ARABIC LOCATION NAME ANNOTATIONS AND APPLICATIONS

Omar ASBAYOU

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

ABSTRACT

This paper show how location named entity (LNE) extraction and annotation, which makes part of our named entity recognition (NER) systems, is an important task in managing the great amount of data. In this paper, we try to explain our linguistic approach in our rule-based LNE recognition and classification system based on syntactico-semantic patterns. To reach good results, we have taken into account morpho-syntactic information provided by morpho-syntactic analysis based on DIINAR database, and syntactico-semantic classification of both location name trigger words (TW) and extensions. Formally, different trigger word sense implies different syntactic entity structures. We also show the semantic data that our LNE recognition and classification system can provide to both information extraction (IE) and information retrieval(IR).The XML database output of the LNE system constituted an important resource for IE and IR. Future project will improve this processing output in order to exploit it in computer-assisted Translation (CAT).

KEYWORDS

Location name annotations, Location named entities, Information retrieval, Information extraction

1. INTRODUCTION

Geographic document pose a problem to IR. For this reason, it is very important to particularly exploit lexical entitiesreferring to location instances (location named entities/proper names). The provided data and classification are important in IR especially in location proper name contextualisation in text. Text geo-parsing, which means identifying place names in corpus, is one of the experimental approaches in Digital Humanities, particularly in Geospatial Humanities.I. Gregory (2015) emphasises the importance of spatial humanities and the necessity of the creation of corresponding databases for geographic information.J.L. Leidner, (2007) worked on LNE for geographic information system (GIS). Text geo-parsing is part of location named entity recognition (LNER), which is a subfield in NLP. A great deal of research has been done on named entity recognition (NER) in general. However, this task, as we will particularly show in this paper, is a sub-task that can actually play an important role in many natural language processing (NLP) applications, especially in IR. This is actually realised by providing annotated index, which is an XML database (LNER system output) enriched with semantic annotations, which can provide a NE class filter to research engine. It also plays a vital role in ontology enrichment for semantic web. LNE interrelations and their relation with other NE of different classes (PERSON, ORGANISATION, EVENT, TIME etc.) represent an interesting resource for IR and IE etc. Generally, NE are lexical entities with very important informative value. M. Asharef*et.al* [2012], for example, used NE extraction from crime text. The recognition and classification of these entities can provide valuable economic, social and political information

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 51-65, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101405

associated to locations. This is due to what M. Herrmann [2008] calls ‘‘referential unicity’’ to define and characterise NE and particularly proper names. « LOCATION », beside « PERSON », « ORGANISATION », « EVENT » etc. is one of the most important NE classes. Fine annotations of LNE are central in many tasks such as improving geographic text and general corpus analysis. They offer the possibility to associate different entity classes to different location sub-classes.

2. METHODOLOGY: LINGUISTIC DESCRIPTION

LNE belongs to different lexical and lexico-syntactic categories:

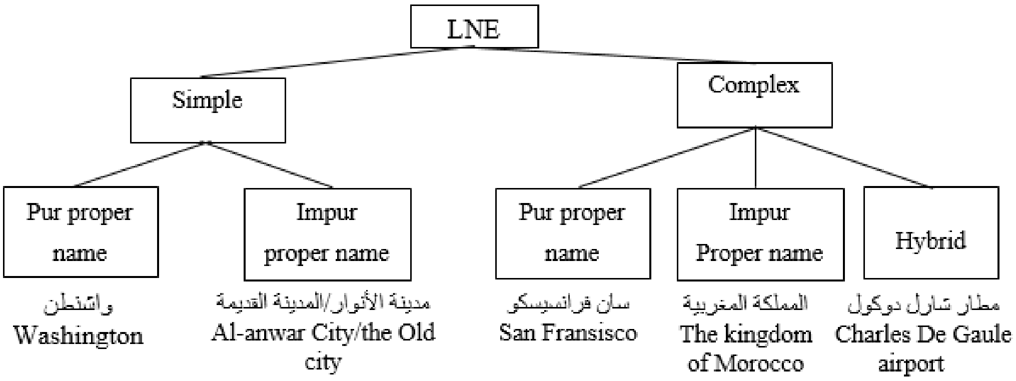


Figure 1.LNE lexico-syntactic categories

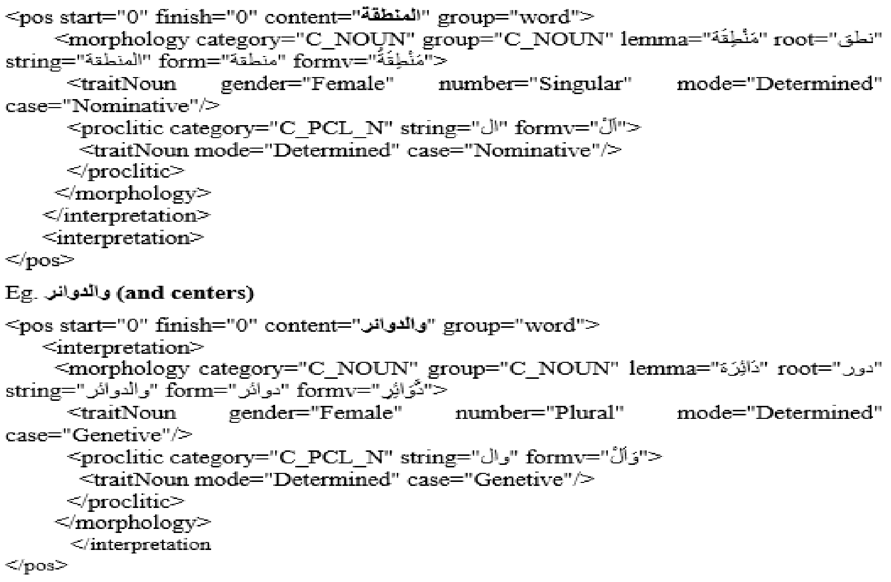
Our rule-based system takes into account these structure properties in their recognition and classification by fine annotations. Therefore, we used a set of lexical, syntactic, and semantic classifications to build correct syntactico-semantic rules of our system.

