Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic

Mohammed Attia, Lamia Tounsi, Josef van Genabith
Outline

• Introduction
• Principles of Lexicography
• Modern Standard Arabic vs. Classical Arabic
• Review of Arabic lexicographic work
• Project Description
• Lexical Markup Framework (LMF)
• Automatic Acquisition Architecture
• Results and evaluation
• Conclusion
Introduction

• Importance of lexical resources for NLP tasks
  – The backbone of morphological analysers
  – Principal factor for coverage
  – Unknown words in parsing cause a problem and we want to minimize them as much as possible

• Advantages of automatic acquisition vs. Manual construction of dictionaries
  – Time and effort
  – Speed and efficiency
  – Consistency and quality
  – It is impractical to manually analyse large and ever-growing amounts of data

• How the resource will fit in the annotation/parsing tools
  – DCU Parser
  – DCU annotation tools
Principles of Lexicography

Definition of a dictionary
- A description of the vocabulary used by members of a speech community. A dictionary deals with:
  - conventions not idiosyncrasies
  - norms not rarities
  - probable not possible

- Lexical evidence
  - Subjective evidence
    - introspection
    - informant-testing
  - Objective evidence
    - A corpus provides typifications of the language
      - A typical lexical entry means it is both “frequent” or “recurrent” and “well-dispersed” in a corpus.
      - A typical lexical entry belongs to the stable “core” of the language.

Principles of Lexicography

- **Corpora and Dictionaries**
  - Brown Corpus, 1 million words, 1960s,
    → Citations for *American Heritage Dictionary*
  - Birmingham corpus, 20 million words, 1980s
    → Cobuild English Dictionary.
  - British National Corpus (BNC), 100 million words, 1990s set the standard (balance, encoding)
  - The Oxford English Corpus, one billion words, 2000s
    → Oxford English Dictionary
  - Longman Corpus Network, 330 million word
    → Longman Dictionaries
Principles of Lexicography

- **Dictionaries before Corpora**
  - Citation banks
    - A citation is a short extract providing evidence for a word usage or meaning in authentic use.
  - Disadvantages
    - labour-intensive
    - instances of usage are authentic, but there is a big subjective element in their selection.
      - People tend to notice what is remarkable and ignore what is typical
      - bias towards the novel or idiosyncratic usages
Principles of Lexicography

• Characteristics of a reliable corpus
  – The corpus does not favour high class language
  – The Corpus should be large and diverse
  – The corpus should be either synchronic or diachronic
  – The corpus should be well-balanced using “stratified sampling”
  – The corpus should avoid skewing
Principles of Lexicography

• Lexical Profiling
  – Word POS
    v, n, adj, adv, conj, det, interj, prep, pron
  – Valency Information
  – Collocations
    commit a crime, sky blue, lame duck
  – Colligational preferences
    was acquitted, trials (difficult experiences)
Principles of Lexicography

- Lexical Profiling Software
- Concordancers
- Word Sketch (Sketch Engine) - Adam Kilgarriff
Concordancer
Sketch Engine

<table>
<thead>
<tr>
<th>object_of</th>
<th>count</th>
<th>frequency</th>
<th>a_modifier</th>
<th>count</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>strike</td>
<td>61</td>
<td>43.38</td>
<td>hard</td>
<td>23</td>
<td>25.99</td>
</tr>
<tr>
<td>drive</td>
<td>26</td>
<td>27.56</td>
<td>real</td>
<td>20</td>
<td>23.43</td>
</tr>
<tr>
<td>get</td>
<td>27</td>
<td>16.38</td>
<td>best</td>
<td>14</td>
<td>19.31</td>
</tr>
<tr>
<td>seal</td>
<td>5</td>
<td>14.82</td>
<td>good</td>
<td>19</td>
<td>18.01</td>
</tr>
<tr>
<td>make</td>
<td>26</td>
<td>13.6</td>
<td>bad</td>
<td>8</td>
<td>15.31</td>
</tr>
<tr>
<td>find</td>
<td>8</td>
<td>7.81</td>
<td>better</td>
<td>8</td>
<td>14.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>modifies</th>
<th>count</th>
<th>frequency</th>
<th>n_modifier</th>
<th>count</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>basement</td>
<td>22</td>
<td>38.62</td>
<td>plea</td>
<td>26</td>
<td>40.62</td>
</tr>
<tr>
<td>hunter</td>
<td>22</td>
<td>37.23</td>
<td>wage</td>
<td>6</td>
<td>16.8</td>
</tr>
<tr>
<td>price</td>
<td>54</td>
<td>33.65</td>
<td>credit</td>
<td>6</td>
<td>14.68</td>
</tr>
<tr>
<td>bookshop</td>
<td>11</td>
<td>26.73</td>
<td>sale</td>
<td>5</td>
<td>10.47</td>
</tr>
</tbody>
</table>

Part of the Word Sketch for the noun *bargain*
Modern Standard Arabic vs. Classical Arabic vs. Colloquial Arabic

• Modern Standard Arabic
  – The language of modern writing, prepared speeches and the language of the news

• Classical Arabic
  – The language of Arabia before Islam and after Islam until the Medieval Times
  – Present religious teaching, poetry and scholarly literature.

