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

Wajdi Zaghouani, Bruno Pouliquen, Mohamed Ebrahim & Ralf Steinberger

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.newsexplorer.eu/
- 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, …;
  - 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 tags, large-scale dictionaries.
- Sample NER rules:
  1) PERSON_TRIGGER=bUPPERCASEWord
     Swiss world champion Roger Federer
  2) UPPERCASEWord=bUPPERCASEWord
     bMODIFIER=bPERSON_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.

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. معاذ (Mu’adh), خالد (Khalid).

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 part 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.

Online demo: Select Arabic on the page: http://emm.newsexplorer.eu/

Arabic-specific NER rules

(3) KNOWN_NAME=b(w)+bNAME_INFIX* bKNOWN_NAME
(4) (w)+bNAME_INFIX* b(w)+

(5) PERSON_TRIGGER=b(w)+bKNOWN_NAME
(6) NAME_STOP_WORDS=b(w)+bMODIFIER+bPERSON_TRIGGER+

Evaluation

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

Table: Results obtained for the various NE types.

| Category       | Number | Precision | Recall | F-measure |
|----------------|--------|-----------|--------|-----------|
| Person         | 801    | 87%       | 66.54% | 75.40%    |
| Organization   | 514    | 69.96%    | 35.79% | 47.35%    |
| Location       | 492    | 91.52%    | 74.82% | 82.33%    |
| Date and time  | 54     | 96.12%    | 84.11% | 90.10%    |
| Numerical      | 46     | 93.29%    | 89.47% | 91.34%    |
| Overall        | 1581   | 87.51%    | 65.74% | 74.95%    |

Figure: Architecture of the system.

References

- Zaghouani W., Pouliquen B., Ebrahim M. & Steinberger R. (2010). Adapting a resource-light highly multilingual Named Entity Recognition system to Arabic, Proceeding of TALN, Valletta, Malta.
- Steinberger R., Pouliquen B. & van der Baart B. (2009). An Introduction to the Europe Media Monitor Family of Applications. In: Dey, P., Dillenbourg, P. & Ehrmann, M. (Eds.). Proceedings of a Multi-lingual Web - Proceeding of a Multi-lingual Web - Proceeding of a Multi-lingual Web. IOS Press.
- Steinberger R., Ebrahim M. & van der Baart B. (2008). Using language-independent rules to achieve high quality Arabic entity recognition. In: proceedings of the NATEXIS Conference, 20-21 February, Padova, Italy.
- Steinberger R. & Steinberger S. (2008). Using language-independent rules to achieve high quality Arabic entity recognition. In: Proceedings of the 5th Conference on Arabic Language Information Processing, Algiers, Algeria.