2.1. Lexical Information

2.1.1.Morpho-syntactic analysis

Our LNER system makes part of our rule-based NER system, which is processing of six levels. We are not going to expose a detailed description of our system here. Level 0 applies the morpho-syntactic analysis which is the first and basic step in our system. We use a morpho-syntactic analysis system based on DIINAR, a rich Arabic lexical database constructed by many researchers: J. Dichy, from Lumière university Lyon 2, and A. Braham from Manouba University Tunisia, (linguistic aspect), M. Hassoun, ENSSIB Lyon, Research Institute for Computer Science and Telecommunication in Tunis, and S. Ghazali from High Institute of Language in Tunis(computer science aspect).This database provides our morpho-synyactic analysis system with rich morpho-syntactic data. The morpho-syntactic-analysis represents the first pre-treatment step for our LNE extraction system (Level 0). For example, the words المنطقة (the region) and والدوائر (and centres) aremorpho-syntactically analysed as follows:

E.g.1:**المنطقة (the region)**



```
<pos start="0" finish="0" content="المنطقة" group="word">
    <morphology category="C_NOUN" group="C_NOUN" lemma="مَنْطِقَة" root="نطق"
string="المنطقة" form="منطقة" formv="مَنْطِقَةُ">
     <traitNoun    gender="Female"    number="Singular"    mode="Determined"
case="Nominative"/>
     <proclitic category="C_PCL_N" string="ال" formv="أَلْ">
      <traitNoun mode="Determined" case="Nominative"/>
     </proclitic>
    </morphology>
   </interpretation>
   <interpretation>
</pos>
```

Eg. **والدوائر (and centers)**

```
<pos start="0" finish="0" content="والدوائر" group="word">
   <interpretation>
    <morphology category="C_NOUN" group="C_NOUN" lemma="دَائِرَة" root="دور"
string="والدوائر" form="دوائر" formv="دَوَائِرِ">
     <traitNoun    gender="Female"    number="Plural"    mode="Determined"
case="Genetive"/>
     <proclitic category="C_PCL_N" string="وال" formv="وَأَلْ">
      <traitNoun mode="Determined" case="Genetive"/>
     </proclitic>
    </morphology>
     </interpretation
</pos>
```

Figure 2. Examples of morpho-syntactic analysis

This Lexical resource constitute the base for our syntactic rules in that it provides lexical information that are basic data in our NE recognition and classification system.

2.1.2 Semantic classification

Our sub-classification of the class « «LOCATION» is based on many approaches: field, organisation location, organisation building, geographic location, address and facility.

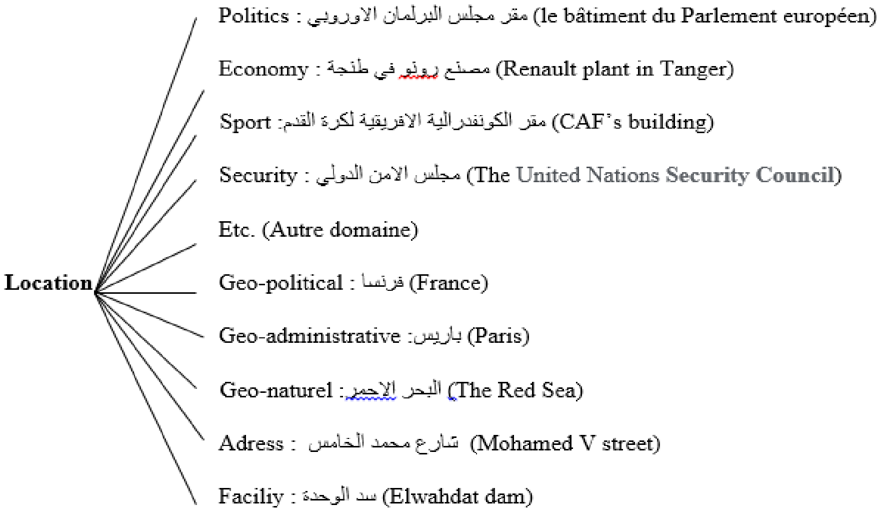
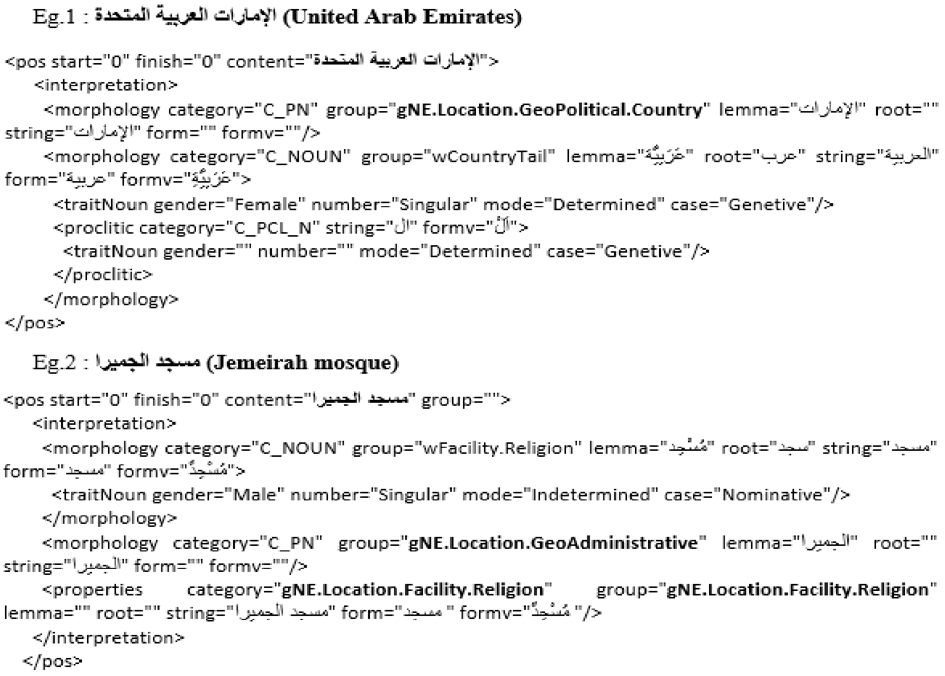

Location
Politics : مقر مجلس البرلمان الاوروبي (le bâtiment du Parlement européen)
Economy : مصنع رونو في طنجة (Renault plant in Tanger)
Sport :مقر الكونفدرالية الافريقية لكرة القدم (CAF's building)
Security : مجلس الامن الدولي (The United Nations Security Council)
Etc. (Autre domaine)
Geo-political : فرنسا (France)
Geo-administrative :باريس (Paris)
Geo-naturel :البحر الاحمر (The Red Sea)
Adress : شارع محمد الخامس (Mohamed V street)
Faciliy : سد الوحدة (Elwahdat dam)

Figure 3. Location sub-categories

The figure above show that LOCATION class can be defined by:

- Field (politics, security, sport, economy .etc.): it functions as a distinctive feature based on a well-defined semantic field classification.
- Geography: this allows us to distinguish geographic proper names form other types of location. This sub-class is subdivided on many subclasses : geopolitical, which includes « country », « city », « region », « department » etc. and geo-natural, in which we put « sea », « mountain », « river » etc.
- Facility: this information is associated to different facility proper names (dam, motorway, stadium etc.).

