرملين في بيان على موقعه الإلكتروني " جرى تبادل الرأي بالتفصيل بشأن الوضع في سوريا بما في ذلك الاستعداد لعقد مؤتمر جنيف2 " ، مضيفا أنه " جرى التطرق إلى الفعاليات الإيجابية في الجهود الدولية لتسوية الملف النووي الإيراني " . وتدعم روسياالرئيس السوري بشار الأسد في صراعه مع المعارضة المسلحة في بلاده والذي أودى بحياة أكثر من مائة ألف شخص منذ منتصف مارس / آذار 2011 ، بينما تدعم السعودية المعارضة المسلحة التي تسعى للإطاحة بالأسد . وكان مصدر دبلوماسي - طلب عدم كشف اسمه - قال في وقت سابق ليوناينتد برس إنترناشونال ، إن بن سلطان موجود في موسكو لإجراء مباحثات مع المسؤولين الروس تتركز حول الأوضاع في سوريا ، بالإضافة إلى احتياجات المملكة من الأسلحة الروسية . " جواد ظريف إذا وجهت الدعوة لنا للمشاركة في اجتماعات جنيف2 فإننا سنشارك في هذا الاجتماع ، وإننا نؤكد أن حل الأزمة السورية لا يمكن أن يتم من خلال الخيار العسكري " مشاركة إيران وتأتي هذه التطورات في الوقت الذي قالت فيه المتحدثة باسم الخارجية الإيرانية مرضية أفخم إن مشاركة بلادها في مؤتمر جنيف2 بشأن الأزمة السورية ستكون دون شروط مسبقة . ونقلت وكالة " مهر " للأنباء الإيرانية عن أفخم قولها في مؤتمر صحفي أمس الثلاثاء إن " الجمهورية الإسلامية

Figure 4. Example of automatically annoted text

The style-sheet colours represent different sub-classes set up in our classification. Generally, the different colours reflect the diversity and the richness of the sub-categorisation in the system. For example:

- PERSON NE_PN / PERSON NE_Politics / PERSON NE_Economy / PERSON NE_Security etc.
- ORGANISATION NE_Politics / ORGANISATION NE_Economy / ORGANISATION NE_Security etc.
- LOCATION NE_Politics / LOCATION NE_Economy/ LOCATION NE_Security etc.
- EVENT NE_Politics / EVENT NE_Economy / EVENT NE_Security etc.

Each class is sub-categorised into the 13 fields and we note no error in our results as far as the sub-categrisation is concerned.

The followind example shows the XML output annotation of the two first NE in the text above:
الرئيس الروسي فلاديمير بوتي (The russan president Vladimir Putin) =>**gNE.Person.Politics**

```
<pos start="0" finish="0" content="الرئيس الروسي فلاديمير بوتين">
  <interpretation>
    <morphology category="C_NOUN" group="gNE.Person.Politics" lemma
        ="رَئِيس" root="رأس" string="الرئيس" form="رئيس" formv
        ="رَّئِيسَ">
      <traitNoun gender="Male" number="Singular" mode="Determined"
          case="Accusative"/>
      <proclitic category="C_PCL_N" string="ال" formv="اَلْ">
        <traitNoun gender="" number="" mode="Determined" case
            ="Accusative"/>
      </proclitic>
    </morphology>
    <morphology category="wNationalityMasculin" group="gNE.Person
        .Politics" lemma="الروسي" root="" string="الروسي" form=""
        formv=""/>
    <morphology category="gNE.Person" group="gNE.Person.Politics"
        lemma="فلاديمير" root="" string="فلاديمير" form="" formv=""/>
    <morphology category="C_NOUN" group="gNE.Person.Politics" lemma
        ="وَتِين" root="وتن" string="بوتين" form="وتين" formv
        ="وَتِينِ">
      <traitNoun gender="Male" number="Singular" mode="Annexion" case
          ="Genetive"/>
      <proclitic category="C_PCL_N" string="ب" formv="بِ">
        <traitNoun gender="" number="" mode="Annexion" case="Genetive"
            />
      </proclitic>
    </morphology>
    <properties category="gNE.Person.Politics" group="gNE.Person
        .Politics" lemma="" root="" string="الرئيس الروسي فلاديمير
        بوتين" form="الرئيس  بوتين" formv="اَلرَّئِيسَ  بِوَتِينِ">
      <traitNoun gender="Male" number="Singular" mode="Annexion" case
          ="Genetive"/>
    </properties>
  </interpretation>
</pos>
```

