Extractive Text-Based Summarization of Arabic videos: Issues, Approaches and Evaluations


Nancy, France

16, 17 October 2019
Introduction: Context and objectives

- How a user can access to the information which is expressed in a foreign language?
- Understanding a video in a foreign language is first step to answer this question.

- Develop and evaluate a system for automatic summarization of Arabic videos.
How a user can access to the information which is expressed in a foreign language?

Understanding a video in a foreign language is first step to answer this question.

Develop and evaluate a system for automatic summarization of Arabic videos.
Automatic Speech Recognition -ASR-

Modern Standard Arabic and dialect cases
ASR: From the signal to the text
Extract the acoustic features (MFCC, PLP ...).
ASR: From the signal to the text

- Extract the acoustic features (MFCC, PLP ...).
Acoustic Modeling:

- DNN-HMM model is used for the acoustic modeling.
Acoustic modeling: DNN-HMM acoustic modeling
ASR: Modern Standard Arabic case

- Acoustic modeling: DNN-HMM acoustic modeling

- 44 hours of MSA spoken data are used for training the neural network: Nemlar\(^a\) and NetDC\(^b\);

- 440-dimensional input layer (11*40-dimensional acoustic features);

- 6 hidden layers with 2048 nodes by layer;

- and 4264-dimensional output layer (number of HMM states).

\(^a\)http://catalog.elra.info/product_info.php?products_id=874

ASR: Modern Standard Arabic case

Language modeling:
- n-gram model is used for the language modeling.

$P(W) = \prod_{i=1}^{M} P(w_i | w_{i-1} \ldots w_{i-n-1})$  \hspace{1cm} (1)

$p^{FinalLM}(W) = \lambda_1 P^{LM1}(W) + \lambda_2 P^{LM2}(W)$  \hspace{1cm} (2)
Priming modeling:

- Select 100k most frequent words from the textual data.
- Use an external lexicon\(^1\) to generate pronunciation.

<table>
<thead>
<tr>
<th></th>
<th>#Words</th>
<th>#Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSA</td>
<td>95K</td>
<td>485K</td>
</tr>
</tbody>
</table>

Table: Statistics about the MSA lexicon.

\(^1\)http://alt.qcri.org/resources/msa-dictionary/
This Algerian dialect is highly impacted by the MSA and French language.

The Algerian dialect is mainly spoken, there are no data to train the different model.
This Algerian dialect is highly impacted by the MSA and French language.

The Algerian dialect is mainly spoken, there are no data to train the different model.

Explore data that impact the Algerian dialect, namely MSA and French to enhance models for the dialect.
Textual data collection:

- two corpora containing Algerian dialects are constituted: PADIC$^2$ and CALYOU$^3$ corpora.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>#Words</th>
<th>#Unique words</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALYOU</td>
<td>10M</td>
<td>512k</td>
</tr>
<tr>
<td>PADIC</td>
<td>25k</td>
<td>6.6K</td>
</tr>
</tbody>
</table>

Table: Statistics about textual data.


Spoken data:

- The aligned dialectal spoken corpus is created by having native Algerian people reading 4.6k sentences extracted from PADIC and CALYOU corpora.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Dur</th>
<th>Female spkrs</th>
<th>Male spkrs</th>
<th>Total spkrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>240 min</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Dev</td>
<td>40 min</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Test</td>
<td>75 min</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Table: Some characteristics of the dialectal corpus.
Acoustic modeling:

- The dialectal corpus is quite small to train a robust AM.

The amount of data is optimized iteratively on the dialectal Dev corpus.
ASR: Algerian dialect case

- Acoustic modeling:
  - The dialectal corpus is quite small to train a robust AM.

- The amount of data is optimized iteratively on the dialectal Dev corpus.

```
<table>
<thead>
<tr>
<th>MSA data</th>
<th>French data</th>
<th>Dialectal data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