This information is provided by the recognised LNE constituent information and expressed by fine semantic annotations. The figure bellow show how LNE are annotated by our system and the classifying information provided by these semantic annotations (for visibility, I put the entity and the annotation in bold):



Eg.1 : الإمارات العربية المتحدة **(United Arab Emirates)**

```
<pos start="0" finish="0" content="الإمارات العربية المتحدة">
    <interpretation>
     <morphology category="C_PN" group="gNE.Location.GeoPolitical.Country" lemma="الإمارات" root=""
string="الإمارات" form="" formv=""/>
     <morphology category="C_NOUN" group="wCountryTail" lemma="عَرَبِيَّة" root="عرب" string="العربية"
form="عربية" formv="عَرَبِيَّةٌ">
       <traitNoun gender="Female" number="Singular" mode="Determined" case="Genetive"/>
       <proclitic category="C_PCL_N" string="ال" formv="أَلْ">
         <traitNoun gender="" number="" mode="Determined" case="Genetive"/>
       </proclitic>
     </morphology>
</pos>
```

Eg.2 : مسجد الجميرا **(Jemeirah mosque)**

```
<pos start="0" finish="0" content="مسجد الجميرا" group="">
    <interpretation>
     <morphology category="C_NOUN" group="wFacility.Religion" lemma="مَسْجِد" root="سجد" string="مسجد"
form="مسجد" formv="مَسْجِدٌ">
       <traitNoun gender="Male" number="Singular" mode="Indetermined" case="Nominative"/>
     </morphology>
     <morphology  category="C_PN"  group="gNE.Location.GeoAdministrative"  lemma="الجميرا"  root=""
string="الجميرا" form="" formv=""/>
     <properties      category="gNE.Location.Facility.Religion"      group="gNE.Location.Facility.Religion"
lemma="" root="" string="مسجد الجميرا" form="مسجد " formv="مَسْجِدٌ "/>
    </interpretation>
  </pos>
```

Figure 4. Examples of LNE annotation output

We note that the second annotation example contains two LNE categories: gNE.Location.GeoAdministrative and gNE.Location.Facility.Religion. Based on entity constituents translated by our lexical semantic classification (e.g. wFacility.Religion), each of these annotations provide some location information (location, geo-administrative, facility, religion). To morpho-syntactic data, we add and semantic information .We classified the linguistic entities involved in LNE structure, in our syntactico-semantic rule system, according to different semantic (semantic fields) and conceptual relations (TW classes). We put this TW lexicon into different categories according to common semantic, conceptual and lexico-syntactic criteria.

A. TW (Trigger words) :

TW are word marking NE initial position. For example, مجلس (council), جمعية (association), هيئة (committee) etc. are *generic TW* belonging to the class LOCATION/ORGANISATION. These are distinguished form *specific TW* like قنصلية (consulate) ,مطار (airport) for syntactico-semantic reasons. These semantically different TW classes belong to different LOCATION sub-classes since they participate in different syntactico-semantic patterns (rules). In Level 1, we added this lexico-semantic information to morpho-syntactic output of level 0.

B. Proper names :

We have also enrichedLevel 1 with a set of:

- ***Lists of sub-category locations***: these lists of simple LNE of countries, cities, rivers, mountains etc. aims to enrich le lexical database. For example:

a.المغرب(Morocco), فرنسا (France) etc. =>gNE.LOCATION.GeoPolitical.Country
b. باريس (Paris), الرباط(Rabat) etc. =>gNE.LOCATION.GeoAdministrative.City

- ***Syntactic entities:*** syntactic entities are extracted and annotated using our syntactico-semantic rules combining constituents.

a. الحدود الجنوبية الشرقية (The *southeastern frontiers)* =>gNE.LOCATION.GeoPolitical
*b.*الملعب الاولمبي الياباني (The Japanese Olympic Stadium) =>gNE.LOCATION.Facility.Sport
c. مدينة أكادير (the city of Agadir) =>gNE.LOCATION.GeoAdministrative.City

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

Sentence 1:
التقى الرئيس الروسيفلاديمير بوتين رئيسالاستخبارات العامة السعوديةالأميربندر بن سلطانأمس الثلاثاء في موسكو
Sentence 2:
أكدرئيسالمجلس الوطنيلتنظيمالقطاع الخصوصي وتشجيع المبادرات الشيخ سلمان بن علي بهذا الخصوص على أن تعمل على زيادةمساهمةالطاقة المتجددة في خليطالطاقة الكلية

In these two examples, the NER system extracted and classified the LNE:موسكو (Moscow) (gNE.LOCATION.GeoAdministrative.Country), in sentence 1, and the complex TW المجلس الوطني (the national council) in of the NE المجلس الوطنيلتنظيمالقطاع الخصوصي وتشجيع المبادرات. LOCATION/ORGANISATION ambiguity is resolved in the following levels exploiting contextual elements such as the prepositionsفي (in) and ORGANISATION attributes such asرئيس (president) in *رئيسالمجلسالوطني* لحقوق الإنسان (the president of the National Council of Human Rights) which disambiguates the NE class into gNE_PERSON_Society. Here, we combine ‘‘*organisation*’’ with ‘‘*person function*’’ attribute.

Our study deals with NE linguistic specificities and the elements involved in their syntactico-semantic structures. These linguistic data highlight several levels of analysis.

2.2. Syntactico-Semantic Information

Our study starts from the principles that TW represent the head of the noun phrase NE (NP)and it is accompanied with one or several modifiers or complements. Second, Each LNE class has defined set of attributes (generally extracted in Level 1 and 2), which are put in a well-determined distribution in NE extraction patterns. Third, Complex LNE structures are composed of many linguistic entities combined in a defined order (immediate constituent analysis). Fourth, fine annotation depends on trigger word and extension information. Therefore, in the syntactico-semantic level, we will shed light on two important aspects: NE structure constituents and NE class attributes.

2.2.1. LNE Structure Constituents

LNE constituents are crucial in LNE constituent combinations and classification. The structure of NE is divided into two parts: trigger words and extensions.