Figure 5. Annotation output of the NE الرئيس الروسي فلاديمير بوتن

رئيس الاستخبارات العامة السعودية الأمير بندر بن سلطان (the president on the intelligent sevice Bandar Ibn Sultan) =>**gNE.Person.Security**

```
<pos start="0" finish="0" content="رئيس الاستخبارات العامة السعودية الأمير
    بندر بن سلطان">
      <interpretation>
        <morphology category="C_NOUN" group="gNE.Person.Administration"
            lemma="رَئِيس" root="رأس" string="رئيس" form="رئيس" formv
            ="رَئِيسٌ">
          <traitNoun gender="Male" number="Singular" mode="Indetermined"
              case="Nominative"/>
        </morphology>
        <morphology category="C_PN" group="gNE.Organisation.Security"
            lemma="الاستخبارات" root="" string="الاستخبارات" form="" formv
            =""/>
        <morphology category="C_NOUN" group="gNE.Organisation.Security"
            lemma="عَامّ" root="عمم" string="العامة" form="عامة" formv
            ="عَامَّةٌ">
          <traitNoun gender="Female" number="Singular" mode="Determined"
              case="Genetive"/>
          <proclitic category="C_PCL_N" string="ال" formv="اَلْ">
            <traitNoun gender="" number="" mode="Determined" case
                ="Genetive"/>
          </proclitic>
        </morphology>
        <morphology category="C_PN" group="gNE.Organisation.Security"
            lemma="السعودية" root="" string="السعودية" form="" formv=""/>
        <morphology category="C_NOUN" group="gNE.Person.PN" lemma="أَمِير"
            root="أمر" string="الأمير" form="أمير" formv="أَمِيرَ">
          <traitNoun gender="Male" number="Singular" mode="Determined"
              case="Accusative"/>
          <proclitic category="C_PCL_N" string="ال" formv="اَلْ">
            <traitNoun gender="" number="" mode="Determined" case
                ="Accusative"/>
          </proclitic>
        </morphology>
```

Figure 6. Annotation output of the NE رئيس الاستخبارات العامة السعودية الأمير بندر بن سلطان

The output in this figuer shows that the NE رئيس الاستخبارات العامة السعودية الأمير بندر بن سلطان is a person NE (gNE.Person.Administration) into which a secutity organization NE (الاستخبارات العامة السعودية =>gNE.Orgnasation.Security) is embedded.

We exploits the output of our system of NE extraction and classification in both information retrieval and extraction:

- *Information retrieval* : the search engine uses the output of our system of NE extraction and classification to contextualise the query. The follwing figures show some aspects of the recognised NE contribution in information retrieval :

