Winter Institute in Digital Humanities 2020
Text Analytics for Arabic Course

Hands-on Exercise Overview

Part Baa (Session #3 and #4)

Ossama Obeid, Salam Khalifa, Dima Taji, and Nizar Habash
CAMeL Lab, New York University Abu Dhabi

Outline

The hands-on exercise is about the automatic processing of four different Arabic corpora, and collecting statistics and calculating textual similarity for them.

The exercise has four parts:

1. **Data Extraction and Cleaning**
2. **Running the MADAMIRA Arabic Morphological Disambiguator**
3. **Vocabulary Frequency**
4. **Vocabulary Overlap**

File Formats

- Data files come in different formats.
- We will show examples of four different files that have different combinations of encodings and file structures, and how to process each one of them.
- The files will vary also in genre (to keep things interesting).

Genre	Variant	Data Type	Encoding
Gigaword News	MSA	Raw text	Arabic Script (UTF8)
Hindawi Novels	MSA	Raw text	Buckwalter transliteration
MADAR Travel Domain	Dialects	Tab Separated Files	Arabic Script (UTF8)
UN Corpus	MSA	XML directory	Arabic Script (UTF8)

Raw Text in Arabic Script

Excerpt from Gigaword:

وافادت وكالة انباء الشرق الاوسط المصرية ان الرئيسين المصرى والفلسطينى عقدا جلسة ثنائية عقبتها جلسة موسعة حضرها من الجانب المصرى عاطف عبيد رئيس الوزراء والمشير حسين طنطاوى وزير الدفاع والانتاج الحربى وصفوت الشريف وزير الاعلام واحمد ماهر وزير الخارجية واسامة الباز المستشار السياسى للرئيس المصرى .

Raw Text in Buckwalter Transliteration

Excerpt from Taha Hussein's novel "Curlew's Prayer"

wkAn fy >vnA' *lk rbmA dEAny <lY grfth w>x* ytHdv <ly wysmE mny، wkAnt Almdynp w$}wn >hlhA mwDwE HdyvnA fy kvyr mn Al>HyAn، kmA kAnt AlqAhrp w$}wnhA mwDwE HdyvnA >HyAnFA >xrY.

Tab Separated Files (TSV)

Excerpt from the MADAR 6 Corpus: <Text><tab><Dialect>

بغيت شي حاجة توقف الغيلوف . RAB
ما عندكش صرف كتر من هادا ؟ RAB
لأ . شارع واحد من هون . BEI
هي شنطة زرقة . DOH
ابغي ام الربيان نية . DOH
يكلفني كام لسياتل ؟ CAI
فين تخدم ؟ TUN
تذكرتين للنهاردة بليل ، لو سمحت . CAI

XML Files

Excerpt from The United Nations Parallel Corpus v1.0

```
<text>
   <body>
      <p id="36">
            <s id="36:1" lang="ar">2001 مراجعة مفهوم &amp;quot;الدولة المطلقة&amp;quot; بصيغته
الواردة في اتفاقية المسؤولية الدولية عن الأضرار التي تحدثها الأجسام الفضائية (مرفق قرار الجمعية العامة 2777
واتفاقية تسجيل الأجسام المطلقة في الفضاء الخارجي (مرفق قرار &amp;quot;اتفاقية المسؤولية&amp;quot;) ((د-26
29-د) 3235 (الجمعية العامـة &amp;quot;اتفاقية التسجيل&amp;quot;)، حسبما تطبقه الدول والمنظمات الدولية.</s>
      </p>
   </body>
</text>
```

Exercise #1: Data Extraction and Cleaning

This exercise provides the scripts to:

- Extract the data from each of the different previously mentioned file formats
- Clean the data, including:
  - Unicode cleaning
  - Transliteration
  - Arabic encoding cleaning
- Produce a clean text file that you can use in the remaining exercise partsd

Exercise #2: Running MADAMIRA

- MADAMIRA is a tool for Arabic text disambiguation.
  - Out of all the possible readings for each word, MADAMIRA selects the best meaning for this word in context.
- The provided MADAMIRA configuration creates the following output files:
  - <file>.mada — Morphological Analysis and Disambiguation of Arabic
  - <file>.ATB.tok — Arabic Treebank Tokenization
  - <file>.D3.tok — D3/Full Decliticization Tokenization
- We provide extra scripts to extract additional representations from <file>.mada
  - <file>.mada.diac — Diacritized text
  - <file>.mada.undiac — Undiacritized text (spelling adjusted from raw input)
  - <file>.mada.lex — Lemmas
  - <file>.mada.lexPOS — Lemmas+Part-of-speech (34 tags)

Exercise #3: Vocabulary Frequency

This exercise provides you with a script that reads a file produces a frequency-token list in addition to reporting file statistics.

The script is independent of the specific types of tokens. File tokens can be raw words, lemmas, tokenized words, lemma+POS tokens, etc.

Run the script on the different files you generated from Exercise #2.

- How do the vocabulary frequencies vary by genre?
- How do the vocabulary frequencies vary by token form (lemmas vs undiac vs diac, e.g.) ?

Exercise #4: Vocabulary Overlap

Vocabulary overlap is a measure of similarity between two files.

- There are many metrics for text similarity. We use the weighted Jaccard metric.
- The two files can be of the same or different genres, in same or different representations.
- Interpretation of the similarity measure depends on what the input files are.

This exercise provides you with a script that reads two files and produces the text similarity score.

- Run the script on different file pairs.
- Compare the output of different file pairs. Are there surprises?