A. Trigger word :

We have taken into account different perspectives in constructing the typology of trigger words:

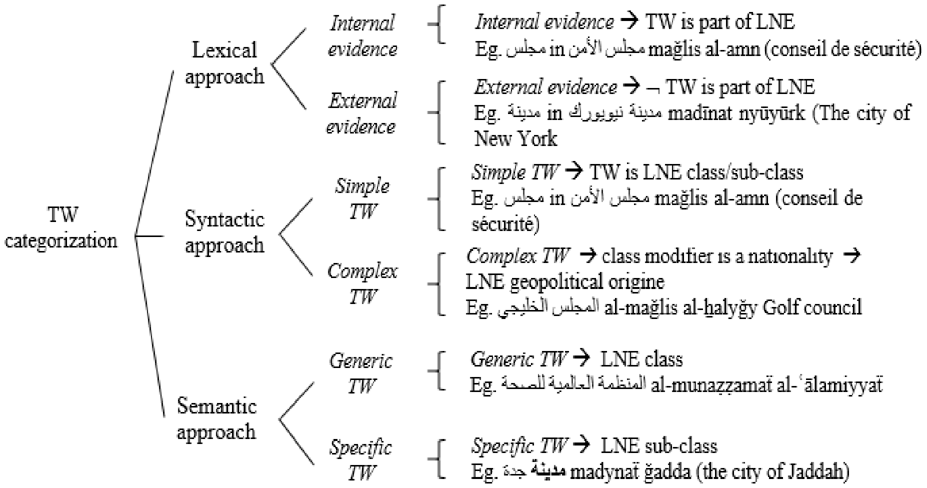

TW categorization
Lexical approach
Internal evidence
Internal evidence → TW is part of LNE
Eg. مجلس in مجلس الأمن mağlis al-amn (conseil de sécurité)
External evidence
External evidence → ¬ TW is part of LNE
Eg. مدينة in مدينة نيويورك madīnat nyūyūrk (The city of New York
Syntactic approach
Simple TW
Simple TW → TW is LNE class/sub-class
Eg. مجلس in مجلس الأمن mağlis al-amn (conseil de sécurité)
Complex TW
Complex TW → class modifier is a nationality → LNE geopolitical origine
Eg. المجلس الخليجي al-mağlis al-ẖalyğy Golf council
Semantic approach
Generic TW
Generic TW → LNE class
Eg. المنظمة العالمية للصحة al-munaẓẓamaẗ al-ʿālamiyyaẗ
Specific TW
Specific TW → LNE sub-class
Eg. مدينة جدة madynaẗ ğadda (the city of Jaddah)

Figure 5.Trigger word typology

As the figure shows, TW sub-categorization provides interesting information for LNE recognition and classification.

B. 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. Information provided by the extension is very useful in solving the problems of NE frontiers and classification.

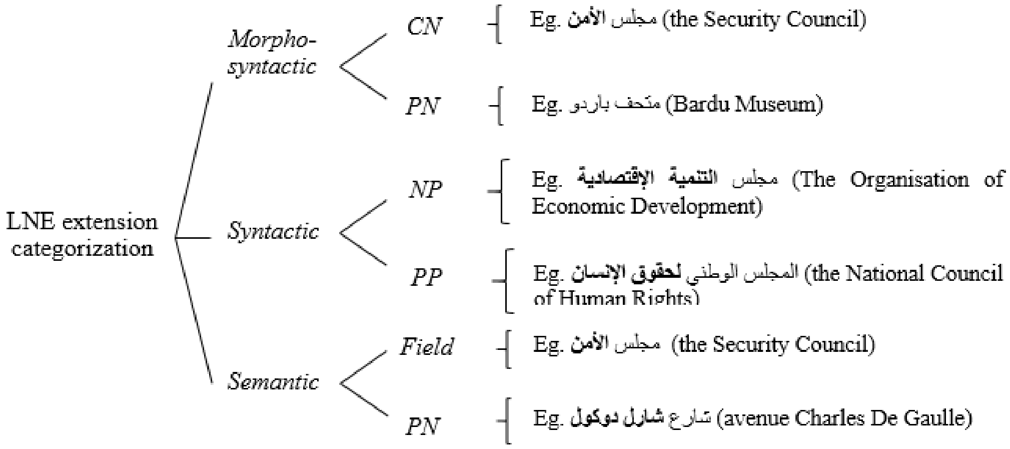

LNE extension categorization
Morpho-syntactic
CN
Eg. مجلس الأمن (the Security Council)
PN
Eg. متحف باردو (Bardu Museum)
Syntactic
NP
Eg. مجلس التنمية الإقتصادية (The Organisation of Economic Development)
PP
Eg. المجلس الوطني لحقوق الإنسان (the National Council of Human Rights)
Semantic
Field
Eg. مجلس الأمن (the Security Council)
PN
Eg. شارع شارل دوكول (avenue Charles De Gaulle)

Figure 6. LNE extension categorisation

2.2.2. LNE class semantic attributes and relation with NE

LNE semantic attributes are classes that participate in the formation of their syntactico-semantic patterns. They have values, which are recognised and represented by different types of linguistic entities classified by our NER system. For example, the LNE attribute « person proper name» in **شارع شارل دوكول**šāriʿšārldūgul (Charles De Gaulle) has the value «**شارل دوكول** » šārldūgul (Charles De Gaulle). LNE attribute position is « after TW » and LNE are, in their turn, are attributes in other named entities.

A. Attributes after TW

In this case, class attribute values do not change LNE class. The formal description is a semantic feature structure: LNE « x attribute » has « y value ». For example, the LNE شارع **شارل دوكول**šāriʿ šārl dūgul (Avenue Charles De Gaulle) « person proper name attribute» has « **دوكولشارل šārl** dūgul (Charles De Gaulle) value ».

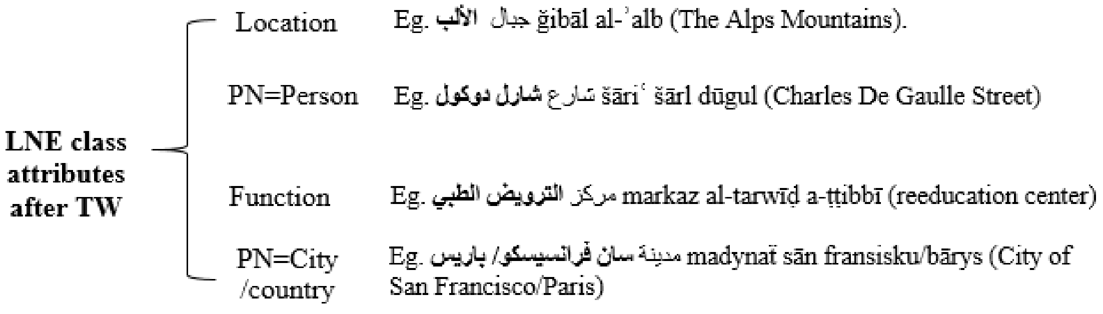

LNE class attributes after TW
Location
Eg. جبال الألب ğibāl al-ʾalb (The Alps Mountains).
PN=Person
Eg. شارع شارل دوكول šāriʿ šārl dūgul (Charles De Gaulle Street)
Function
Eg. مركز الترويض الطبي markaz al-tarwīḍ a-ṭṭibbī (reeducation center)
PN=City /country
Eg. مدينة سان فرانسيسكو/ باريس madynaī sān fransisku/bārys (City of San Francisco/Paris)