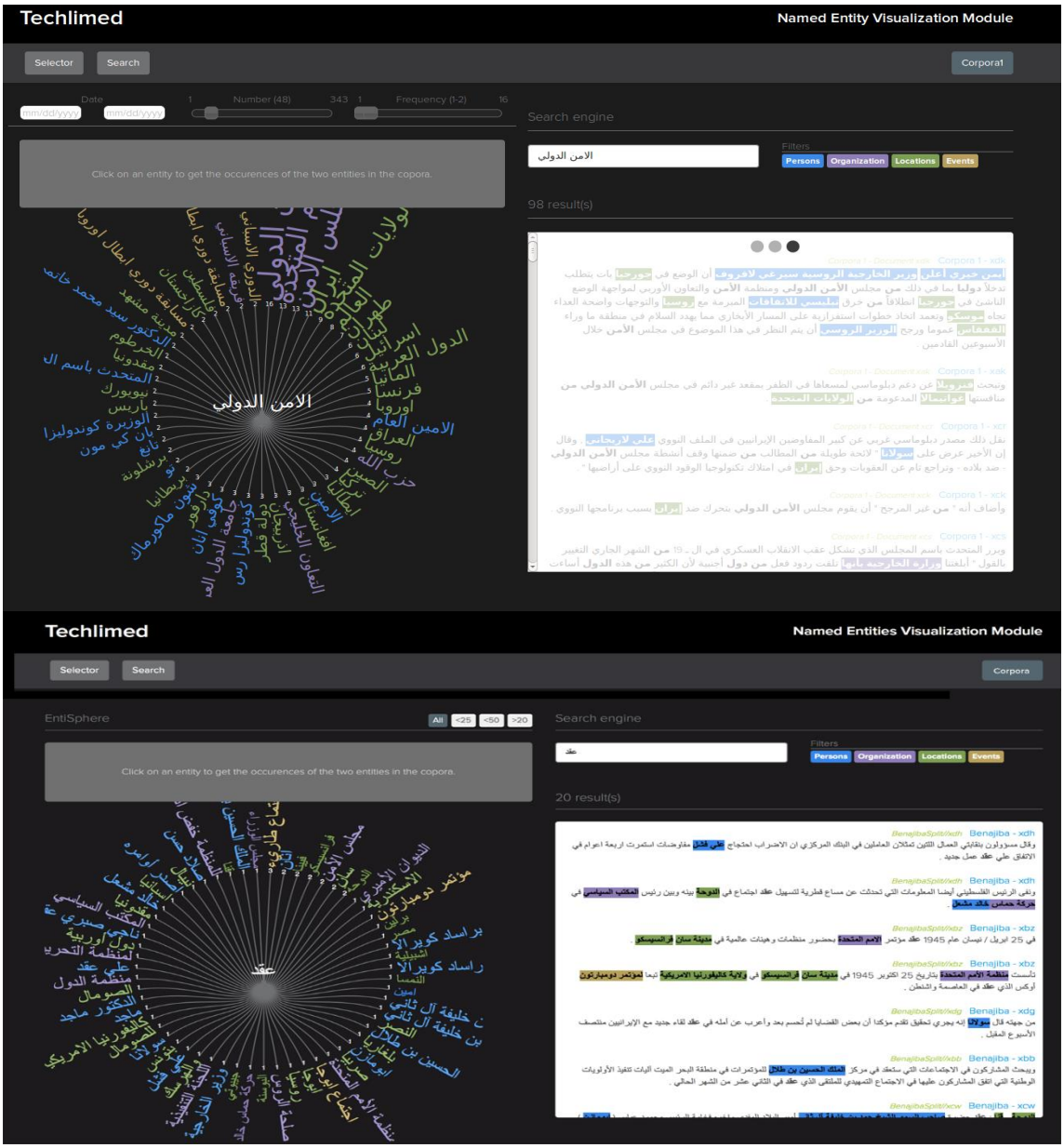


Figure 7. Research engine contextualisation of the query الامن الدولي (the international security) with NEs extracted by our system

The displayed results associate NEs with the query (the diagram on the left of the screen), and highlight NEs in the text in the right side. In addition, the search engine uses the NE classification to filter by classes and subclasses. We can see that we can filter by class (ex : PERSON, ORGANISATION, LOCATION, EVENT) and by field (Politics, science, religion etc.) (figure 9).

