

# **A resource-light\* Arabic Named Entity Recognition system**

**Wajdi Zaghouni**

University of Pennsylvania  
Linguistic Data Consortium

**Ralf Steinberger and Bruno Pouliquen**

European Commission  
Joint Research Center

\* Just 30 calories per serving

# ***Agenda***

## **◆ 1. Introduction and General goal**

# ***Agenda***

- ◆ **1. Introduction and General goal**
- ◆ **2. Related work**

# *Agenda*

- ◆ **1. Introduction and General goal**
- ◆ **2. Related work**
- ◆ **3. Method**

# ***Agenda***

- ◆ **1. Introduction and General goal**
- ◆ **2. Related work**
- ◆ **3. Method**
- ◆ **4. Evaluation**

# ***Agenda***

- ◆ **1. Introduction and General goal**
- ◆ **2. Related work**
- ◆ **3. Method**
- ◆ **4. Evaluation**
- ◆ **5. Conclusion**

# ***What is NER?***

## **Named Entity Recognition (NER) Task:**

→ Locate named entities in a given text

# ***What is NER?***

## **Named Entity Recognition (NER) Task:**

→ Locate named entities in a given text

### **Entities of interest include:**

- ◆ Person names
- ◆ Place names
- ◆ Company names
- ◆ Dates, times
- ◆ Dollar/currency amounts
- ◆ Percentages
- ◆ Ratios, averages, sums, other statistics



# ***Named Entities classification history***

## ◆ The Message Understanding Conference (MUC-6) distinguished 3 categories

### ■ ENAMEX (“entity name expression”)

- ◆ people, organization and locations

### ■ NUMEX (“numeric expression”)

- ◆ currency and percentages

### ■ TIMEX (“time expression”)

- ◆ temporal expressions – dates and times

## ***Why identifying Named Entities matters?***

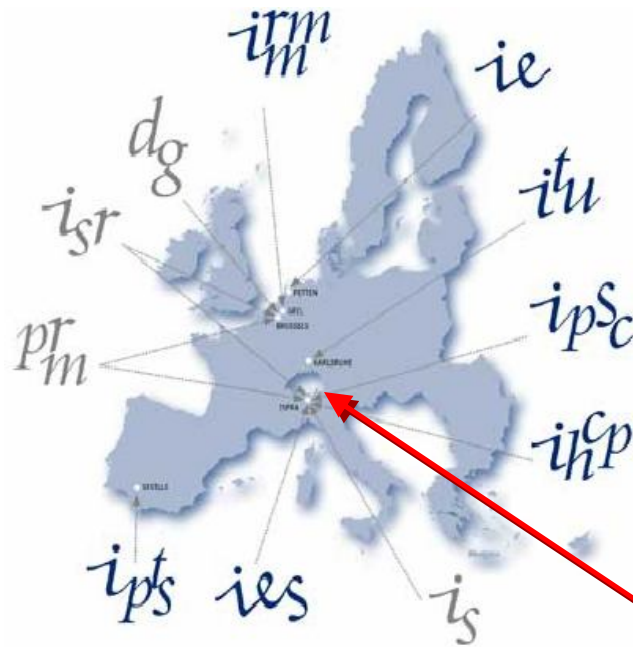
**Identifying NE is useful for many applications like the following:**

- ◆ Document classification/clustering
- ◆ Machine Translation
- ◆ Text Summarization
- ◆ Indexing and Search
- ◆ Tool or component of IE and IR
- ◆ Question Answering (QA)

## ***Goal and Project Context***

- **Goal** : Adapting an existing language independent multilingual NER system to the Arabic language.
- In cooperation with the Open Source Text Information Mining and Analysis Group (OPTIMA)
- A language Technology research group at the Joint Research Centre in Ispra, Italy and part of the European Commission

# Joint Research Center



## **BRUSSELS (BE)**

The Directorate General (DG)

The Institutional and Scientific Relations Directorate (ISR)

The Programme and Resource Management Directorate (PRM)

## **GEEL (BE)**

The Institute for Reference Materials and Measurements (IRMM)

## **KARLSRUHE (DE)**

The Institute for Transuranium Elements (ITU)

## **ISPRA (IT)** Download the Ispra site Brochure ([English](#) - [Italian](#))

The Institute for the Protection and Security of the Citizen (IPSC)

The Institute for Environment and Sustainability (IES)

