Phonetic Search of Speech

Dictionary Normalization and Query Expansion
Overview

- We seek to develop a deeper understanding of the parameters for effective **phonetic search**—the search for content in large audio archives—and apply it to the creation of query formation aids for users of speech search technology (specifically, the Nexidia speech search tool.)
Goals

To help Foreign Language Professionals:

- stay focused on the meaning and content of their queries
- frame valid queries (e.g., attested words or phrases of the dialect in question)
- frame more effective queries (e.g., queries with greater semantic coverage, and more compatible with the search tool’s capabilities)

By allowing them to:

- Access pronunciations through enhanced dictionary resources
- Access all word forms through automatic morphological query expansion
- Access pronunciation variants through phonological expansion
Language-Specific Challenges: Arabic

• It is written **without vowels**.

• It has a complex **morphology**. Every word stem generates many forms.

• Even the consonants vary across **dialects**.
Language-Specific Challenges: Arabic

- It is written **without vowels**.

- It has a complex **morphology**. Every word stem generates many forms.

- Even the consonants vary across **dialects**.

- قنبلة qunbulah “bomb”
  - القنبلة Al-qunbulah “the-bomb”
  - قنابل qanaAbil “bombs”
  - القنابل Al-qanaAbil “the-bombs”
  - قنبلتي qunbulatiy “my-bomb”
  - قنابلتي qanaAbiliy “my-bombs”
  - قنابلتك qunbulatak “your-bomb (to a man)”
  - قنابلك qanaAbilak “your-bombs (to a man)”
  - قنابلتك qunbulatik “your-bomb (to a woman)”
  - قنابلك qanaAbilik “your-bombs (to a woman)”
  - قنبلتكم qunbulatkum “your-( plural)- bomb”
  - قنابلتكم qanaAbilkum “your-( plural)- bombs”
  - قنبليلة qunaybila “lil’ bomblet”…
  - many, many others
Language-Specific Challenges: Arabic

• It is written without vowels.

• It has a complex morphology. Every word stem generates many forms.

• Even the consonants vary across dialects.

• قنبلة, قمبلة qnblh, qmblh (Educated Urban)
  • ق ق
  • گمبلا gmbla (Iraq)
    • گ گ
  • غنبله, غمبله ghnblh, ghmblh (Sudan)
    • غ غ
  • أنبلة, أمبلة 'nblh, 'mblh (Cairo)
    • أ أ
Dictionary Normalization

• Arabic colloquial dialects have traditionally had no stable “standard” written form.
  – *De facto* standards are now emerging have largely not been adopted by available lexica.
  – The Nexidia pronunciation lexicon is not a reliable substitute for a robust dictionary.

• By adding information about pronunciation to a stand-alone dictionary, we have a chance to impose good lexicographic standards, making the tool more transparent.
CASL’s Arabic lexicographer has been tracking standard colloquial spellings as they emerge and proposes a standardization of Arabic dictionaries that will:

- Create a protocol for standard, unified dictionary entries that will allow users to search for a term by multiple input methods and get back results for all of the term’s common phonological variants in the representation used by Nexidia.
- Provide a mechanism (similar to Google’s “Did you mean…?” spelling help) whereby users who are unsure of the proper spelling for a search term can find it through approximate spellings.
Dictionary Normalization - Examples

- There are often many different ways to spell a word in Arabic- and they are *all* right

*How do you write ‘dog’ in Arabic?*

<table>
<thead>
<tr>
<th></th>
<th>SATTS</th>
<th>Native Unvocalized</th>
<th>Native Vocalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSA</td>
<td>KLB</td>
<td>كلب</td>
<td>كَلْبْ</td>
</tr>
<tr>
<td>Iraqi (1)</td>
<td>#CLB, J-LB</td>
<td>جَلَبْ</td>
<td>جَلْبَةْ</td>
</tr>
<tr>
<td>Iraqi (2)</td>
<td>JLB</td>
<td>جَلَبْ</td>
<td>جَلْبَةْ</td>
</tr>
<tr>
<td>Jordanian</td>
<td>T:LB</td>
<td>تَشْلَبْ</td>
<td>تُشَلْبَةْ</td>
</tr>
</tbody>
</table>
Spellings for “dog” in Iraqi lexicons:
J-LB, J-ALB, JLB, JALB

