Arabic Language Challenges

Walid Magdy
This lecture is not

- About Arabic language technologies
- Description of the state-of-the-art
- Highly technical
- Duplicate to other presentations (I hope)
- Boring (promise)
This lecture is about

- Why Arabic Language is Important
- Arabic orthographic nature
- Arabic morphological nature
- Arabic phonetic nature
- Challenges stem from this nature
This sentence is written in Arabic language

هذه الجملة مكتوبة باللغة العربية

هذه الجملة مكتوبة باللغة العربية

هذه الجملة مكتوبة باللغة العربية
Language Technology

Technology Related to the Language People Speak

- Information retrieval (Google)
- Translation (Google-translate)
- Question Answering
- Sentiment Analysis
- Automatic Speech Recognition (ASR, e.g. Siri)
- Optical Character Recognition (OCR)
Arabic Language

- Arabic is the largest living member of the Semitic language family
- It is classified as a macro-language with 27 sub-languages
- It is spoken by over 280 million people in 28 countries (middle-east)
- The language of Quran (over 1.6 billion Muslims)
Arabic Language (Internet)

Growth in Internet Users by Language (2000-2010)

- English
- Chinese
- Spanish
- Japanese
- Portuguese
- German
- Arabic
- French
- Russian
- Korean
- Rest of the Languages

Institute for Language, Cognition and Computation
ILCC

THE UNIVERSITY OF EDINBURGH
Arabic Language (Types)

- Current written Arabic is the *modern standard Arabic*
  - Unified across all Arabic countries (news, political speeches)
  - Easy to understand by all Arabs
  - Not spoken by people!

- Spoken Arabic (dialectic Arabic)
  - Different across Arabic countries (regions)
  - Semi-understandable by different Arabic dialectic
  - For informal use (on social media)

- Classic Arabic (Language of Quran)
  - Contains ancient Arabic words
  - Mostly understandable by Arabic people
  - Previously used different version of Arabic scripts
Arabic Language Nature

- **Orthographical nature:**
  The way to write Arabic letters
  OCR

- **Morphological nature:**
  The way to construct Arabic sentences
  NLP, IR, MT, QA

- **Phonetic nature:**
  The way to pronounce Arabic letters and words
  ASR, T2S, S2S
Orthographical Nature

- Written from right to left (letters only)
- 15 of the 28 letters contain dots
- Characters are connected or semi-connected
- Character shape depends on position
- Printed text may include ligatures and kashida
- Optional diacritics may be present
15 of the 28 letters contain dots

deď
th d

اَآِ
th t t b

ر ز
z r

س ش
sh s

ص ض
D S

ع غ

ط ظ
Z T

kh g
Character shape depends on position

<table>
<thead>
<tr>
<th>middle</th>
<th>begin</th>
<th>end</th>
<th>isolated</th>
</tr>
</thead>
<tbody>
<tr>
<td>بـ بـ بـ</td>
<td>بـ بـ بـ</td>
<td>بـ بـ بـ</td>
<td></td>
</tr>
<tr>
<td>حـ حـ حـ</td>
<td>حـ حـ حـ</td>
<td>حـ حـ حـ</td>
<td></td>
</tr>
<tr>
<td>سـ سـ سـ</td>
<td>سـ سـ سـ</td>
<td>سـ سـ سـ</td>
<td></td>
</tr>
<tr>
<td>عـ عـ عـ</td>
<td>عـ عـ عـ</td>
<td>عـ عـ عـ</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>middle</th>
<th>begin</th>
<th>end</th>
<th>isolated</th>
</tr>
</thead>
<tbody>
<tr>
<td>نـ نـ نـ</td>
<td>نـ نـ نـ</td>
<td>نـ نـ نـ</td>
<td></td>
</tr>
<tr>
<td>يـ يـ يـ</td>
<td>يـ يـ يـ</td>
<td>يـ يـ يـ</td>
<td></td>
</tr>
<tr>
<td>هـ هـ هـ</td>
<td>هـ هـ هـ</td>
<td>هـ هـ هـ</td>
<td></td>
</tr>
<tr>
<td>رـ رـ رـ</td>
<td>رـ رـ رـ</td>
<td>رـ رـ رـ</td>
<td></td>
</tr>
</tbody>
</table>
Presence of kashida and ligatures
Optional diacritics may be present

إن الله يأمر بالعدل والإحسان وإيتاء ذي القربى وينهى عن الفحشاء والمنكر والبغي يعظكم لعلكم تذكرون
It was very ambiguous

ان الله نامر بالعدل والاحسان واساء دى الفری ومسی عن الفحساء والمنكر والمعي بعطكم لعلكم بذكررون

ووجه بضرورة إنشاء مقر دائم لنقابة المهن التمشيلية بمدينة الإنتاج الإعلامي واتخاذ الإجراءات القانونية اللازمة لحماية حق الفنان في الحصول على حقوق الأداء العامي ودراسة إقرار العقد الموحد الذي يتطلب به النقابات الفنية لضمان حقوق الفنانين، ووجه كذلك بدعم صناديق العلاج للفنانين بنقاباتهم الفنية.
What about Arabic OCR?