The Institute for Health and Consumer Protection (IHCP)

The Ispra site Directorate (IS)

## **PETTEN (NL)**

The Institute for Energy (IE)

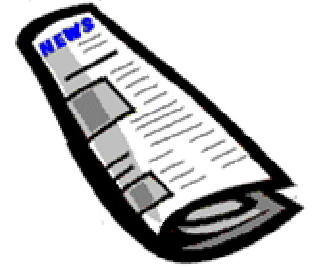
## **SEVILLE (E)**

The Institute for Prospective Technological Studies (IPTS)

<http://langtech.jrc.it/>



## ***Europe Media Monitor (EMM) News Explorer System***





- Monitors ~ 2,200 news sources
- Gathers 80,000 – 100,000 news articles per day
- In about 50 languages
- Visits some sites every 5 minutes
- Extracts text from the web page
- **Extract and to display NE to the end users**

**Free access Available online at <<http://emm.newsexplorer.eu/>>**



# EMM News Explorer – Multilingual daily news overview


EMM NewsBrief
EMM NewsExplorer



## NewsExplorer

RSS RSS feed for the latest news summary  
Daily News Analysis, across languages and over time

**Main Menu**

News Summary  
About EMM NewsExplorer

**News language and date**

Language or country:

en - English

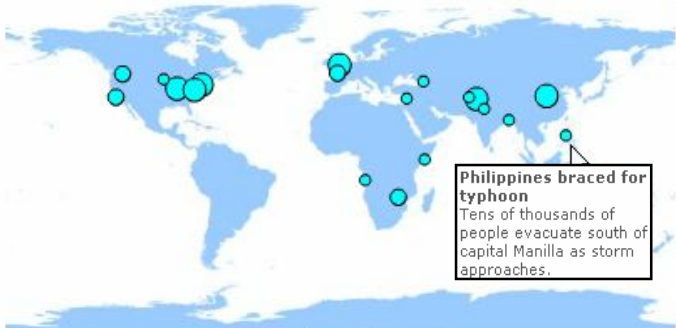
- ar - Arabic
- bg - Български
- es - Español
- de - Deutsch
- da - Dansk
- en - English
- et - Eesti keel
- fa - Farsi
- fr - Français
- it - Italiano
- nl - Nederlands
- no - Norsk
- pl - Polski
- pt - Português
- ro - Română
- ru - Russian
- sl - Slovenščina
- sv - Svenska
- tr - Türkçe

**Countries**

- AT - Austria
- BE - Belgium
- DE - Germany
- ES - Spain
- FR - France
- GB - United Kingdom
- IT - Italy
- NL - Netherlands

### Clustered news for Thursday, November 22, 2007

Read more...



View with Google Earth

**Countries**

- United States (492)
- United Kingdom (196)
- Russian Federation (83)
- Zimbabwe (67)
- France (62)
- China (58)
- Italy (58)
- Afghanistan (56)
- Uganda (54)
- Angola (48)
- South Africa (45)
- Pakistan (42)
- India (38)
- Netherlands (38)
- Canada (36)
- Congo, The Democratic Republic Of The (36)
- Philippines (34)
- Australia (33)
- Spain (33)
- Saudi Arabia (31)
- Turkey (30)
- Korea, Republic Of (30)
- Korea, Democratic People's Republic Of (29)
- Greece (28)

**Pakistan court dismisses final Musharraf challenge [32]**

de es fr it nl ar bg da pl pt ru sl sv tr

Pakistan's new-look Supreme Court has, as expected, dismissed the last challenge to President Pervez Musharraf's re-election.  
*euronews-en 2:31:00 PM CET*

**Iran heeding transparency pledge, more needed - IAEA [32]**

de es fr it da pt tr

efforts on schedule, countering Western doubts, but Tehran must step up cooperation to resolve remaining questions this year. Mohamed ElBaradei summarised findings of an International Atomic Energy Agency report on Iran at a debate of the IAEA's governing board, where differences simmered over....  
*austrianews 6:23:00 PM CET*

**Chinese bluster on Tibet and Taiwan [29]**

de es fr it nl da pt sv

Contrary to expectations, China is not doing much to soften its image ahead of the Beijing Olympics by allowing its domestic critics to speak their minds or championing human rights in Sudan. Instead, Chinese leaders are defending authoritarian rule at home and abroad and waiving aggressive diplomacy against those who