<table>
<thead>
<tr>
<th>Phonetic</th>
<th>SATTS</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ich-chalib</td>
<td>ALJ-LB</td>
<td>the dog</td>
</tr>
<tr>
<td>2 hach-chalib</td>
<td>?ALJ-ALB</td>
<td>this dog</td>
</tr>
<tr>
<td>3 ch-chalib</td>
<td>ALJ-ALB</td>
<td>the dog</td>
</tr>
<tr>
<td>4 chalbi</td>
<td>J-ALBI</td>
<td>my dog</td>
</tr>
<tr>
<td>5 chalibna</td>
<td>J-ALBNA</td>
<td>our dog</td>
</tr>
<tr>
<td>6 il-chalib</td>
<td>ALJLB</td>
<td>the dog</td>
</tr>
<tr>
<td>7 chalibhum</td>
<td>--------</td>
<td>their dog</td>
</tr>
<tr>
<td>8 lich-chalib</td>
<td>-------</td>
<td>to the dog</td>
</tr>
<tr>
<td>9 chalba</td>
<td>JLB?</td>
<td>(fem.) dog</td>
</tr>
<tr>
<td>10 chalba</td>
<td>JALBA</td>
<td>(fem.) dog</td>
</tr>
</tbody>
</table>
Mapping variant orthographic forms to a single canonical form

<table>
<thead>
<tr>
<th>Native Variants</th>
<th>SATTS Variants</th>
<th>Canonical Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>كلب</td>
<td>KLB</td>
<td>كلب</td>
</tr>
<tr>
<td>#CLB</td>
<td>J-LB</td>
<td>كلب</td>
</tr>
<tr>
<td>JLB</td>
<td>T:LB</td>
<td>“dog”</td>
</tr>
</tbody>
</table>
Morphological Query Expansion

• Because Arabic is morphologically complex, finding all the relevant forms of a single word often takes multiple queries (even dozens of queries for complete coverage of certain words).

• We are developing a morphological query expansion (MQE) tool that, given a query, will
  – identify the root of the queried word forms
  – generate a list of morphological variants (reflexes) of this root.
  – allow automatic filtering to user-specified criteria
Morphological Query Expansion

• We want to use Arabic linguistic structure
  – Roots, patterns, CV templates
  – Well-known inflectional classes
  – Shared patterns between MSA, dialects

• MAGEAD (Habash et al. 2006) does this
  – Built on AT&T FST, LexTools (Sproat)
  – Currently implements MSA, Levantine Verbs
  – We propose to extend it to Iraqi Nouns, Adj.
MAGEAD

• Extension of two-level morphology

• Five tiers:
  – Root (radical consonants) + features
  – Pattern (Vowels, non-radical consonants)
  – CV template (including length of Cs and Vs)
  – Phonological Tier
  – Orthographic Tier (Unicode Native Script)

• Uses multiple inheritance hierarchy for handling inflectional classes
Phonological Query Expansion

Other Iraqi Arabic stumbling blocks:

- **Phonotactic variation**: Words spoken differently depending on the surrounding sounds
- **Stylistic variation**: Stylistic shifts, and words and pronunciations borrowed from MSA

As these variants (especially the phonotactic variants) may be predictable through automatic rules, a companion tool to generate phonologically expanded query sets will be designed.
Phonological Query Expansion: Examples

How do you say “dog” in Arabic?

Main pronunciations:

(1) /kalb/      [k a l b]    MSA
(2) /čalib/     [tS a l i b]  Iraqi

Variant pronunciations:

(3) /kalib/     [k a l i b]   MSA
(4) /čalb/      [tS a l b]    Iraqi
Questions?