• Colloquial Arabic
  – Variety of Arabic spoken regionally and which differs from one country or area to another. They are to a certain extent mutually intelligible.

Code Shifting – Diglossia – multi-layered diglossia
Modern Standard Arabic vs. Classical Arabic vs. Colloquial Arabic

• Modern Standard Arabic
  – Tendency for simplification
    • Some CA structures to die out
    • Structures marginal in CA started to have more salience
    • no strict abidance by case ending rules
  – A subset of the full range of structures, inflections and derivations available in CA
  – MSA conforms to the general rules of CA
  – How “big” or how “small” the difference (on morphological, lexical or syntactic levels) need more research and investigation
Review of Arabic lexicographic work

- *Kitab al-'Ain* by al-Khalil bin Ahmed al-Farahidi (died 789) (refinement/expansion/organizational Improvement)
- *Tahzib al-Lughah* by Abu Mansour al-Azhari (died 980)
- *al-Muheet* by al-Sahib bin 'Abbad (died 995)
- *Lisan al-'Arab* by ibn Manzour (died 1311)
- *al-Qamous al-Muheet* by al-Fairouzabadi (died 1414)
- *Taj al-Arous* by Muhammad Murtada al-Zabidi (died 1791)
- *Muheet al-Muheet* (1869) by Butrus al-Bustani
- *al-Mu'jam al-Waseet* (1960)
Review of Arabic lexicographic work

• Bilingual Dictionaries
  – Edward William Lane's *Arabic-English Lexicon* (1876) indebted to *Taj al-Arous* by al-Zabidi
  – Hans Wehr's Dictionary of Modern Written Arabic (1961)
    • Size: 45,000 entries
    • Aim: Using scientific descriptive principles to describe present-day vocabulary through wide reading in literature of every kind
  • Application
    – Selection of works by high flying poets and literary critics such as Taha Husain, Taufiq al-Hakim, Mahmoud Taimur, al-Manfalauti, Jubran Khalil Jubran
    – Use of secondary sources (dictionaries) for expansion
    – Inclusion of rarities and classisms that no longer formed a part of the living lexicon
Review of Arabic lexicographic work

• Bilingual Dictionaries
  – Landau and Brill (1959) *A Word Count of Modern Arabic Prose*
    • A word count based on 270,000 words based on the news and 60 contemporary books on: fiction, literary criticism, history, biography, political science, religion, social studies, economics, travels and historical novels
    • 6,000 words in the news
    • 11,000 words in literature
    • 12,400 words in the combined list (does not include proper nouns)
Review of Arabic lexicographic work

• Bilingual Dictionaries
    • COBUILD-style, Corpus-based (3 million words)
    • Manually constructed
    • Covers the whole range of the actual vocabulary in the corpus with 17,000 entries compared to 45,000 entries in Hans Wehr
    • 5% of frequent new words not found in Hans Wehr
Review of Arabic lexicographic work

- Bilingual Dictionaries
  - Buckwalter Arabic Morphological Analyzer (2002)
    - Size: 40,222 lemmas (including 2,034 proper nouns)
    - Includes many obsolete lexical items
      (But how many?)

<table>
<thead>
<tr>
<th>#</th>
<th>Meaning</th>
<th>Classical Word</th>
<th>Google</th>
<th>MSA Word</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sully</td>
<td>قلعط</td>
<td>8</td>
<td>لطخ</td>
<td>29,600</td>
</tr>
<tr>
<td>2</td>
<td>caulk</td>
<td>قلفط</td>
<td>9</td>
<td>أفسد</td>
<td>205,000</td>
</tr>
<tr>
<td>3</td>
<td>wear</td>
<td>استكد</td>
<td>4</td>
<td>أنهك</td>
<td>37,100</td>
</tr>
<tr>
<td>4</td>
<td>fickle</td>
<td>غملج</td>
<td>7</td>
<td>متقلب</td>
<td>189,000</td>
</tr>
<tr>
<td>5</td>
<td>erosion</td>
<td>انتكال</td>
<td>7</td>
<td>تأكل</td>
<td>1,700,000</td>
</tr>
</tbody>
</table>