**This Week's New Stories**

Claimants warned of fraud fears  
November 20, 2007 - November 22, 2007  
Search resumes at house after double body find  
November 15, 2007 - November 17, 2007  
Mbeki seeks Mugabe deal for summit  
November 15, 2007 - November 22, 2007  
Ex-Rhodesia leader Ian Smith dies  
November 20, 2007 - November 22, 2007  
Tapas Nine witness says 'Mediterranean man took Madeleine'  
November 17, 2007 - November 21, 2007  
Queen and Duke mark anniversary  
November 17, 2007 - November 20, 2007  
Philippines in 'separatist deal'  
[Read more...](#)

**This Month's New Stories**

Cyclone leaves 242 dead in Bangladesh  
November 14, 2007 - November 22, 2007  
Georgia declares state of emergency  
October 31, 2007 - November 22, 2007  
France gripped by massive strike  
November 12, 2007 - November 22, 2007  
Space crew fixes solar wing  
October 23, 2007 - November 14, 2007  
Vegas showdown: Clinton 'on the hot seat'  
October 23, 2007 - November

**Related People**

- Steve McClaren (42)
- George W. Bush (29)
- Pervez Musharraf (26)
- Brian Barwick (20)
- Nicolas Sarkozy (18)
- Mohamed ElBaradei (14)
- Kevin Rudd (14)
- Martin O'Neill (14)
- Vladimir Putin (14)
- John Howard (13)
- Geoff Thompson (12)
- José Mourinho (12)
- Gordon Brown (11)
- Hugo Chavez (11)
- Mahmoud Abbas (10)
- Sven Goran Eriksson (10)
- Jaap de Hoop Scheffer (10)

# *Agenda*

- ◆ 1. Introduction and General goal
- ◆ 2. **Related work**
- ◆ 3. Method
- ◆ 4. Evaluation
- ◆ 5. Conclusion

# ***Named entities recognition methods***

- ◆ **Machine Learning methods**
  - **Supervised learning**



# ***Named entities recognition methods***

- ◆ **Machine Learning methods**
  - **Supervised learning**
  - **Lazy learning**

# ***Named entities recognition methods***

## **◆ Machine Learning methods**

- Supervised learning**
- Lazy learning**
- weakly supervised**

# ***Named entities recognition methods***

## ◆ **Machine Learning methods**

- **Supervised learning**
- **Lazy learning**
- **weakly supervised**
- **Unsupervised learning**

# ***Named entities recognition methods***

## **◆ Rule based methods**

- Based on Lexicon (e.g : known names list)

# ***Named entities recognition methods***

## **◆ Rule based methods**

- Based on Lexicon (e.g : known names list)
- Local grammar (It describes patterns to match NEs)

# ***Named entities recognition methods***

## ◆ **Rule based methods**

- Based on Lexicon (e.g : known names list)
- Local grammar (It describes patterns to match NEs)
- Internal and external features (or evidences)
  - Capitalization
    - ◆ Not applicable for arabic
  - Trigger words

*“The president Barack Obama”*

## ◆ Hybrid methods

- Combination between the machine learning method and the rule based method

## ◆ Hybrid methods

- Combination between the machine learning method and the rule based method
- Lexicon



## ◆ Hybrid methods

- Combination between the machine learning method and the rule based method
- Lexicon
- Local grammar (Hand written rules)

## ***Existing Arabic Named entity recognition system (1)***

- ◆ Many of the available systems are commercial systems and do not have any publications on their systems with the exception of *IdentiFinder* from BBN (Bikel et al. 1999)
  
- ◆ 1. *NetOwlExtractor* (NetOwl)
- ◆ 2. *Siraj* (Sakhr)
- ◆ 3. *Clear Tags* (ClearForest)
- ◆ 4. *InXight-Smart-Discovery-Entity-Extractor* (InXight)
- ◆ 5. *NameFinder* (AppTek)
- ◆ 6. *Rosette Named Entity Extractor* (BasisTech)
- ◆ 7. *IdentiFinder* from (BBN)
- ◆ 8. *InfoSphere Global Name Recognition* (IBM)
- ◆ 9. *Nfinder* (Nstein)
- ◆ 10. *Temis TM360* (TEMIS)

## ***Existing Arabic Named entity recognition system (2)***

- ◆ **Traboulsi (2009)** discussed the possible use of a local grammar-approach to build an Arabic named entity system (just a prototype)

## ***Existing Arabic Named entity recognition system (2)***

- ◆ **Traboulsi (2009)** discussed the possible use of a local grammar-approach to build an Arabic named entity system (just a prototype)
- ◆ **Shalaan and Raza (2008)** presented (**NERA**) a Name Entity Recognition tool for Arabic using a rule-based approach, dictionaries and a local grammar.