- Word Error Rates (WER) are considerably high
- Good Arabic OCR: 30-40% WER on average
- Trained on similar font: <10% WER
- Old fonts: >70% WER
- Average WER for English: <5%
Morphological Nature

- Language is built of 10k roots
- Short vowels are not written (diacritics)
- Words contain **prefix**, **infix**, and **suffix** (pronouns, others) (the, and, his, her, their, it, him, them, will …) are attached to the main word
- Word spelling can change according to grammatical position
- No rule for plural words
- 60 billion possible surface forms
Short vowels are not written

- In the Arabic text we do not write its short vowels and the pronouns are attached to the words.
- In the Arabic text we do not write its short vowels and the pronouns are attached to the words.
- In the Arabic text we do not write its short vowels and the pronouns are attached to the words.

- كتَبَ (kataba) write
- كتب (kotub) books
- كتَبَ (kattaba) let someone write
- كتب (kuttiba) forced to write
Words contain prefix, infix, and suffix

They are Peter’s children
The children behaved well
Her children are cute
My children are funny
We have to save our children
Patents and children are happy
He loves his children
His children loves him

هواء أبناء بيتر
الأبناء تصرفوا جيدا
أبناءها لطاف
أبنائي ظريفاء
علينا أن نحمي أبناءنا
الأباء والأبناء سعداء
هو يحب أبناءه
ابناوئ يحبونه

وسيكتبونها
wasaya+ktub+unahaa
and will + write + they it
= and they will write it

كتب
(kataba)
write

كاتب
(kateb)
writer

كتاب
(ketab)
book
No rule for plural

<table>
<thead>
<tr>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>رجل</td>
<td>رجال</td>
</tr>
<tr>
<td>كاتب</td>
<td>كتاب</td>
</tr>
<tr>
<td>مكتب</td>
<td>مكاتب</td>
</tr>
<tr>
<td>مكتبة</td>
<td>مكتبات</td>
</tr>
<tr>
<td>هاتف</td>
<td>هاتف</td>
</tr>
<tr>
<td>مصلي</td>
<td>مصليين</td>
</tr>
<tr>
<td>إمام</td>
<td>أئمة</td>
</tr>
<tr>
<td>man</td>
<td>men</td>
</tr>
<tr>
<td>writer</td>
<td>Writers</td>
</tr>
<tr>
<td>office</td>
<td>offices</td>
</tr>
<tr>
<td>library</td>
<td>libraries</td>
</tr>
<tr>
<td>telephone</td>
<td>telephones</td>
</tr>
<tr>
<td>prayer</td>
<td>prayers</td>
</tr>
<tr>
<td>leader</td>
<td>leaders</td>
</tr>
</tbody>
</table>
What about Arabic IR?

- Some characters are normalized
- Diacritics (short vowels) are removed (if existed)
- Later approaches for search
  - Search with words
  - Apply light stemming for words
  - Apply morphological stemming for words
  - Simple character n-grams representation

- New Methods are being developed for Social Arabic

ايشلونك؟، ازاي؟، صج؟، زمبقولك كدة، لول
we lessa ba2a el3arabi elli maktoob bel7rof el inglizi :D
Phonetic Nature

Some phonemes are in Arabic don’t exist in other language (‘ein, ghain, ha, kha, Dad, Sad, Ta, Hamza)

Examples:
Mohamed (ha)
Attia (‘ein, Ta)
Khalid (kha)
Ghada (ghain)
Baraa (Hamza)
Diaa (Dad, Hamza)
What about Arabic ASR?

- Needs special training and decoding
- Requires huge amount of training
- Requires diacritisation as a pre-processing step
- State-of-the-art is not bad (for MSA)
- Again for dialect, it is too bad
## State-of-the-art / Areas of research

<table>
<thead>
<tr>
<th>Language Technology</th>
<th>MSA</th>
<th>Dialect Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stemming (Segmentation)</td>
<td>Good</td>
<td>Needs work</td>
</tr>
<tr>
<td>POS</td>
<td>Good</td>
<td>Good for some</td>
</tr>
<tr>
<td>NER</td>
<td>Good</td>
<td>Can be improved</td>
</tr>
<tr>
<td>Search (IR)</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>ASR</td>
<td>Good</td>
<td>Needs work</td>
</tr>
<tr>
<td>Sentiment analysis</td>
<td>Needs work</td>
<td>Not working!</td>
</tr>
<tr>
<td>Sarcasm detection</td>
<td>NA</td>
<td>HELP!!</td>
</tr>
<tr>
<td>Syntactic tree parsing</td>
<td>kind of</td>
<td>What is it?</td>
</tr>
</tbody>
</table>
Conclusion

- Language technology requires deep algorithms to overcome language challenges
- Arabic language is full of challenges
- Huge amount of work already done
- Huge amount of work is still needed
- Some languages are just harder to deal with in NLP than others!
Thank you

شكرا

(shokran)