*Google score for Classical vs. MSA entries*
Review of Arabic lexicographic work

• Bilingual Dictionaries
  – Buckwalter Arabic Morphological Analyzer (2002)
    • Searching for Buckwalter on Aljazeera (40,205 reduced to 31,359 after removing diacritics)
    • Both false positives and false negatives are possible but the figures are still indicative

<table>
<thead>
<tr>
<th>Frequency Range</th>
<th>0</th>
<th>1-100</th>
<th>101-1000</th>
<th>Over 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Occurrences</td>
<td>7312</td>
<td>13563</td>
<td>6606</td>
<td>3878</td>
</tr>
<tr>
<td>Per Cent</td>
<td>23.31%</td>
<td>43.25%</td>
<td>21.06%</td>
<td>12.36%</td>
</tr>
</tbody>
</table>
Project Description

• Acquisition of Arabic lexical resources
  – from corpora
  – modern language
  – varied domains
  – inducing the lexical profile for each lemma (frequency, inflections, derivations, citation, etc.)

• Production of new lexical sets
  – accumulated in a MySQL database
  – meeting Lexical Markup Framework specifications
  – exported into Lexical Markup Framework format
Project Description

• How our lexical database will be different from Buckwalter’s. We include
  – only entries attested in a corpus
  – subcategorization frames
  – +/-human semantic information for nouns
  – detailed information about derived nouns/adjectives
    (active or passive participle or a verbal noun, *masdar*)
  – multi-word expressions
  – classification of proper nouns: person, place, organization, etc.
Lexical Markup Framework (LMF)

ISO 24613:2008

• Aims:
  – Managing lexical resources
  – Providing a metamodel (or a super-hierarchy) to accommodate lexical information at all levels
  – Provides specifications, encoding format and naming conventions to ensure consistency
  – Enable the merger of individual electronic lexical resources
  – Allows instantiation of monolingual, bilingual or multilingual lexical resources
  – Allows work at a small scale or large scale
  – Tries to cover all natural languages (including languages with rich and complex morphology such as Arabic)
Lexical Markup Framework (LMF)

ISO 24613:2008

• History:
  – It started in 2003
  – Earlier lexicon standardization projects include GENELEX, EDR, EAGLES, MULTEXT, PAROLE, SIMPLE and ISLE
  – Project team:
    • Nicoletta Calzolari (Italy)
    • Gil Francopoulo (France)
    • Monte George (US)
    • + A panel of 60 experts
  – Published officially as an International Standard in 2008
  – Uses Unified Modeling Language (UML)
  – LMF is considered the state of the art in NLP lexicon management field
Lexical Markup Framework (LMF)

**Architecture**
1. Morphology extension
3. NLP syntax extension
4. NLP semantics extension
5. NLP multiword expression patterns extension
Lexical Markup Framework (LMF)

Architecture

Global Information → Lexical Resource → Lexicon → Lexical Entry → Sense → Link to Arabic WordNet Synsets

- Form (Morphology)
- Syntactic Behaviour (Subcat Frames)
- Multiword Expressions
Lexical Markup Framework (LMF)

Simple Example: Noun
Lexical Markup Framework (LMF)

Example: Verb
Lexical Markup Framework (LMF)

Example: Arabic Root Management
Lexical Markup Framework (LMF)

**Sources of Lexical Information**

- **Morphological information (ATB, Buckwalter, FST):**
  - word root, lemma, form, diacritics, frequency, citations. This information will be extracted from the Arabic Treebank
- **Syntactic information:** Subcategorization frames (Arabic Annotation Algorithm)
- **Semantic information:** linking to Arabic WordNet.
- **Dictionary information:** translation in English (Buckwalter, Online Dictionaries, Landau’s Word Count)
- **Multi-word Expression (MWE) and named entity.** FST and Arabic Named Entity Lexicon extraction project
Automatic Acquisition Methodology

• Methodology
  – Starting with the annotated data to build a core lexicon
  – Moving toward un-annotated data for extension in domain and size

• Lemmatization Tools
  – Buckwalter
  – MADA-TOKAN
  – FST-Guesser
Automatic Acquisition Methodology

- Lemmatization is an essential prerequisite due to
  - derivational and inflectional nature of Arabic
  - lack of diacritics (vowel marks)
  - the employment of cliticization (affixation of function words to content words)
  - 2,552 well-formed forms for transitive verbs (شكر “shakar” (to thank))
  - 519 valid forms for regular nouns (معلم “mu’allim” (teacher))

