An Automatically Built Named Entity Lexicon for Arabic

M. Attia*, A. Toral*, L. Tounsi*, M. Monachini^, J.v. Genabith*

*Dublin City University (Ireland)
^Istituto di Linguistica Computazionale - CNR, Pisa (Italy)
Contents

• Introduction
  – NLP acquisition bottleneck, MINELex
• Methodology
  – Mapping, Extraction, Identification, Diacritisation, ...
• Results
• Conclusions
NLP apps make extensive use of LRs
Big effort during last 15 years to build resources
  – e.g. lexica: WordNet, EuroWordNet, SIMPLE, etc.

Enough coverage?
  – ~OK → verbs, adjs, advs, common nouns
  – ¬OK → NEs, domain terms, multiwords

“humans cannot manually structure the available knowledge at the same pace as it becomes available” (Philpot 05)
  – Automatic procedures needed!
Intro

- Step forward → 3 ingredients
  - Web 2.0, LRs, interoperability
- MINELex: Multilingual, Interoperable NE Lexicon
  - Derived automatically from Wikipedia and LRs
  - General approach, applied to:
    - English WN: 975k NEs
    - Spanish WN: 137k NEs
    - Italian SIMPLE-CLIPS: 125k NEs
  - NEs linked to LRs and ontologies
  - Extrinsic eval, QA → 28% increment accuracy
- Is the approach applicable to other lang families?
  - Arabic (arWN, arWK)
Methodology

An Automatically Built Named Entity Lexicon for Arabic. Dublin City University, Istituto di Linguistica Computazionale
Methodology: Mapping

- Identify senses of arWN that can be extended with NEs, i.e. instantiable nouns
- arWN (and enWN) do not have this info but have instance_of relations, i.e. instantiated nouns
  - country1 has_instance Malta
- Union of instantiated nouns from both resources
  - A: arWN i.n. + recursive hyponyms → 384
  - B: enWN i.n. + recursive hyponyms → mapping arWN + recursive hyponyms → 1,475
  - Final set: A U B → 1,572 senses, 1,187 nouns
- Lemma matching: i.n. ↔ arWK cats
  - 40.6%
Methodology: Extraction

• Extract articles from mapped categories
• ...and hyponym subcategories → pattern:
  – ^category_
    • From “سياسيون” (politicians)
      – “سياسيون_حسب_الحزب” (politicians by nationality)
      – “سياسيون_بريطانيون” (British politicians)

• Discard administrative categories
Methodology: NE Identification

- Original approach relied on capitalisation norms
  - Look for occurrences of title in body, check percentage it occurs with lowercase vs. uppercase
- … but Arabic does not follow them → exploit inter-lingual links to obtain equivalent article in 10 langs that follow cap. norms (en, es, fr, it, …)
  - Drawback: covers only 62.5% of articles
- Further heuristics to improve recall
  - Keywords from abstracts
    - LOC (16): abstract begins with “city”, “country”, etc
    - PER (60 + exclusion list 160): abstract contains “born in”, “studied in”, etc
  - Geonames: lexicon of geographic NEs
Methodology: Postprocessing

- Cross-fertilisation
  - Further ar NEs can be obtained by exploiting
    - Links between en, es, it NEs and their LRs
    - Interconnections among these LRs
  - E.g. NE extracted for es has equivalent in arWK but has not been extracted
    - Extract and connect to arWN following mapping esWN → enWN → arWN
Diacritisation

- Diacritics: Short marks above or under letters
  - الإمارات العربية المتحدة / al-imaratu al-arabiyyatu al-muttahidatu / “United Arab Emirates”
- Why needed? Speech, Syntactic disambiguation, WordNet
- Approach for restoring diacritics:
  - Checking available diacritised lists
  - Using a diacritisation tool
  - Using heuristics
Diacritisation

- Diacritised lists: geonames.de, geonames.org
  - 3,5k NEs matched (10%)
- Diacritisation tool: MADA
  - 29% coverage, mainly due to OOV (NEs)
- Using heuristics
  - Most unknown words are foreign names
  - Transliteration of foreign names usually employs long vowels
  - Native Arabic names do not follow this assumption and must be excluded
  - 59% coverage
- Combination: 73% coverage
Evaluation

• Data used
  – arWN (connected to enWN 2.0)
  – enWN 2.1
  – Automatic mapping enWN 2.1 ↔ enWN 2.0
  – arWK dump Feb 2010. 234k articles, 33k categories

• Test set
  – 1k arWK articles that belong to the categories mapped
  – Annotated as [NE, not NE]