Figure 7. LNE class attributes after TW

B. LNE=NE attributes

The LNE do not have attributes before the TW but can be attributes of other NE classes.Nonetheless, they should not be lost within other NE in which they are constituents providing some information such as event/location relation (a and b) and event/organisation relation (c):

- EVENT NE attribute ATW = GEOPOLITICAL LOCATION NE

a) **جنيف** مؤتمر (Geneva Meeting) is EVENT NE with «location attribute» whose value is جنيفǧunif (Geneva).

b) **الصين** الدورة التاسعة والعشرين للألعاب الأوليمبية في (The twenty-ninth Olympic Games 2008 in China) is EVENT NE with «location attribute» whose value is **الصين**(China).

- ORGANISATION NE attribute BTW = LOCATION BUILDING NE

c) **مبنى** مجلس الأمن الدولي (The international Security Council Building) is ORGANISATION NE with «location building attribute» whose value is the whole NE **مبنى** مجلس الأمن الدولي (The international Security Council Building) because the attribute is before the ORGANISATION NE TW.This contributes in solving the problem of some cases of metonymy typical of ORGANISATION NE.

C. The role of « nationality » modifier and of « geopolitical location » complement

« Nationality » and « geopolitical location » are geo political (location) attributes in NE whose value is respectively an adjective (modifier) or geopolitical proper name (complement); both entities denote a geo political information of the NE in which they are constituent after TW.

a. رئيس الحكومة **المغربية** سعد الدين العثماني (the **Moroccan** government president saʿd a-ddīn al-ʿuṯmānī) =>gNE_PERSON_Politics

b. الرئيس **الامريكي** دونالد ترامب(The **American** President Donald Trump) =>gNE_PERSON_Politics

c. البنك المركزي **الأوروبي**(the **European** Central Bank)

d. *مهرجان **مراكش** الدولي للفيلم 2019* (Marrakech International Film Festival 2019)

From annotations provided by our NER system, geopolitical origin is one of the information that can be extracted:

Table 1. NE semantic constituents

NE	Annotation	Class	Field	Geopolitical origine	Proper name	Date
الرئيس الامريكي دونالد ترامب	gNE_PERSON_Politics	PERSON	Politics	American (America)	دونالد ترامب	…
رئيس الحكومة المغربية سعد الدين العثماني	gNE_PERSON_Politics	PERSON	Politics	المغربية (Moroccan/ morocco)	سعد الدين العثماني	…
البنك المركزي الأوروبي	gNE_ORGANISATION_Finance	ORGANISATION	Finance	الأوروبي (European, Europe)	البنك المركزي الأوروبي	…
مهرجان مراكش الدولي 2019 للفيلم	gNE_EVENT_Art	EVENT	Art	*مراكش* (Marrakech)	*مهرجان مراكش الدولي للفيلم*	*2019*

Location information provided by « nationality » as well as city and country names can be exploited in to enrich different kind of databases for many purposes and to establish relation with other NE classes.