## ***Existing Arabic Named entity recognition system (2)***

- ◆ **Traboulsi (2009)** discussed the possible use of a local grammar-approach to build an Arabic named entity system (just a prototype)
- ◆ **Shalaan and Raza (2008)** presented (**NERA**) a Name Entity Recognition tool for Arabic using a rule-based approach, dictionaries and a local grammar.
- ◆ **Benajiba et al. (2007)** developed **ANERsys**, a NER system based on a Maximum Entropy-based statistical learning model.

## ***Existing Arabic Named entity recognition system (2)***

- ◆ **Traboulsi (2009)** discussed the possible use of a local grammar-approach to build an Arabic named entity system (just a prototype)
- ◆ **Shalaan and Raza (2008)** presented (**NERA**) a Name Entity Recognition tool for Arabic using a rule-based approach, dictionaries and a local grammar.
- ◆ **Benajiba et al. (2007)** developed **ANERsys**, a NER system based on a Maximum Entropy-based statistical learning model.
- ◆ **Samy et al. (2005)** used a parallel corpus in Arabic and in Spanish and an NE tagger in Spanish to tag the names in the Arabic corpus

## ***Existing Arabic Named entity recognition system (2)***

- ◆ **Traboulsi (2009)** discussed the possible use of a local grammar-approach to build an Arabic named entity system (just a prototype)
- ◆ **Shalaan and Raza (2008)** presented (**NERA**) a Name Entity Recognition tool for Arabic using a rule-based approach, dictionaries and a local grammar.
- ◆ **Benajiba et al. (2007)** developed **ANERsys**, a NER system based on a Maximum Entropy-based statistical learning model.
- ◆ **Samy et al. (2005)** used a parallel corpus in Arabic and in Spanish and an NE tagger in Spanish to tag the names in the Arabic corpus
- ◆ **Abuleil (2004)** developed a rule-based system that makes use of hand-written rules and trigger words.

## ***Existing Arabic Named entity recognition system (2)***

- ◆ **Traboulsi (2009)** discussed the possible use of a local grammar-approach to build an Arabic named entity system (just a prototype)
- ◆ **Shalaan and Raza (2008)** presented (**NERA**) a Name Entity Recognition tool for Arabic using a rule-based approach, dictionaries and a local grammar.
- ◆ **Benajiba et al. (2007)** developed **ANERsys**, a NER system based on a Maximum Entropy-based statistical learning model.
- ◆ **Samy et al. (2005)** used a parallel corpus in Arabic and in Spanish and an NE tagger in Spanish to tag the names in the Arabic corpus
- ◆ **Abuleil (2004)** developed a rule-based system that makes use of hand-written rules and trigger words.
- ◆ **Maloney and Niv (1998)** presented **TAGARAB**, an Arabic name recognizer that combines the pattern matching module with a morphological analyzer to improve performance



# *Agenda*

- ◆ **1. Introduction and General goal**
- ◆ **2. Related work**
- ◆ **3. Method**
- ◆ **4. Evaluation**
- ◆ **5. Conclusion and future work**

## ***What approach have we used to build the system ?***

- ◆ We have used a rule based approach

## ***What approach have we used to build the system ?***

- ◆ We have used a rule based approach
- ◆ We have created various lexicons of (person names, organizations, locations...)
  - → We use the lexicon entries to match directly the named entities
  - → We also use the lexicon entries in our local grammar to match partially unknown entities

## ***What approach have we used to build the system ?***

- ◆ We have used a rule based approach
- ◆ We have created various lexicons of (person names, organizations, locations...)
  - → We use the lexicon entries to match directly the named entities
  - → We also use the lexicon entries in our local grammar to match partially unknown entities
- ◆ We use local grammar to detect potential Named entities not in our lexicon

## ***What approach have we used to build the system ?***

- ◆ We have used a rule based approach
- ◆ We have created various lexicons of (person names, organizations, locations...)
  - → We use the lexicon entries to match directly the named entities
  - → We also use the lexicon entries in our local grammar to match partially unknown entities
- ◆ We use local grammar to detect potential Named entities not in our lexicon
- ◆ We built a list of Trigger words and Stopwords

## ***But why a rule based NER system?***

- ◆ This approach is motivated by the characteristics of Arabic language.