Wa-sa-ya-shkur-una-hu
And-will-thank-they-him
And they will thank him.
Automatic Acquisition

• From annotated data: Arabic Treebank (ATB)
• Advantages
  – morphologically/syntactically annotated
  – modern texts
  – tokenized and diacritized,
  – manually reviewed by human annotators
• Disadvantages
  – Not large: Only ½ million words
  – Not diverse: taken from the newswire
Automatic Acquisition

• From annotated data: Arabic Treebank (ATB)

Results

<table>
<thead>
<tr>
<th></th>
<th>Types</th>
<th>Unique Lemmas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>41,183</td>
<td>7,184</td>
</tr>
<tr>
<td>Adjectives</td>
<td>14,044</td>
<td>2,540</td>
</tr>
<tr>
<td>Verbs</td>
<td>17,888</td>
<td>2,315</td>
</tr>
<tr>
<td>Total</td>
<td>73,115</td>
<td>12,039</td>
</tr>
</tbody>
</table>
Testing and Evaluation

From annotated data: Arabic Treebank (ATB)
- Testing on Aljazeera Search Engine, Why?
  - The web is polluted with noisy data
  - The type of application (lexicon) has a narrow threshold for noise
  - Technically API does not allow more than 1000 searches per day

<table>
<thead>
<tr>
<th>Misspellings</th>
<th>Google Score</th>
<th>CNN Score</th>
<th>Right Form</th>
<th>Google Score</th>
<th>CNN Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>arround</td>
<td>1,200,000</td>
<td>3</td>
<td>around</td>
<td>780,000,000</td>
<td>44,555</td>
</tr>
<tr>
<td>vedio</td>
<td>4,450,000</td>
<td>0</td>
<td>video</td>
<td>2,590,000,000</td>
<td>131,845</td>
</tr>
<tr>
<td>resaercher</td>
<td>6,200</td>
<td>0</td>
<td>researcher</td>
<td>26,500,000</td>
<td>19,729</td>
</tr>
<tr>
<td>possebility</td>
<td>31,100</td>
<td>0</td>
<td>possibility</td>
<td>95,100,000</td>
<td>38,163</td>
</tr>
<tr>
<td>bilieve</td>
<td>29,200</td>
<td>0</td>
<td>believe</td>
<td>349,000,000</td>
<td>44,330</td>
</tr>
<tr>
<td>perfromance</td>
<td>195,000</td>
<td>0</td>
<td>performance</td>
<td>459,000,000</td>
<td>17,085</td>
</tr>
<tr>
<td>mesjudge</td>
<td>80</td>
<td>0</td>
<td>misjudge</td>
<td>278,000</td>
<td>196</td>
</tr>
<tr>
<td>gtfrde</td>
<td>1,750</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ghgh</td>
<td>233,000</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Testing and Evaluation

From annotated data: Arabic Treebank (ATB)
- Testing on Aljazeera Search Engine, Why?
  - Aljazeera is more than just news.
Testing and Evaluation

From annotated data: Arabic Treebank (ATB)

- Testing results
- No. of Lemmas from the ATB: 12340
- After removing diacritics: 10071
- No. of Not found on Al-Jazeera: 208 (2%)

Error analysis

- Mistagging in the annotation process (baloqA' should be bi-liqA')
- Buckwalter gives the wrong lemma it should be >abora$iy~+ap not >abora$iy~
- Not found in Al-Jazeera (>adokan nor dakonA')
Automatic Acquisition Architecture

From un-annotated data

• Advantages
  – Large:
    • CCA: ½ million
    • Wiki: 40 million
    • Gigaword: 200 million
  – Diverse: taken from the various domains

• Disadvantages
  – No morphological or syntactical annotation
Automatic Acquisition Architecture

From un-annotated data

• Results for Corpus of Contemporary Arabic (CCA)
  – Nominals: 240236
  – Unique nominals: 12502
  – Verbals: 67812
  – Unique verbals: 4245
  – Total: 16747
  – Not intersected with ATB: 6312
Automatic Acquisition Architecture

From un-annotated data

• Testing on Aljazeera Search Engine
  – New lexical items: 6312
  – lemmas not found: 543 (9%)
  – full forms: not found: 941 (15%)
  – Neither lemma nor form: 240 (4%)
Automatic Acquisition Architecture