3. RESULTS AND EVALUATION

The result of our LNER system is an XML database with annotated place proper names. Here are some examples:

```
<pos start="0" finish="0" content="نفق الشندغة" group="">
<interpretation>
<morphology category="C_NOUN" group="wFacility" lemma="نُفُق" root="نفق" string="نفق"
form="نفق" formv="نُفُقَ">
<traitNoun gender="Female" number="Singular" mode="Annexion" case="Accusative"/>
</morphology>
<morphology category="C_PN" group="cPN" lemma="الشندغة" root="" string="الشندغة" form=""
formv=""/>
<properties category="gNE.Location.Facility" group="gNE.Location.Facility" lemma="" root=""
string="نفق الشندغة" form="نفق "formv="نُفُقَ "/>
</interpretation>
</pos>
```

```
<pos start="0" finish="0" content="الجمهورية التونسية">
<interpretation>
<morphology category="C_NOUN" group="gNE.Location.GeoPolitical.Country"
lemma="جُمْهُورِيَّة" root="جمهر" string="الجمهورية" form="جمهورية" formv="جُمْهُورِيَّةَ">
<traitNoun gender="Female" number="Singular" mode="Determined" case="Accusative"/>
<proclitic category="C_PCL_N" string="ال" formv="اَلْ">
<traitNoun gender="" number="" mode="Determined" case="Accusative"/>
</proclitic>
</morphology>
</pos>
```

```
<pos start="0" finish="0" content="دبي">
<interpretation>
<morphology category="C_VB" group="gNE.Location.GeoAdministrative.City"
lemma="دَبَّ/يَدَبُّ" root="دبب" string="دبي" form="دبي" formv="دَبِّي">
<traitVerb pronoun="2PFS" tens="IMP_SPL_ACT"/>
</morphology>
<properties category="gNE.Location.GeoAdministrative.City"
group="gNE.Location.GeoAdministrative" lemma="دَبَّ/يَدَبُّ" root="دبب" string="دبي" form="دبي"
formv="دَبِّي">
<traitVerbpronoun="2PFS" tens="IMP_SPL_ACT"/>
</properties>
</interpretation>
</pos>
```

```
<posstart="0" finish="0" content="حديقة الخور" group="">
<interpretation>
<morphology category="C_NOUN" group="wFacility" lemma="حَدِيقَة" root="حدق" string="حديقة"
form="حديقة" formv="حَدِيقَةٌ">
<traitNoun gender="Female" number="Singular" mode="Indetermined" case="Nominative"/>
</morphology>
```

```
<morphology category="C_NOUN" group="gNE.Location.GeoAdministrative" lemma="خَوَر"
root="خور" string="الخور" form="خور" formv="خَوَرِ">
<traitNoun gender="Male" number="Singular" mode="Determined" case="Genetive"/>
<proclitic category="C_PCL_N" string="ال" formv="اَلْ">
<traitNoun gender="" number="" mode="Determined" case="Genetive"/>
</proclitic>
</morphology>
        <properties category="gNE.Location.Facility" group="gNE.Location.Facility"
lemma="" root="" string="حديقة الخور" form="حديقة الخور" formv="حَدِيقَةٌ اَلْخَوَرِ">
<traitNoun gender="Male" number="Singular" mode="Determined" case="Genetive"/>
</properties>
</interpretation>
</pos>
```

```
<pos start="0" finish="0" content="لمصرف الإمارات">
<interpretation>
<morphology category="C_NOUN" group="wEconomicOrgIndet" lemma="مُصَرِّف" root="صرف"
string="لمصرف" form="مصرف" formv="مٌّصُرفَ">
<traitNoun gender="Male" number="Singular" mode="Indetermined" case="Nominative"/>
<proclitic category="C_PCL_N" string="ل" formv="لَ">
<traitNoun gender="" number="" mode="Indetermined" case="Nominative"/>
</proclitic>
</morphology>
<morphology category="C_PN" group="gNE.Location.GeoPolitical.Country"
lemma="الإمارات" root="" string="الإمارات" form="" formv=""/>
<properties category="gNE.Organisation.Economy" group="gNE.Organisation.Economy"
lemma="" root="" string="لمصرف الإمارات" form="لمصرف "formv="لَمٌّصُرفَ "/>
</interpretation>
</pos>
```

```
<pos start="0" finish="0" content="جبال حجر" group="">
<interpretation>
<morphology category="C_NOUN" group="wGeoNaturalLocation" lemma="جِبَال" root="جبل"
string="جبال" form="جبال" formv="جِبَالٌ">
<traitNoun gender="Male" number="Singular" mode="Indetermined" case="Nominative"/>
</morphology>
<morphology category="C_NOUN" group="cNoun" lemma="حِجْر" root="حجر" string="حجر"
form="حجر" formv="حِجْرَ">
<traitNoun gender="None" number="Singular" mode="Annexion" case="Accusative"/>
</morphology>
<properties category="gNE.Location.GeoNatural" group="gNE.Location.GeoNatural"
lemma="" root="" string="جبال حجر" form="جبال حجر" formv="جِبَالٌ حِجْرَ">
<traitNoun gender="None" number="Singular" mode="Annexion" case="Accusative"/>
</properties>
</interpretation>
</pos>
```

```
<pos start="0" finish="0" content="جزيرة مسندم" group="">
<interpretation>
<morphology category="C_NOUN" group="wGeoNaturalLocation" lemma="جَزِيرَة" root="جزر"
string="جزيرة" form="جزيرة" formv="جَزِيرَةٍ">
<traitNoun gender="Female" number="Singular" mode="Indetermined" case="Genetive"/>
```

```
</morphology>
<morphology category="C_PN" group="cPN" lemma="مسندم" root="" string="مسندم" form=""
formv=""/>
<properties category="gNE.Location.GeoNatural" group="gNE.Location.GeoNatural"
lemma="" root="" string="جزيرة مسندم" form="جزيرة "formv="جَزِيرَةٍ "/>
</interpretation>
</pos>

<pos start="0" finish="0" content="الساحل الإماراتي" group="">
<interpretation>
<morphology category="C_NOUN" group="wCardinalPoint" lemma="سَاحِل" root="سحل"
string="الساحل" form="ساحل" formv="سَّاحِلِ">
<traitNoun gender="Male" 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="wNationalityMasculin" group="wNationalityMasculin"
lemma="الإماراتي" root="" string="الإماراتي" form="" formv=""/>
<properties category="gNE.Location.GeoPolitical" group="gNE.Location.GeoPolitical"
lemma="" root="" string="الساحل الإماراتي" form="الساحل "formv="اَلْسَّاحِلِ "/>
</interpretation>
</pos>

<pos start="0" finish="0" content="مطار شارل دوكول" group="">
<interpretation>
<morphology category="C_NOUN" group="wFacility" lemma="مَطَار" root="مطر" string="مطار"
form="مطار" formv="مَطَارٌ">
<traitNoun gender="Male" number="Singular" mode="Indetermined" case="Nominative"/>
</morphology>
<morphology category="gNE.Person" group="gNE.Person.PN" lemma="شارل دوكول " root=""
string="شارل دوكول " form="" formv=""/>
<properties category="gNE.Location.Facility" group="gNE.Location.Facility" lemma="" root=""
string="مطار شارل دوكول" form="مطار  "formv="مَطَارٌ "/>
</interpretation>
</pos>
```

Figure 8. Examples of our LNE recognition andannotation

Our LNE system made good results. We used two corpora for evaluation:*ANERCorp* (154 674 words), which is available online, and French Press Agency (FPA) *Corpus* (30 000 words).Here is the evaluation table (recall, precision and f-measure):

$$\text{Recall} = \frac{\text{Number of correctly annotated entities} \times 100}{\text{Number of entities in the corpus}}$$

$$\text{Precision} = \frac{\text{Number of correctly annotated entities} \times 100}{\text{Numberof annotated entities}}$$

$$\text{F-mesure} = \frac{2\text{*recall *precision}}{\text{Recall + precision}}$$

Table 2. LNER results

Corpus	NE in the corpus	correctly annotated NE	Recall	Precision	F-measure
ANERCorp	4008	3709	92,53 %	96,46%	94,45
FPA *Corpus*	2523	2344	94,96 %	97,82 %	96,36%

We exploited the result of our LNER system in the construction of Techlimed research engine filtering by different LNE class (geo-political, geo-administrative, facility, geo-natural), and different fields (politics, economy, social, sport, justice, health, science, religion, administration, etc.). Figure 9 is the research engine interface showing a cloud of NE including LNE recognised by our system:

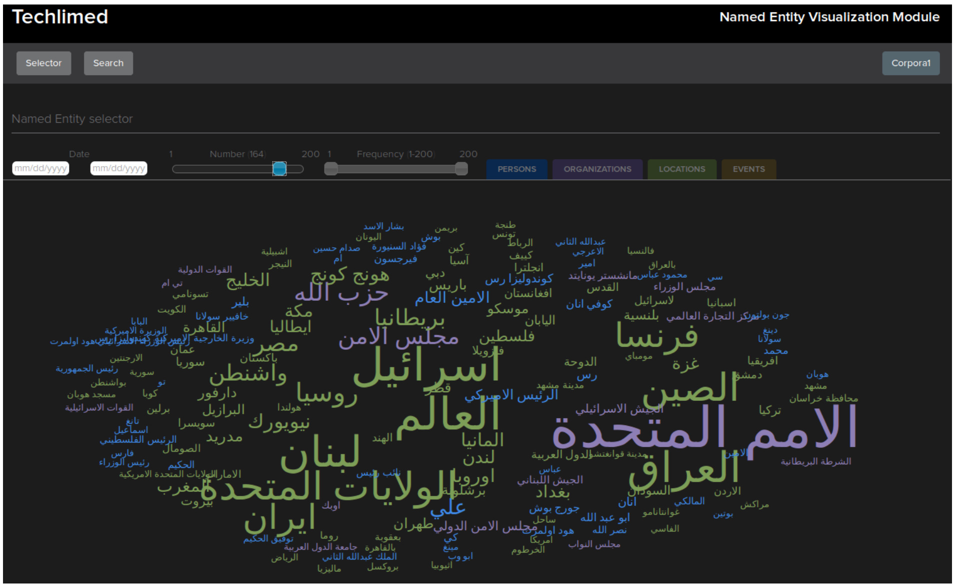


Figure 9. The LNE output in Techlimed research engine

The most frequent are bigger and clearer (bigger size): e.g.الأمم المتحدة (United Nation), الولايات المتحدة (USA), اسرائيل (Israel), فرنسا (France), الصين (China), ايران (Iran) etc. The less frequent are smaller: e.g. مجلس الأمن الدولي (the International Security Council), جامعة الدول العربية (the Arab League), مركز التجارة العالمية (the International Trade Centre), سويسرا (Switzerland) etc. The research engine uses the output of our system of NE extraction and classification to contextualise the query. The figures below (research engine pages) show some aspects of our LNER contribution in IR:

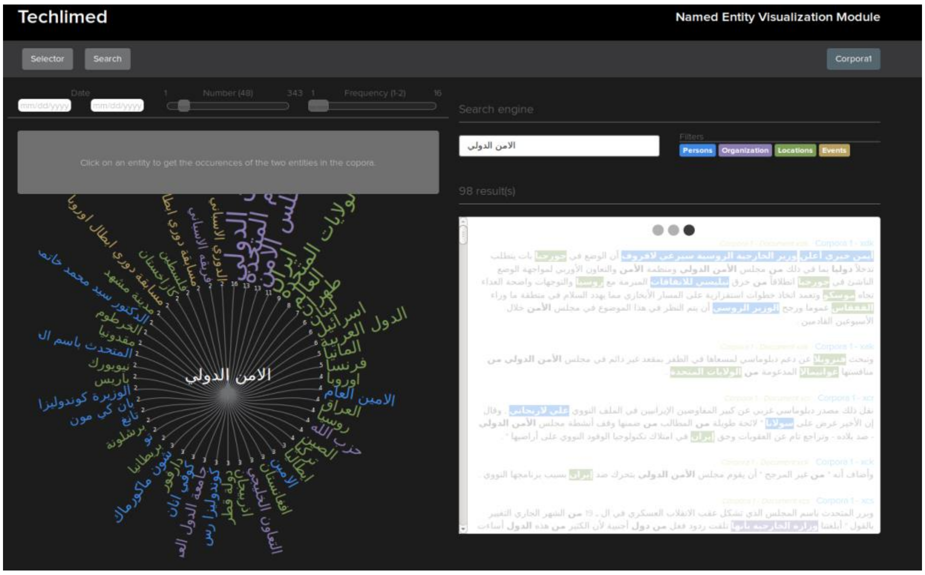


Figure 10. Research engine contextualisation of the query الامن الدولي (the international security) with NE extracted by our system

The figure shows the obtained results of the query الامن الدولي (the international security)**.** The diagram in the retrieved page on the left of the screen shows that the query is associated not only with LNE but also with the rest of NE recognised by our system; and the click on any NE on the left diagram gives access to the text with the corresponding context on the right side of the page. The research engine also uses the NE classification to filter by classes and subclasses. The following figure is an example a query contextualisation:

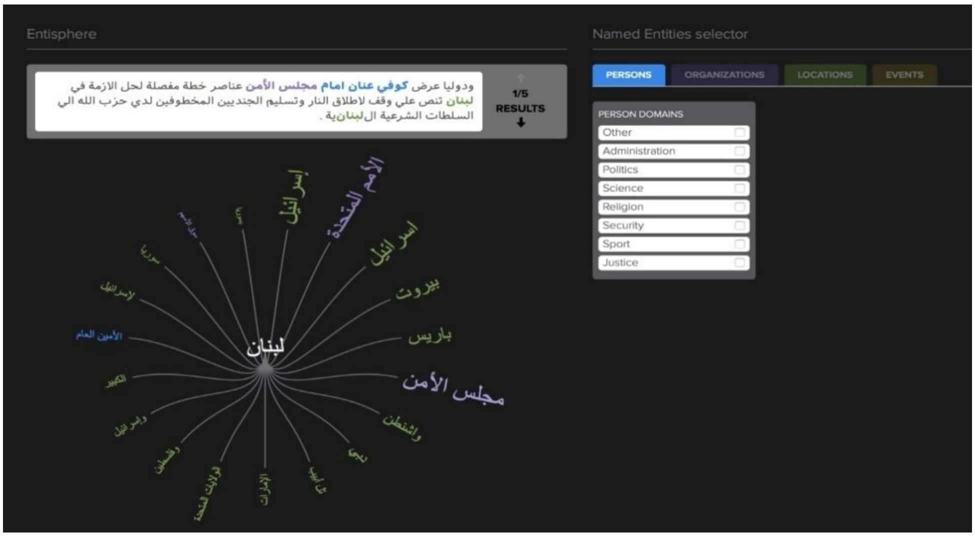
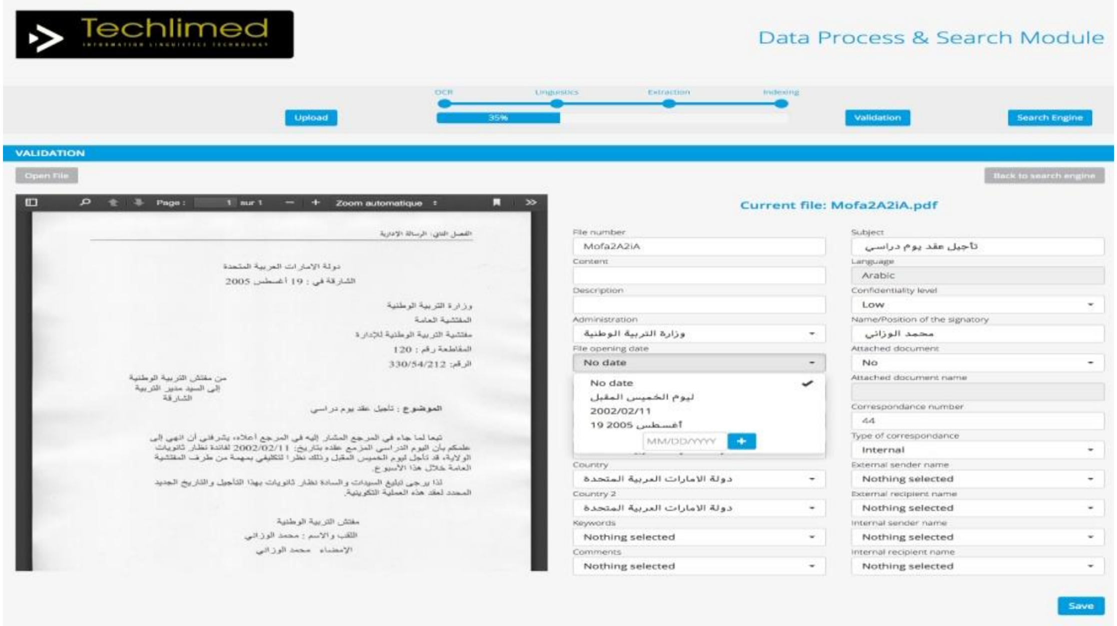