## ***But why a rule based NER system?***

- ◆ This approach is motivated by the characteristics of Arabic language.
- ◆ Practical and easy to implement when compared to machine learning approaches that usually require large tagged corpora, dictionaries.

## ***But why a rule based NER system?***

- ◆ This approach is motivated by the characteristics of Arabic language.
- ◆ Practical and easy to implement when compared to machine learning approaches that usually require large tagged corpora, dictionaries.
- ◆ The EMM system is optimized for rule based systems



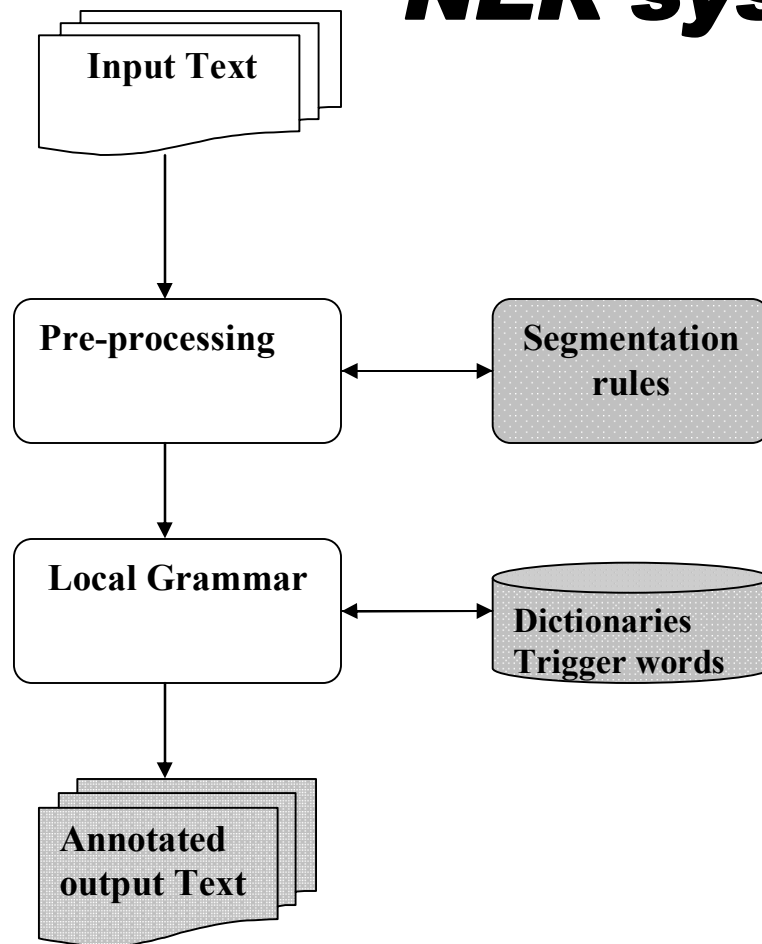
## ***But why a rule based NER system?***

- ◆ This approach is motivated by the characteristics of Arabic language.
- ◆ Practical and easy to implement when compared to machine learning approaches that usually require large tagged corpora, dictionaries.
- ◆ The EMM system is optimized for rule based systems
- ◆ This open architecture approach provides flexibility and adaptability features in our system and it can be easily configured to work with different languages and domains

# ***General architecture***

- ◆ Our named entities recognition process takes two cycles, using the existing lexicon component and then applying the grammar rules as shown in the following figure

# ***The proposed Arabic NER system***



## **Figure 1 : Architecture of the System**

- ◆ The complex inflection system of the Arabic language can be dealt with by using hand-crafted rules, which enable to strip off **possible prefixes** and **suffixes** from the word stem before applying our local grammar.