Figure 8. Search engine filtering by Person class and sub-classes

- *Information extraction* : Techlimed has also developed an information extraction application using the output of our NE extraction and classification system. The figures bellow show the contribution of our NE recognition system in information extraction :

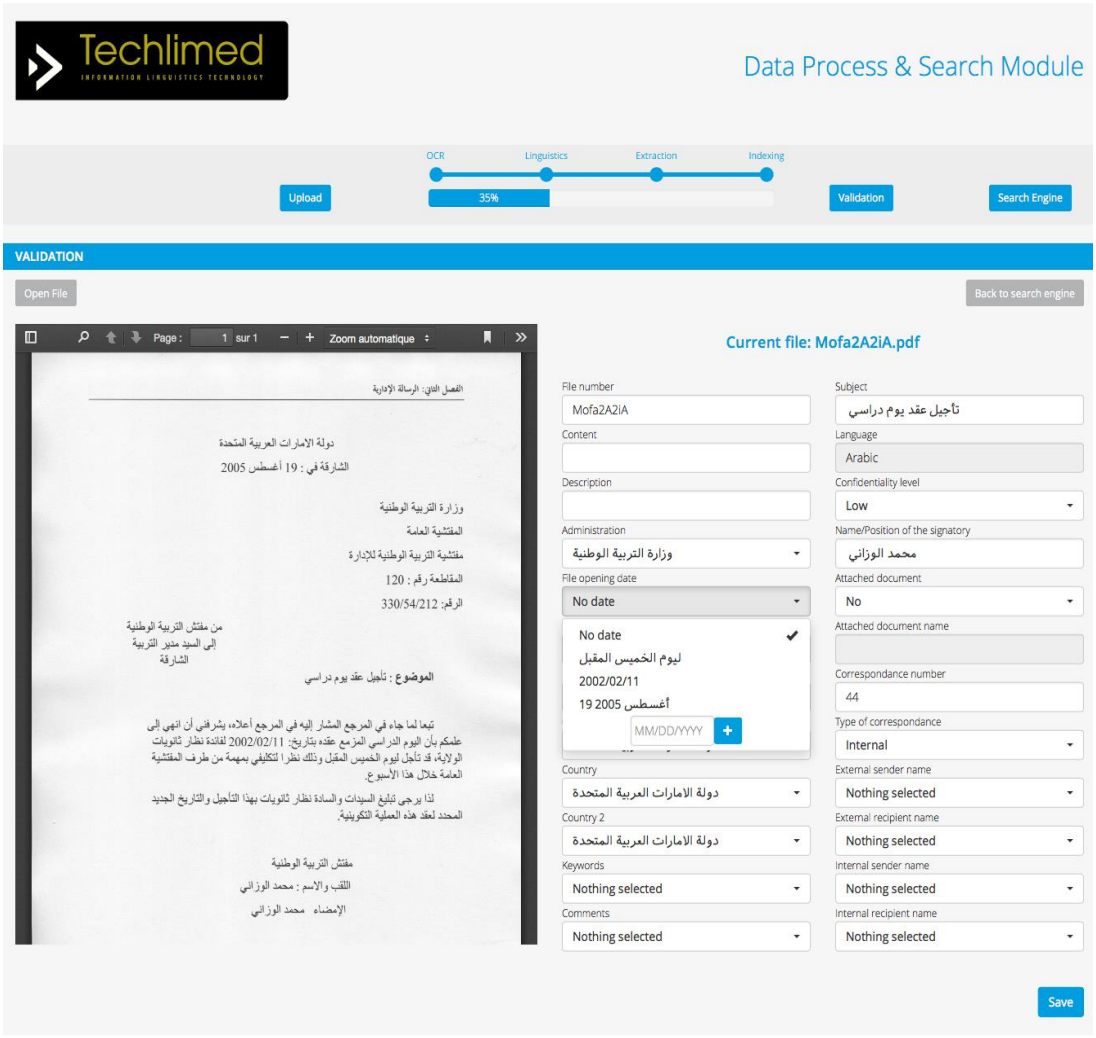


Figure 9. Information extraction based on our NE extraction and classification

6. CONCLUSION AND PERSPECTIVES

Our system highlights the importance and the necessity of the linguistic approach in NE recognition and classification. The morpho-syntactic analysis is based on DIINAR with is a rich Arabic lexical database. The syntactic rule construction is based on the binary combinations inspired from X-bar theory and from the notion ‘‘immediate constituents’’. This necessitates different levels of rule construction (Level 1, 2, 3, 4, 5). The semantic classifications of NE constituents lie on the semantic relations of synonymy, hyperonymy and semantic fields. And the syntactico-semantic rules take into account the notion of NE class attributes. These linguistic information (morpho-syntactic, syntactic, semantic, and syntactico-semantic) are essential in our method of Arabic NE extraction and classification. The output of our system were exploited