ودوليا عرض كوفي عنان امام مجلس الأمن عناصر خطة مفصلة لحل الازمة في لبنان تنص علي وقف لاطلاق النار وتسليم الجنديين المخطوفين لدي حزب الله الي السلطات الشرعية اللبنانية .

Figure 11. Contextualisation of the query لبنان (Lebanon) and filtering by NE classes and subclasses

As we can see, we can filter by class (e.g. PERSON, ORGANISATION, LOCATION, and EVENT) and by field (politics, science, religion, sport, economy etc.).

We also developed an information extraction application using the output of our LNE extraction and classification system. The figures bellow shows the contribution of our NE recognition system in information extraction from administrative letters:



الفصل الثاني: الرسالة الإدارية

دولة الإمارات العربية المتحدة
الشارقة في : 19 أغسطس 2005

وزارة التربية الوطنية
المفتشية العامة
مفتشية التربية الوطنية للإدارة
المقاطعة رقم : 120
الرقم: 330/54/212

من مفتش التربية الوطنية
إلى السيد مدير التربية
الشارقة

الموضوع : تأجيل عقد يوم دراسي

تبعا لما جاء في المرجع المشار إليه في المرجع أعلاه، يشرفني أن أنهي إلى علمكم بأن اليوم الدراسي المزمع عقده بتاريخ: 2002/02/11 لفائدة نظار ثانويات الولاية، قد تأجل ليوم الخميس المقبل وذلك نظرا لتكليفي بمهمة من طرف المفتشية العامة خلال هذا الأسبوع.

لذا يرجى تبليغ السيدات والسادة نظار ثانويات بهذا التأجيل والتاريخ الجديد المحدد لعقد هذه العملية التكوينية.

مفتش التربية الوطنية
اللقب والاسم : محمد الوزاني
الإمضاء محمد الوزاني

Current file: Mofa2A2iA.pdf

File number: Mofa2A2iA
Content:
Description:
Administration: وزارة التربية الوطنية
File opening date: No date
No date ✓
ليوم الخميس المقبل
2002/02/11
19 2005 أغسطس
MM/DD/YYYY +
Country: دولة الامارات العربية المتحدة
Country 2: دولة الامارات العربية المتحدة
Keywords: Nothing selected
Comments: Nothing selected

Subject: تأجيل عقد يوم دراسي
Language: Arabic
Confidentiality level: Low
Name/Position of the signatory: محمد الوزاني
Attached document: No
Attached document name:
Correspondance number: 44
Type of correspondance: Internal
External sender name: Nothing selected
External recipient name: Nothing selected
Internal sender name: Nothing selected
Internal recipient name: Nothing selected

Figure 12 : Information extraction based on our NE extraction and classification

Here les NE annotation are used to fill in the form: administration, country, country2 etc. We can use the same LNE system in any other IE system.

4. Conclusion

This paper shows our LNR system and the importance of a linguistic approach in this task. The process is composed of many processing levels going form lexical to syntactico-semantic analysis. The LNE syntactic and semantic information are vital to their extraction and classification. We exploited the output of our system within Techlimed in the information retrieval and extraction applications. We integrated the obtained fine annotations made of LNE classes and subclasses in the systems of indexations for an efficient information retrieval and extraction. We can also extend our analysis to sentence annotation using verb classification as predicates. The objective is, to exploit the output enriched by these semantic annotations, to develop the project of the extraction of relations between different extracted NE classes within sentences. That is to say, wecan extract and classify predicate relationship between sentence NE arguments. The obtained results can be used in ontology enrichment and semantic analysis of sentences for many NLP applications like CAT and Controlled Arabic.

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] Attia, M., Toral, A., Tounsi, L., Monachini M., Van Genabith, J. (2010). "An automatically built NE lexicon for Arabic". In *LREC 2010, 7th conference on International Language Resources and Evaluation*. Valletta, Malta.

[3] Beaudet S. (2002).Extraction et analyse sémantique automatique des entités spatiales géographiques. Intership report for computer science master.

[4] Bodenhamer, D.J., Harris, T.M., and Corrigan, J., (2013). "Deep mapping and the spatial humanities". In *International Journal of Humanities and Arts Computing*, 7 (1–2), 170–175.

[5] Cooper, D., Donaldson, C., and Murrieta-Flores, P., eds., (2016). *Literary mapping in the digital age. Digital research in the arts and humanities*. London, UK: Routledge.

[6] Daille B., Fourour N., Morin E. (2000). ‘‘Catégorisation des noms propres : une étude en corpus’’. In *Cahiers de Grammaire*, Vol 25, pp. 115-129.

[7] Ehrmann, M. (2008). Les entités nommées, de la linguistique au TAL: statut théorique et méthodes de désambiguïsation. Thesis (PhD). Paris 7 University.

[8] Ehrmann, M., Jacquet, G. (2006). ‘‘Vers une double annotation des entités nommées’’ . In *Traitement Automatique du Langues*, Vol. 47, pp. 63-88.

[9] El Maarouf, I., Villaneau, J., Rosset, S. (2011). ‘‘Extraction de patrons sémantiques appliquées à la classification d'entités nommées’’. In *TALN*. Montpellier.

[10] Gregory, I., Donaldson, D., Murrieta-Flores, P., Rayson, P., (2015). ‘‘Geoparsing, GIS, and textual analysis: current developments in spatial humanities research’’.In *International Journal of Humanities and Arts Computing*, 9 (1), 1–14. doi:10.3366/ijhac.2015.0135.

[11] Leidner, J.L., (2007). Toponym resolution in text: annotation, evaluation and applications of spatial grounding of place names. Thesis (PhD). The University of Edinburgh.

AUTHOR

I am **Omar ASBAYOU**, a teacher in Lumière University Lyon 2, and a memberin CRTT laboratory. My research focuses on Arabic language processing. I was an engineer researcher in Techlimed, a company specialised inThe automatic processingof Arabic