- ◆ The complex inflection system of the Arabic language can be dealt with by using hand-crafted rules, which enable to strip off **possible prefixes** and **suffixes** from the word stem before applying our local grammar.
- ◆ By doing this, particles attached to words are stripped, **allowing a better match between the words found in the text and those in the dictionaries.**

## *Morphological pre-processing (2)*

- ◆ For example, the attached conjunction ***wa***, the preposition ***li*** and the definite article ***al*** are stripped off all words and names, like in the example shown in the next page

◆ For instance, the attached conjunction **wa**, the preposition **li** and the definite article **al** are stripped off all words and names, like in the example shown in the next page

◆ وللرئيس /**walilraais** / « and for the president».

**Transformation 1** : removal of the attached conjunction و wa  
« and ».

- ◆ For example, the attached conjunction **wa**, the preposition **li** and the definite article **al** are stripped off all words and names, like in the example shown in the next page

- ◆ **واللرئيس /walilraais / « and for the president».**

**Transformation 1** : removal of the attached conjunction و wa « and ».

- ◆ **للرئيس /lilraais/ « for the president ».**

**Transformation 2** : removal of the attached preposition ل /li/ « for » and of the definite article ال / il / « the ».



- ◆ For example, the attached conjunction **wa**, the preposition **li** and the definite article **al** are stripped off all words and names, like in the example shown in the next page

- ◆ والرئيس /**walilraais** / « and for the president».

**Transformation 1** : removal of the attached conjunction و wa « and ».

- ◆ للرئيس /**ilraais**/ « for the president ».

**Transformation 2** : removal of the attached preposition ل /li/ « for » and of the definite article ال / il / « the ».

- ◆ رئيس /**raais**/ « president ».

## **Corpora**

- ◆ For training and testing purposes, we have compiled various Arabic online sources containing texts which are diverse in terms of domain, format, style and genre (Aljazeera, Alarabiya, BBC Arabic, weblog, Wikipedia articles...)

### **Trigger Words**

They have been created manually, but they are mostly built semi-automatically, by looking at the most frequent left and right contexts of known Arabic NEs, or at the context of NEs that are found using the rules with initial lists of seed words

Trigger Words helps in identifying a NE within text but does not get recognized itself

## ***Building Resources (3)***

- ◆ Sample person names trigger words that we built

قال
أشار
ذكر
أضاف
أعلن
اقترح
علق
خبر
أفاد
خلص
حدث

## *Dictionaries (1)*

- ◆ Extensive list or common first names (not only Arabic firstnames but also common international names like John, Jean, Juan written in Arabic).
- ◆ Similarly we compiled a list of common family names (mostly Arabic).

- ◆ The wordlists for known first and last names contain altogether 19,600, names: 17,000 first names and 2,600 last names of people, collected from various Arabic Internet resources, including the Arabic version of Wikipedia.

## ***Dictionnaires (3)***

- ◆ The organization lexicon is limited to a list of 4k famous company and organization names

## *Dictionaries (4)*

- ◆ Geographical locations are recognized through a simple gazetteer lookup procedure, without any grammatical patterns.
- ◆ The gazetteer consists of around 2K names of countries, cities, towns and villages found mainly in the multilingual KNAB gazetteer produced by the *Institute of Estonian Language*[\[1\]](http://www.eki.ee/knab/p_mm_en.htm).

[1] <[http://www.eki.ee/knab/p\\_mm\\_en.htm](http://www.eki.ee/knab/p_mm_en.htm)>,



## *StopWords list*

- ◆ Stop words are frequent words of which we know that they **cannot be proper names parts**
- ◆ Consist mainly of (Pronouns, Particles, Verbs, Demonstratives...):
- ◆ هما , خلال , أكل , في , من , إن , انه
- ◆ We use name-stop words at the beginning or at the end of the rules to limit the right and left border of the detected entity

## ***Local Grammars (1)***

- ◆ The rules to guess person names use a list of local patterns based on cues or trigger words like :
- ◆ **Titles:** السيد , السيدة
- ◆ **Professions or positions:** الدكتور
- ◆ **Country adjectives:** الإماراتي , مصري
- ◆ **Religious and ethnic groups:** الشيعي , السني

## *Local Grammars (2)*

◆ {PERSON\_TRIGGER}\b(\w+)\b{LASTNAME}

The following name would be recognized by this rule

◆ السيد أحمد عيسى

◆ *Mr. Ahmed Issa*

A person name preceded by a title and composed of an unknown string assumed to be a first name followed by a last name based on a preceding person trigger words.

