Auto Generation of Arabic News Headlines

Project team: Omar Alhadlaq (hadlaq@) Saud Alsaif (salsaif@) and Yaya Khoja (ykhoja@)

Project Motivation
- Automating the generation of useful headlines for journalists
- Understanding the specific challenges of an Arabic NLP project
- Potentially can be expanded to summarizing Arabic articles

Data Source
We used 31,030 Arabic news articles and headline pairs in the Saudi Newspapers Arabic Corpus, SaudiNewsNet, that were extracted from 14 different news websites covering topics like politics, local news, culture, and sports with the following fields:

- source: the newspaper from which the article was extracted
- url: a full URL to the article
- date_extracted: a timestamp of the date on which the article was extracted
- title: the headline of the article (can be empty)
- author: the author of the article (can be empty)
- content: the body of the article

Data Preprocessing
- Data cleanup: remove annotations and punctuation marks
- Data splitting into 80% train, 8% validation and 12% test sets
- Adding special tokens:
  - PAD: to pad articles and headlines below max length
  - UNK: for words not in the vocabulary
  - START and STOP: to start and stop the model decoder

Challenges
- The same Arabic word might be spelled differently due to annotation of conjugation of the word
- Training the model requires high computation power
- The complexity of the model made it harder to debug

Sequence-to-sequence model

Hyperparameters

<table>
<thead>
<tr>
<th>Learning Rate</th>
<th>Max Gradient Norm</th>
<th>Batch Size</th>
<th>Embeddings Size</th>
<th>Vocab Size</th>
<th>Article Length</th>
<th>Headline Length</th>
<th>Dropout Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>5.0</td>
<td>64</td>
<td>64</td>
<td>100K</td>
<td>200</td>
<td>10</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Adding Attention

Motivation: Using attention can help when the input contains long sentences, such as in the case of articles.

Examples and Observations

Examples:
- Human title (translated):
  6 Died and Injured in a Traffic Accident East of Quwaieyah
- Standard title (translated):
  8 of One Family Died and Injured in a Car Flip Over on Khurais
- Attention title (translated):
  Three Died and Seven Injured in a Traffic Accident to Arar

General Observations:
- Can’t strongly affirm which model is better
- Generated titles overall seem related to the article
- They tend to make up some of the details mentioned in the article
- In the example given above, both of the models talk about the car accident. But, they also made up the numbers and the location of the accident. The standard model also decided all of these people are related and the accident was, in fact, a flip over

Next Steps
- Train the attention model for more epochs
- Try a copying mechanism to avoid unks and better incorporate article details (numbers, nouns, etc) in the title

Results

Results on numeric evaluation metrics:

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev Ppl.</th>
<th>Dev BLEU</th>
<th>Test Ppl.</th>
<th>Test BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard (30K)</td>
<td>3969</td>
<td>0.9697</td>
<td>3663</td>
<td>1.053</td>
</tr>
<tr>
<td>Standard (10K)</td>
<td>1570</td>
<td>0.6354</td>
<td>1512</td>
<td>0.6003</td>
</tr>
<tr>
<td>Attention</td>
<td>2700</td>
<td>0.6653</td>
<td>2417</td>
<td>0.7408</td>
</tr>
</tbody>
</table>

Results on numeric evaluation metrics:

Examples:
- Human title (translated):
  6 Died and Injured in a Traffic Accident East of Quwaieyah
- Standard title (translated):
  8 of One Family Died and Injured in a Car Flip Over on Khurais
- Attention title (translated):
  Three Died and Seven Injured in a Traffic Accident to Arar

General Observations:
- Can’t strongly affirm which model is better
- Generated titles overall seem related to the article
- They tend to make up some of the details mentioned in the article
- In the example given above, both of the models talk about the car accident. But, they also made up the numbers and the location of the accident. The standard model also decided all of these people are related and the accident was, in fact, a flip over

Next Steps
- Train the attention model for more epochs
- Try a copying mechanism to avoid unks and better incorporate article details (numbers, nouns, etc) in the title