• Measures: P, R, F1, F0.5
## Evaluation: NE identification

<table>
<thead>
<tr>
<th>Heur.</th>
<th>Threshold</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>F0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>0.91</td>
<td><strong>99.25</strong></td>
<td>42.39</td>
<td>59.40</td>
<td>78.25</td>
</tr>
<tr>
<td></td>
<td>0.41</td>
<td>98.33</td>
<td>50.16</td>
<td>66.43</td>
<td><strong>82.49</strong></td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>94.70</td>
<td><strong>51.33</strong></td>
<td><strong>66.57</strong></td>
<td>81.01</td>
</tr>
<tr>
<td>yes</td>
<td>0.91</td>
<td><strong>99.28</strong></td>
<td>58.68</td>
<td>73.76</td>
<td>87.21</td>
</tr>
<tr>
<td></td>
<td>0.41</td>
<td>98.55</td>
<td>65.07</td>
<td><strong>78.38</strong></td>
<td><strong>89.35</strong></td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>95.83</td>
<td><strong>66.13</strong></td>
<td>78.26</td>
<td>87.94</td>
</tr>
</tbody>
</table>
Evaluation: NE extraction

<table>
<thead>
<tr>
<th>Heur.</th>
<th>Threshold</th>
<th>NEs</th>
<th>Relations</th>
<th>Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>0.91</td>
<td>23,910</td>
<td>27,422</td>
<td>24,887</td>
</tr>
<tr>
<td></td>
<td>0.41</td>
<td>28,048</td>
<td>32,287</td>
<td>29,451</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>30,354</td>
<td>34,901</td>
<td>32,205</td>
</tr>
<tr>
<td>yes</td>
<td>0.91</td>
<td>31,284</td>
<td>36,271</td>
<td>32,386</td>
</tr>
<tr>
<td></td>
<td>0.41</td>
<td>35,423</td>
<td>41,136</td>
<td>36,940</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>37,729</td>
<td>43,750</td>
<td>39,693</td>
</tr>
</tbody>
</table>

Postprocessing:
- 11.7k en, 6.8k it, 6.9k es NEs have ar equivalent
- Discard duplicates + NEs extracted for ar → 6.5k NEs
- Added to MINELex → contains 44k ar NEs
### Output Example

#### FormRepresentation

<table>
<thead>
<tr>
<th>LE id</th>
<th>written form</th>
<th>v. type</th>
<th>script</th>
<th>orthog. n.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ar.le.المملكة</td>
<td>الأمة المتحدة</td>
<td>full</td>
<td>Arab</td>
<td>arabicUnpointed</td>
</tr>
<tr>
<td>ar.le.المملكة</td>
<td>الأمة المتحدة</td>
<td>full</td>
<td>Arab</td>
<td>arabicPointed</td>
</tr>
<tr>
<td>ar.le.nnZm</td>
<td>الأمة المتحدة</td>
<td>full</td>
<td>Latin</td>
<td>arabicPointed</td>
</tr>
<tr>
<td>en.le.United_Nations</td>
<td>en.United_Nations</td>
<td>full</td>
<td>Latin</td>
<td>arabicPointed</td>
</tr>
</tbody>
</table>

#### Sense

<table>
<thead>
<tr>
<th>S id</th>
<th>LE id</th>
<th>res.</th>
<th>res. id</th>
</tr>
</thead>
<tbody>
<tr>
<td>الأمة المتحدة</td>
<td>ar.le.المملكة</td>
<td>ar.WK</td>
<td>2270</td>
</tr>
<tr>
<td>الأمة المتحدة</td>
<td>ar.le.nnZm</td>
<td>ar.WN</td>
<td>109710501</td>
</tr>
<tr>
<td>الأمة المتحدة</td>
<td>en.le.United_Nations</td>
<td>en.WK</td>
<td>31769</td>
</tr>
<tr>
<td>en.s.United_Nations</td>
<td>en.United_Nations</td>
<td>en.WK</td>
<td>31769</td>
</tr>
</tbody>
</table>

#### SenseAxis

<table>
<thead>
<tr>
<th>SA id</th>
<th>element</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ar.s.المملكة</td>
</tr>
<tr>
<td>1</td>
<td>en.s.United_Nations</td>
</tr>
</tbody>
</table>

#### SenseAxisExternalRef

<table>
<thead>
<tr>
<th>SA id</th>
<th>resource</th>
<th>resource id</th>
<th>relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SUMO</td>
<td>PoliticalOrganization</td>
<td>at</td>
</tr>
</tbody>
</table>

#### SenseRelation

<table>
<thead>
<tr>
<th>source id</th>
<th>target id</th>
<th>relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>الأمة المتحدة</td>
<td>ar.s.109710501</td>
<td>instanceof</td>
</tr>
</tbody>
</table>

#### Confidence (NE id)

<table>
<thead>
<tr>
<th>S id</th>
<th>mode</th>
<th>occurrences</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ar.s.المملكة</td>
<td>wiki10</td>
<td>250</td>
<td>0.996</td>
</tr>
</tbody>
</table>
Conclusions

- Adapted and extended generic methodology to build a NE lexicon to Arabic: arWN and arWK
- Challenges: NE identification and diacritisation
- Result: 44k NE lex
  - Connected to
    - Intralingual: arWN synsets
    - Interlingual: equivalent NEs in en, es, it + ontologies
  - Can be used with different levels of granularity
  - Compliant with ISO LMF format
- Available at
  - www.ilc.cnr.it/ne-repository
Thank you very much!

Questions?