- ◆ Arabic proper names have the particularity to use common middle names (patronyms) like ( عبد, أبو, بن ) so we use them to recognize proper names in rules such as the following:

## Local Grammars (4)

◆ ({FIRSTNAME})(\b(\w+)\b{MIDDLENAME})  
+\b{LASTNAME})

محمد علي بن حليلة

◆ This person name will be recognized by this rule as long as *Mohammed* is a known firstname and *Halima* is a known last name

## Local Grammars (5)

`{NAME_STOP_WORDS}\b(\w+)(\b{PERSON_TRIGGER})*\b{MODIFIER}+`

ان حامد كرزاي الرئيس الأفغاني الجديد

*Indeed Hamid Karzai, the new Afghani president*

(..) : Expression found between parantheses will be detected as Named entity

{..} : Expressions between accolades will not be displayed as a Named Entity

\b : White space

\w: Unknown word between 2 white spaces

+ : Expression replaced one or more times

\* : Expression replaced zero or more times

## Local Grammars (6)

{({TRIGGER\_ORG\_BEG}\b{FIRSTNAME}\b{MIDDLENAME}\b{LASTNAME}  
\b{TRIGGER\_ORG\_END})}

شركة محمد أبو الحمдاني و إخوان

The company of Mohamed Abu Alhamadani and brothers

(..) : Expression found between parantheses will be detected as Named entity

{..} : Expressions between accolades will not be displayed as a Named Entity

\b : White space

\w: Unknown word between 2 white spaces

+ : Expression replaced one or more times

\* : Expression replaced zero or more times

## Local Grammars (7)

◆ ({number}/b{measure})

◆ 85 كغ « 85 kg » et 130 صم « 130 cm ».

◆ ({DATE-COMPLEX}/b{DATE\_NUMBER})

◆ 2007 عيد القيامة



# *Agenda*

- ◆ **1. Introduction and General goal**
- ◆ **2. Related work**
- ◆ **3. Method**
- ◆ **4. Evaluation**
- ◆ **5. Conclusion**

# ***Evaluation Corpus***

- ◆ The evaluation corpus was built from online news sources covering in equal parts the Tunisian newspaper Assabah<sup>[1]</sup> and the Lebanese newspaper Alanwar<sup>[2]</sup>.
- ◆ We have manually tagged the NEs in the corpus.
- ◆ Our corpus consists of 34K tokens (18K for Assabah and 16K for Alanwar). <sup>[1]</sup>  
<<http://www.assabah.com.tn/>>  
<sup>[2]</sup> <<http://www.alanwar.com/ar/>>

# ***Evaluation Corpus***

Category	Distribution and Ratio	
Persons	804	43,43 %
Locations	433	23,39 %
Organizations	514	27,76 %
Dates	54	2,91 %
Numeric expressions	46	2,48 %
Total	1 851	100 %

**Table I : Distribution and ratio of Named Entities by category**

# ***Evaluation Metrics***

## ◆ *Precision*

- Accuracy of the system versus all it's generated values

## ◆ *Recall*

- Accuracy of the system vs. all values that *should* have been generated

## ◆ *F-Score*

- Weighted harmonic mean of Precision and Recall

# Results

Category	Precision	Recall	F-measure
Person	87 %	66,54 %	75,40 %
Organization	69,96 %	35,79	47,34 %
Location	91,52 %	74,82 %	82,33 %
Date and time	96,13 %	94,11 %	95,10 %
Numeric expression	93,29 %	89,47 %	91,34
Overall	87,17 %	65,74 %	74,95 %

**Table 2.** Results obtained with the various categories of Named entities

## ***Results***

- ◆ The achieved system performance results were satisfactory when evaluated against the standard measures; precision, recall, and f-measure

## ***Difficulties and challenges***

- ◆ One major challenge was the handling of ambiguity (a given Arabic verb can also act as proper noun for example *Hamad*)
- ◆ Another challenge was to predict the boundary of the named entity especially with long and composed Arabic names.
- ◆ We found cases where the full name is composed of 8 words. Moreover, it was very difficult to cover with our local grammar all regional variants of the names of organizations as labels and standards differ from one country to another and from one culture to another.

## ***Conclusion***

- ◆ We proposed
  - Rule-based Named entity recognition tool
  - Resource-light system
- ◆ Needs better tuning
  - Rules
  - Dictionary
  - StopWords
  - Handling ambiguous cases
  - Adding a POS tagger will help



# النهاية

شكرا لكم على حسن الإستماع

Thank you!