



Adapting a resource-light highly multilingual Named Entity Recognition system to Arabic

Wajdi Zaghouni¹, Bruno Pouliquen², Mohamed Ebrahim² & Ralf Steinberger²

(1) Linguistic Data Consortium LDC, USA
(2) European Commission – Joint Research Centre, Italy

Objective

- Adapt an existing language-independent named entity recognition (NER) system to cover Arabic.
- Use it for daily large-scale information extraction and media monitoring within the *Europe Media Monitor* family of applications.

Europe Media Monitor (EMM) – Multilingual news analysis

- The EMM family of applications was developed at the European Commission's *Joint Research Centre*.
- Freely accessible** at: <http://emm.newsbrief.eu/overview.html>;
- EMM gathers an average of **100,000 news articles per day** in **50 languages**.
- EMM clusters and classifies the articles, follows topics over time (topic detection and tracking) and detects trends (alerting).
- EMM-NewsExplorer **applies text mining tools to 20 languages**. Tools include:
 - Recognition and disambiguation (grounding) of persons, organisations, locations;
 - Name variant matching, including across languages and scripts e.g. *Javier Solana*, *Khavier Solana*, *خافيير سولانا*, *Хавьер Солана*, ...;
 - Recognition of quotations by and about people;
 - Linking related news across languages, for all language pairs;
 - Produce social networks based on multilingual media information;
 - Gather and aggregate multilingual information about people.

Multilingual NER rules in EMM

- Manually produced NER rules are language-independent, but make reference to language-specific parameter files.
- The **language-specific parameter files** contain long lists of trigger words: titles, professions, ethnic groups, religious groups, modifiers, determiners, stop words, quotation verbs, etc.
- These **trigger word lists** were produced semi-automatically, using machine learning and boot-strapping methods.
- Light-weight**: no use of parsers, POS taggers, large-scale dictionaries.
- Sample NER rules:
 - PERSON_TRIGGER+**b**UppercaseWord**b** UppercaseWord
Swiss world champion Roger Federer
 - UppercaseWord**b**UppercaseWord(**b**MODIFIER)***b** PERSON_TRIGGER+
Hamid Karzai, the newly elected Afghan president
- About 600 new names are detected every day and added to a list of over 1 million known names. Known names are then recognised through a simple lookup.

en death of former Prime Minister Rafik Hariri, blamed by many opposition
 es asesinato del ex primer ministro Rafic al-Hariri, que la oposición atribuyó
 fr l'assassinat de l'ex-dirigeant Rafic Hariri et le départ du chef de la diplom
 nl na de moord op oud-premier Rafiq al-Hariri gingen gisteren bijna een
 de libanesischen Regierungschef Rafik Hariri vor einem Monat wichtige B
 sl danjelega libanonskega premiera Rafika Haririja. Libanonska opozicija si
 et möödumisele ekspeaminister Rafik al-Hariri surma põhjustanud pommipl
 ar اغتيال رئيس الوزراء السابق رفيق الحريري بأحد يهودية وما حدث سابقا
 ru Бывший премьер-министр Ливана Рафик Харирри, который

Figure: NER examples for 9 out of the 20 languages covered.

Pervez Musharraf
 Information about this person was last updated on Tuesday, September 15, 2009

Names	Key titles and phrases	External resources
Pervez Musharraf (de,sv)	président pakistanais (fr - 400)	
General Pervez Musharraf (de,sv)	presidente paquistanés (it - 281)	
پيريز مشرف (ar)	pakistanese president (nl - 239)	
Gen Musharraf (en)	pakistani president gen (en - 181)	
Pervez Múgarraf (tr)	presidente paquistaní (es - 149)	
پيريز مشرف (fa)	président (de - 627)	
Pervez Musharraf (de,sv)	president (de - 245)	
Gen Pervez Musharraf (en)	general (en - 457)	
General Musharraf (de,en)	president (esp - 598)	
Pervez Musharraf (de,en)	president gen (en - 92)	
Pervez Múgarraf (ru,uk)	президент (de - 143)	
پيريز مشرف (ar)	staatschef (de - 92)	
Pervez Músharraf (fr)	président (fr - 312)	
Pervez Músharraf (it)	presidente pakistano (it - 42)	
Pervez Musharraf (en,de)	begleiter (tr - 231)	
Pervez Músharraf (pl)	militärmaschinist (de - 47)	



Figure: Part of the multilingual historical information gathered by EMM-NewsExplorer about over 1 million entities.