```
Output (softmax) layer
```

```
6 hidden layers
```

```
Input layer
```

```
Data merge
```

```
DNN-HMM model
```

```
All data?
```

```
Best model
```

```
4h recorded sentences
```

```
dialectal data
```

Adding 4h of MSA and Fr data

```
4h dialectal data
```

```
+ 12h MSA + 12h Fr data
```

```
no
```

○ Language modeling:
  ○ The LM is a linear interpolation of 4 LMs.
Pronunciation modeling:

- Adapt the approach proposed in\(^4\) to generate the pronunciation of dialectal words.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>#Words</th>
<th>#Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSA</td>
<td>95K</td>
<td>485K</td>
</tr>
<tr>
<td>CALYOU</td>
<td>50K</td>
<td>50K</td>
</tr>
<tr>
<td>PADIC</td>
<td>6.6K</td>
<td>6.6K</td>
</tr>
<tr>
<td>Total</td>
<td>123K</td>
<td>538K</td>
</tr>
</tbody>
</table>

Table: Statistics about lexicons.

The test is carried out on the 75 min of the dialectal data and 5 hours of MSA data:

<table>
<thead>
<tr>
<th>System</th>
<th>AM</th>
<th>LM</th>
<th>Lex</th>
<th>WER_dial (%)</th>
<th>WER_MSA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR-MSA</td>
<td>MSA</td>
<td>MSA</td>
<td>MSA</td>
<td>78.5</td>
<td>14.02</td>
</tr>
<tr>
<td>(S_1)</td>
<td>4h dial</td>
<td>MSA+dial</td>
<td>MSA+dial</td>
<td>40</td>
<td>/</td>
</tr>
<tr>
<td>(S_2)</td>
<td>MSA+Fr+dial</td>
<td>MSA+dial</td>
<td>MSA+dial</td>
<td>37.7</td>
<td>/</td>
</tr>
</tbody>
</table>

Table: Performance of the ASR systems on the Test dialectal corpus.
Automatic text summarization

Sentence Boundary Detection
Sentence Boundary Detection: Architecture
The CNN is trained on 70M words subset extracted from the Gigaword corpus.
The evaluation is carried out on 10.5M samples.

<table>
<thead>
<tr>
<th>class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;SEG&gt;</td>
<td>0.797</td>
<td>0.612</td>
<td>0.684</td>
</tr>
<tr>
<td>&lt;NO SEG&gt;</td>
<td>0.972</td>
<td>0.989</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table: Sentence Boundary Detection performance.
Automatic text summarization

Automatic text summarization
Document preprocessing  The text is represented in a suitable space model.

Global topic vector  An average document vector is built.

Lexical weight  A lexical vector is built for each sentence.

Sentence scoring  A score for each sentence is calculated using their proximity with the global topic vector and their lexical weight.

\[
score(s_i) = (\overrightarrow{s} \times \overrightarrow{b}) \times \overrightarrow{a} = \frac{1}{NP} (\sum_j s_{i,j} \times b_j) \times a_i
\]

Sentence selection  The summary is generated concatenating the sentences with the highest scores following their order in the original document.

---

Tests ans results

Evaluation
French, English and Arabic videos are collected according to a set of controversial Twitter Hashtags such as 

#سوريا, #حقوق_المراة.

More than 1.5K Arabic videos (>100h) are collected. they come from channels such as AlArabiya, France24, EchoroukTV, EnnaharTV, BBC, etc.

<table>
<thead>
<tr>
<th>Count</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Videos</td>
<td>27</td>
</tr>
<tr>
<td>Summary per Video</td>
<td>3</td>
</tr>
<tr>
<td>Channel TV</td>
<td>3</td>
</tr>
<tr>
<td>Evaluators</td>
<td>3</td>
</tr>
<tr>
<td>Size of the shortest summary (in words)</td>
<td>52</td>
</tr>
<tr>
<td>Size of the longest summary (in words)</td>
<td>394</td>
</tr>
</tbody>
</table>

Table: Some figures concerning the subjective evaluation.
**Evaluation: Subjective evaluation, MSA case**

1. Incomprehensible summary
   - Only some events of the original video are found in the summary and overall the text is incomprehensible

2. Only certain segments of the video are understandable
   - A substantial proportion of the transcription is understandable
   - The transcription is very understandable

3. Very good summary and the text is very correct
   - A substantial proportion of the events in the original video are in the summary and overall the text is understandable

4. Excellent summary
   - Only some events of the original video are found in the summary and overall the text is incomprehensible

5. The transcription is not only understandable, but it is fluid

**Figure:** Rating scale for the automatic summarization system assessment.

**Figure:** Rating scale for the automatic speech recognition system assessment.
Figure: The Box plot corresponding to the subjective evaluation of the Arabic ASR and the automatic summarization systems on MSA data.
Figure: The number of responses for each score of the subjective assessment of dialectal data with MSA-ASR system.

Figure: The number of responses for each score of the subjective assessment of dialectal data with the adapted ASR system.
What is the relationship between the scores of the summary $EvalSum$ and:

- the number of words ($ASRWord$);
- the score of the ASR system ($ASRScore$);
- and the number of words of the summary ($SumWord$).

Use the multiple linear regression through the coefficient of determination ($R^2$).

On our data-set of 243 examples, $R^2 = 0.310$, this indicates that 31% of the dispersion is explained by the regression model.
Evaluation: Factors impacting summary

- $H_0 : a_1 = a_2 = a_3 = 0$ and $H1$ at least one of the $a_i$ is different from 0.

\[ F = \frac{R^2}{\frac{p}{1-R^2}} \frac{n-p-1}{p} \]  

(4)

- $R^2 = 0.31$
- $p = 2$ #dependent variables
- $n = 243$ #samples
- $F = 35.899$
- $F > F_{0.95}(2, 240)$

- $d_1 = 2, d_2 = 240$
Describe the development and the evaluation of an automatic video summarization system.

- The ASR system was developed for MSA and adapted for the Algerian dialect.
- Each component performs well separately.
- Several parameters impact the summary, namely the number of words in the original/summarized video and the output of the ASR system.
Thank you for your attention
Questions?