by Techlimed in both applications information retrieval and extraction. The obtained annotations are integrated in the systems of indexations for an efficient information retrieval and extraction. The output part of speech will also allow us to develop the project of establishing relations between the different extracted NE using verb classification. That is to say, argument/predicate relationship can then be extracted and classified. We have already stated this task. The following sentence extaction and annotation by g.Relation.Contact (the first sentence in the text above) is an example:

```
<pos start="0" finish="0" content="التقى الرئيس الروسي فلاديمير بوتين
      رئيس الاستخبارات العامة السعودية الأمير بندر بن سلطان أمس الثلاثاء
      في موسكو">
  <properties category="g.Relation.Contact" group="g.Relation
       .Contact" lemma="" root="" string="التقى الرئيس الروسي
       فلاديمير بوتين رئيس الاستخبارات العامة السعودية الأمير بندر بن
       سلطان أمس الثلاثاء في موسكو " form="التقى الرئيس  بوتين رئيس
       العامة  الأمير بندر بن سلطان أمس  في " formv="إلْتَقَى
       ٱلرَّئِيسَ  بِوتِين رَئِيسُ  ٱلْعَامَّةِ  ٱلْأَمِيرَ
       بَنْدَر بْنُ سَلْطَانِ أَمْسَ  فِي "/>
```

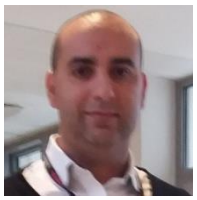
In addition, the obtained annotations can be exploited in the ontology enrichment.