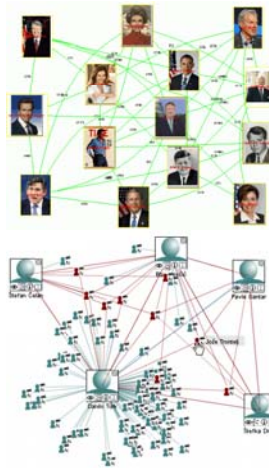


Figure: 2 different social networks generated from information found in multilingual news.

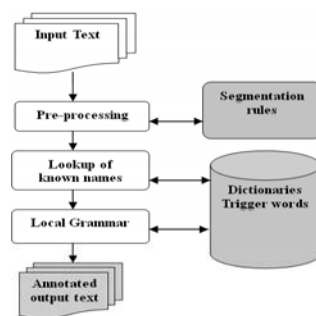


Figure: Architecture of the system.

Major challenge for adapting EMM rules to Arabic

- Many language-independent NER rules rely on case information, but **Arabic does not distinguish upper and lower case**, making it difficult to identify the name boundaries.
- Arabic commonly uses prefixes and suffixes.
- Some titles are also common first names, e.g. سلطان (Sultan), امير (Prince).

How we solved these issues

- Language-specific Arabic NER rules were added to the language-specific parameter file.
 - Longer lists of ~20,000 known name parts were used.
 - Name stop words** help detect the end of the name.
 - Light-weight morphological processing strips prefixes and suffixes from the word stems before applying the grammar.
- والرئيس /walirra'iis/ (and for the president).
 للرئيس /lirra'iis/ (for the president).
 الرئيس /ra'iis/ (president).
- Online demo**: Select Arabic on the page: <http://emm.newsexplorer.eu/>

Arabic-specific NER rules

- KNOWN_NAME+**b**(w+)**b**NAME_INFIX***b** KNOWN_NAME
- (w+)**b**NAME_INFIX+**b**(w+)
محمد علي بن حليلة (Mohammed ali ben Halima)
- PERSON_TRIGGER+**b**(w+)**b**KNOWN_NAME
السيد عيسى أحمد (Mr. Issa Ahmed – Ahmed is known)
- NAME_STOP_WORDS**b**(w+)(**b**MODIFIER)***b** PERSON_TRIGGER+
وقال حامد كرزاي الرئيس الأفغاني المنتخب الجديد (And said Hamid Karzai, the newly elected Afghan president)

Evaluation

- Test corpus: 35 manually tagged online news articles (34,000 tokens) from the newspapers *Assabah* (Tunisia) and *Alanwar* (Lebanon).

Category	Number	Precision	Recall	F-measure
Person	804	87 %	66.54 %	75.40 %
Organization	514	69.96 %	35.79	47.35 %
Location	433	91.52 %	74.82 %	82.33 %
Date and time	54	96.13 %	94.11 %	95.10 %
Numeric expression	46	93.29 %	89.47 %	91.34
Overall	1851	87,17 %	65,74 %	74,95 %

Table: Results obtained for the various NE types.

References

- Zaghouni W., Pouliquen B., Ebrahim M. & Steinberger R. (2010). *Adapting a resource-light highly multilingual Named Entity Recognition System to Arabic*. Proceedings of LREC, Valletta, Malta.
- Steinberger R., Pouliquen B. & Van der Goot E. (2009). *An Introduction to the Europe Media Monitor Family of Applications*. In Gey F., Kando N. & Karlgren J. (eds.): *Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIP 2009)*. Boston, USA.
- Steinberger R., Pouliquen B. and Ignat C. (2008). *Using language-independent rules to achieve high multilinguality in Text Mining*. In Fogelman-Soulé F., Perrotta D., Piskorski J. & Steinberger R. (eds.), *Mining Massive Data Sets for Security*. Amsterdam, The Netherlands, IOS Press.