From un-annotated data
- Dealing with the residue (Words not analysed by MADA)
  - 6106 items from the CCA had no analysis by MADA, potentially useful
  - Error Detection (Language dependent)
    - Not including numbers or non-Arabic letters
    - Not including (".َّ.+) taa marbouta in the middle
    - Not including (".ُ.+) Alif maqsoura in the middle in the middle
    - Not including (".ُّ.") two alifs anywhere
- Error Detection (Language independent)
- We merge MADA with our FST Guesser to filter through these words
Automatic Acquisition Architecture

MadaOutput: *1.000000 wAlmtswqyn=[wAlmtswqyn_0 POS:AJ Al+ +ACC w+ +DEF MOOD:NA +PL]=NO-ANALYSIS

Guesser output:

<table>
<thead>
<tr>
<th>الاسم باللغة العربية</th>
<th>الاسم باللغة الإنجليزية</th>
<th>ترتيب</th>
</tr>
</thead>
<tbody>
<tr>
<td>والمسؤونين</td>
<td>Guess+dual+acc+gen@</td>
<td>2</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>Guess+masc+pl+acc+gen@</td>
<td>3</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>Guess+sg@</td>
<td>1</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>Guess+dual+acc+gen@</td>
<td>0</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>Guess+masc+pl+acc+gen@</td>
<td>0</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>Guess+sg@</td>
<td>0</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+dual+acc+gen@</td>
<td>6</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+masc+pl+acc+gen@</td>
<td>7</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+sg@</td>
<td>5</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+dual+acc+gen@</td>
<td>0</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+masc+pl+acc+gen@</td>
<td>0</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+sg@</td>
<td>0</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+dual+acc+gen@</td>
<td>4</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+masc+pl+acc+gen@</td>
<td>5</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+sg@</td>
<td>3</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+dual+acc+gen@</td>
<td>0</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+masc+pl+acc+gen@</td>
<td>0</td>
</tr>
<tr>
<td>والمسؤونين</td>
<td>conj@+ال+المستوى+Guess+sg@</td>
<td>0</td>
</tr>
</tbody>
</table>
Automatic Acquisition Architecture

- GuessLemma: 7@متسوق@والمتسوقين@AJ
- GuessLemma: 6@متسوق@والمتسوقين@AJ
- GuessLemma: 5@متسوق@والمتسوقين@AJ
- GuessLemma: 5@المتسوق@والمتسوقين@AJ
- GuessLemma: 4@المتسوق@والمتسوقين@AJ
- GuessLemma: 3@المتسوق@والمتسوقين@AJ
- GuessLemma: 3@المتسوق@والمتسوقين@AJ
- GuessLemma: 2@المتسوق@والمتسوقين@AJ
- GuessLemma: 1@المتسوق@والمتسوقين@AJ
Automatic Acquisition Architecture

• Formula for giving weight to the guessing output:

Word Weight =

$$\frac{((\text{# of different forms} \times 2) + (\text{# of form repetitions} \times 1))}{2}$$

Testing the Formula (How useful in cascading good solutions up the list)

- 57% from the top are valid for inclusion in a dictionary as is
- 6% from the bottom are valid for inclusion in a dictionary as is
Automatic Acquisition Architecture

• Arabic Wikipedia:
  • First Portion 2 million words
    – MADA coverage is 96%.
    – 36,000 unique words not found by MADA
    – 22,164 verbs, nouns and adjectives are collected from the first portion
    – 10,712 were not found in the ATB
      7,763 Nominals not found in the ATB
      2,949 verbs not found in the ATB
Automatic Acquisition Architecture

From un-annotated data

• Results for Arabic Wikipedia (Full Corpus)
  – Nominals: 17076178
  – Unique nominals: 22969
  – Verbals: 2991974
  – Unique verbals: 8151
  – Total: 31120
  – No Analysis by MADA: 1,724,200
Automatic Acquisition Architecture

- Al-Jazeera Corpus
- Attia's Morphology
- DCU Arabic Named Entities
- Arabic Wikipedia
- Landau's word count
- Arabic WordNet
- MADA Lemmatizer
- CCA
- Wiki
- Gigaword
- Jazeera

- Arabic Annotation Algorithm
- Penn Arabic Treebank
- Annotated-Text Lemmatization
- Buckwalter Morphology
- Subcat Frames
Conclusion

• Size of acquired lexicon
  – 12,340 from ATB
    • 1,000 from Attia fst
    • 6,312 from CCA
    • 18,823 from Wiki
  – Buckwalter contains irrelevant lexical entries (at least 1/5 is outside of MSA)
    • 23% of the lemmas are not found in Al-Jazeera
    • 20% of the lemmas are not found in Arabic Wiki (40 million words)

• Significance of frequency information to calculate word weight as a method of validation
• Using search engine as a way to flag potentially problematic entries
• Improving error detection techniques