REFERENCES

1. Asharef, M., Omar, N., Albared, M. (2012), "Arabic NE recognition in crime documents", In *Journal of Theoretical and Applied Information Technology*. Vol. 44. N°. 1, pp 1-6.
2. Asbayou, O., (2016), *L'identification des entités nommées en arabe en vue de leur extraction et classification automatiques : La consruction d'un système à base de règles syntactico-sémantiques*, Thèse de doctorat, Université Lumière Lyon 2.
3. Asbayou, O., (2017), ''La détermination desattributs sémantiques internes à la structure syntaxico-sémantique des entités nommées en vue de leur extraction automatique'', *In 20ème Colloque International sur le Document Electronique, ENSSIB* pp227-240.
4. Attia, M., Toral, A., Tounsi, L., Monachini M. et Van Genabith, J. (2010), "An automatically built NE lexicon for Arabic", In *LREC 2010, 7th conference on International Language Resources and Evaluation*, Valletta, Malta.
5. Benajiba, Y., Rosso, P. (2008), "Arabic NE recognition using conditional random fields", In *Proceedings of Workshop on HLT and NLP within the Arabic world, LREC'08*.
6. Brun, C., Ehrmann, M., Jacquet, G. (2007), "A hybrid system for NE metonymy resolution", In *4th International Workshop on Semantic Evaluations, ACL-SemEval*, Prague.
7. Charton, E., Torres-moreno, J. (2009), ''Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées'', In *TALN 2009, Senlis,*Vol. n° 1, pp 24–26.
8. Charton, E. Gagnon, M. et Ozell, B. (2010), ''Extension d'un système d'étiquetage d'entités nommées en étiqueteur sémantique'', In *proceedings TALN*, Montréal.
9. Daille B., Fourour N., et Morin E. (2000), ''Catégorisation des noms propres : une étude en corpus'', In *Cahiers de Grammaire*, Vol 25, pp 115-129.
10. Dichy, J., Braham, A., Ghazali, S., Hassoun, M. (2002), ''La base de connaissances linguistiques DIINAR.1 (DIctionnaire INformatisé de l'Arabe, version1)'', In *Proceedings of the International Symposium on The Processing of Arabic*, Tunis, Université de Manouba.
11. Ehrmann, M. (2008), *Les entités nommées, de la linguistique au TAL: statut théorique et méthodes de désambiguïsation*, Thèse de doctorat. Université Paris 7.
12. Ehrmann, M., Jacquet, G. (2006),''Vers une double annotation des entités nommées'', In *Traitement Automatique du Langues*, Vol. 47, pp 63-88.
13. El Maarouf, I., Villaneau, J., Rosset, S. (2011), ''Extraction de patrons sémantiques appliquée à la classification d'entités nommées'', In *TALN*. Montpellier.
14. Farber, B., Freitag, D., Habash, N., et Rambow, O. (2008), "Improving NER in Arabic Using a Morphological Tagger", In *Proceedings of LREC*, Marrakech.
15. Friburger, N. (2002), *Reconnaissance automatique des noms propres: application à la classification automatique de textes journalistiques*, Thèse de doctorat, Université de Tours.
16. Gary-Prieur, M. N. (1994), *Grammaire du nom propre*, Paris, Presse universitaire de France.
17. Gary-Prieur, M. N. (2009), ''Le nom propre, entre langue et discours'', In *Les Carnets du Cediscor, 11, Publication du Centre de recherches sur la didacticité des discours ordinaires*, pp 153-168.
18. Hagène, C., Roux, C. (2003), ''Entre syntaxe et sémantique : normalisation de la sortie de l'analyse en vue de l'amélioration de l'extraction d'information à partir de texte'', In *Actes de TALN 2003*. Batz-sur-Mer, pp 11-14.
19. Harris, Z. (1951), *Methods in structural singuistics*, Chicago, University of Chicago Press.

20. Harris, Z. (1968), *Mathematical structures of language*, New York, Wiley and Sons.
21. Kosseim, L., Poibeau, T. (2001), ‘‘Extraction de noms propres à partir de textes variés : problématique et enjeux’’, In *8ème conférence nationale sur le Traitement Automatique des Langues Naturelles*. Tours.
22. Le Meur, C., Galliano, S., Geoffrois, E. (2004), ‘‘Conventions d'annotations en entités nommées-ESTER’’, In *Rapport technique de la campagne Ester*, Paris.
23. Mesfar, S. (2008), *Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard*, Thèse de doctorat, Université de Franche-Comté, Besançon.
24. Poibeau, T. (2001), ‘‘Deconstructing Harry, une évaluation des systèmes de repérage d’entités nommées’’, In *Revue de la société d’électronique, d’électricité et de traitement de l’information*, Paris.
25. Poibeau, T. (2005), ‘‘Sur le statut référentiel des entités nommées’’, In *Actes de la conférence Traitement Automatique des Langues Naturelles*, Dourdan, France.
26. Shaalan, K. (2014), ‘‘A Survey of arabic NE recognition and classification’’, In *Computational Linguistics*, Vol. 40, issue 2: MIT Press, Cambridge, Massachusetts, pp 469 - 510.
27. Traboulsi, H. (2009), “Arabic NE extraction: a local grammar-based approach”, In *Proceedings of the International Multiconference Computer Science and Information Technology*, pp 139-143.
28. Zaghouani, W. (2009),*Le repérage automatique des entités nommées dans la langue arabe: vers la création d'un système à base de règles*, Mémoire de Maîtrise, Université de Montréal.

AUTHOR



I am **Omar ASBAYOU**, a teacher in Lumière University Lyon 2, and a member in CRTT laboratory. My reseach focuses on Arabic language processing I was an ingeneer resercher in Techlimed, a company specialised in The automatic processing of Arabic.