Building an Arabic Multiword Expressions (MWE) Repository

Abdela Hawwari, Kfir Bar, Mona Diab
Multiword Expression

- kick the bucket
- spill the beans
- make a decision
- New York
- rain check
- daddy soda
- going screensaver...
- wall street
Compositionality

prime minister
make a decision
... spy the beans
kick the bucket

Compositional

Transperent

Non compositional

Idiomatic
Goal

• To create a repository of Arabic MWEs, annotated with some morphological features
• To enable additional research in the field of Arabic processing involving MWEs
• Highly inflected (Semitic language)
• Words are based on root and pattern and inflected for person, number and gender
• Nouns use suffixes to reflect possessive forms
• Verbs use suffixes to reflect direct objects
• Agreement

>gmDt EynyhA
she closed her eyes
MWE Classes

- Based on syntactic features
  - Verb Noun Construction (VNC)
    - make a decision
    - md aljswr (someone build bridges... as in extending the arms for peace...)
  - Noun Noun Construction (NNC)
    - traffic light
    - Enq {lzjAjp (bottleneck)
  - Verb Particle Construction (VPC)
    - take over
    - mDY fy (continues working on...)
  - Adjective Noun Construction (ANC)
    - the white house
    - >xZ w>ETY (give and take)
Arabic MWE Repository

- Manually collected from dictionaries

- A Dictionary of Arabic Contemporary Idioms
  Fayed, Wafaa Kamel

- A contextual Dictionary of Idioms
  Sieny, Mahmoud Esmail, Mokhtar A. Hussein and Sayyed A. Al-Doush

- A Dictionary of Arabic Contemporary Idioms
  Dawood, Mohammed

- A Dictionary of Arabic Idiomatic Expressions
  Abou Saad, Ahmed
Arabic MWE Repository

• Each MWE was manually augmented with the following information:
  – The correct SAMA [Maamouri, 2010] morphological analysis for every word, since MADA [Roth et al., 2008] didn’t perform well due to short contexts
  – MWE class
MWE Class

- Manually assigned

<table>
<thead>
<tr>
<th>MWE Class</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>VNC</td>
<td>1974</td>
</tr>
<tr>
<td>VPC</td>
<td>670</td>
</tr>
<tr>
<td>NNC</td>
<td>1239</td>
</tr>
<tr>
<td>ANC</td>
<td>285</td>
</tr>
<tr>
<td>VVC</td>
<td>41</td>
</tr>
</tbody>
</table>

4,209
Generic Words

• Placeholders for semantically-related words
  – $fIAn$ – “so and so” – a person
    qr $[fIAn]$ $EynA$ – “pleased someone”
  – $k^*A$ – “something” – an object
    $ElY$ $HsAb$ $[k^*A]$ – “at the expense of that”
  – $<mr$ – “something” – an issue
    $[<mr]$ $Abn$ $ywmh$ - “something very new“

• Generic words are provided with their context-sensitive morphological analysis
Corpus Annotation

• **Preprocessing:**

  AMIRAN [*Diab et al. – to appear*] is a tool for finding context-sensitive morpho-syntactic information

  – Results contain for each word:

    • Clitics / Segmentation
    • Diacritized lemma
    • Stem
    • Full part-of-speech tag
    • Base-phrase tag
    • Named-entity-recognition (NER) tag
Pattern Matching

• Deterministically identifying MWE instances in a running text
  – Considering **morphological variations** of words – matching on the lemma level
  – Matching with generic words
    • \([k^*A]\) is currently matched with noun-phrases
    • \([f\ddot{I}An]\) is matched with noun-phrases and person entities
Gaps

• Additional words, such as modifiers, that are not part of the original MWE words are considered as a gap

wDEt AlHrb <wzArhA
“the war is over”

wDEt AlHrb AIEAlmyp AlvAnyp <wzArhA
“the second world war is over”
Gaps

• Allowing NP chunks to be considered as gaps

\[wDEt\ AlHrb <wzArhA\]

“the war is over”

\[wDEt\ AlHrb \text{ AIEAAlmyp AlvAnyp} <wzArhA\]

“the second world war is over”
Corpus

- Arabic Gigaword 4.0
- 250M tokens were processed
- Found ~480K MWE instances of 2031 MWEs

<table>
<thead>
<tr>
<th>MWE Class</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>VNC</td>
<td>64,504</td>
</tr>
<tr>
<td>VPC</td>
<td>75,844</td>
</tr>
<tr>
<td>NNC</td>
<td>316,393</td>
</tr>
<tr>
<td>ANC</td>
<td>23,814</td>
</tr>
</tbody>
</table>
Corpus Evaluation

• We sampled the corpus and evaluated the results
• We checked whether an instance represents the MWE in that context
  
  For example –

  \texttt{<TiAq AlnAr} (“opening fire” and not just “lighting a fire”)
Evaluation Results

- Evaluation results

<table>
<thead>
<tr>
<th>MWE Class</th>
<th>Number</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>VNC</td>
<td>157 (34 MWEs)</td>
<td>154</td>
</tr>
<tr>
<td>VPC</td>
<td>161 (32 MWEs)</td>
<td>125</td>
</tr>
<tr>
<td>NNC</td>
<td>155 (26 MWEs)</td>
<td>154</td>
</tr>
</tbody>
</table>
Summary

• Arabic MWE Repository is introduced
• It has 4,209 MWEs from 4 different classes
• The Arabic Gigaword 4.0 was deterministically annotated with the MWEs and the results were manually evaluated
• We believe this resource will enable new directions for researches on Arabic processing
Thank you

The repository is available for download—please contact us to get a copy

barkfir@post.tau.ac.il