The QCN System for Egyptian Arabic to English Machine Translation

Presenter: Hassan Sajjad
QCN Collaboration

Ahmed El Kholy
Wael Salloum

Nizar Habash

Ahmed Abdelali
Nadir Durrani
Francisco Guzmán
Preslav Nakov
Hassan Sajjad
Stephan Vogel
NIST Egyptian Arabic-English Dataset

• Three genres
  - SMS, Chat and CTS

• Dataset distribution
  - Approximately 3000 sentences for tuning
  - The rest is used for training

• Development sets provided by NIST
  - Test: devTest
  - TestG: gold devTest
Baseline System Settings

- Phrase-based SMT system with the following settings:
  - MGIZA for alignment
  - Phrase tables with Kneser-Ney smoothing
  - Lexicalized reordering
  - Operation sequence model
  - Tuning using PRO and MIRA
  - Minimum Bayes risk decoding
  - Cube pruning
  - Other Moses defaults…
Important Modules

**Data Preprocessing**
- Arabizi to UTF8 conversion
- Normalization
- Speech markups removal
- Cleaning
- Intended vs. literal meaning
- Egyptian Arabic segmentation
- Egyptian Arabic to MSA conversion

**System Features**
- Class-based models
- Neural network joint model
- Interpolated language model
- Sparse features
- Adaptation
- Unsupervised transliteration model
- System combination
Data Preprocessing

- Arabizi to UTF-8 using 3arrib

- Normalization
  - Emoticons e.g. - > : = P
    - Tokenizer splits them into single units like - > : = P
    - Normalizing emoticons to their original form
  - Fixed character repetitions on both Arabic and English side
    - Map repetitions like hahahahahahah to one from say, haha
    - Convert emphasis repetitions like Yessss to their original form

- Removing markups e.g. %fw, %fp, {laugh}
Data Preprocessing: Egyptian Segmentation

- Segmentation of Egyptian Arabic using **MADAMIRA**
  - ATB, S2, D3

<table>
<thead>
<tr>
<th></th>
<th>SMS Test</th>
<th>SMS TestG</th>
<th>CHT Test</th>
<th>CHT TestG</th>
<th>CTS Test</th>
<th>CTS TestG</th>
</tr>
</thead>
<tbody>
<tr>
<td>No segmentation</td>
<td>21.02</td>
<td>21.64</td>
<td>20.27</td>
<td>22.34</td>
<td>20.60</td>
<td>23.36</td>
</tr>
<tr>
<td>D3</td>
<td>23.68</td>
<td>23.41</td>
<td>23.22</td>
<td>25.97</td>
<td>21.72</td>
<td>24.89</td>
</tr>
<tr>
<td>S2</td>
<td>23.62</td>
<td>23.66</td>
<td>22.82</td>
<td>25.41</td>
<td>21.61</td>
<td>24.67</td>
</tr>
<tr>
<td>ATB</td>
<td>23.57</td>
<td>23.50</td>
<td>22.82</td>
<td>26.01</td>
<td>21.68</td>
<td>24.83</td>
</tr>
</tbody>
</table>
Data Preprocessing: Egyptian to MSA Conversion

• Character-level system to convert Egyptian words to MSA
  • e.g. ِبيتكلم to ِيتكلم

<table>
<thead>
<tr>
<th></th>
<th>SMS Test</th>
<th>SMS TestG</th>
<th>CHT Test</th>
<th>CHT TestG</th>
<th>CTS Test</th>
<th>CTS TestG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egyptian</td>
<td>21.02</td>
<td>21.64</td>
<td>20.27</td>
<td>22.34</td>
<td>20.60</td>
<td>23.36</td>
</tr>
<tr>
<td>Converted MSA</td>
<td>21.54</td>
<td>21.82</td>
<td>20.70</td>
<td>22.77</td>
<td>21.30</td>
<td>23.81</td>
</tr>
<tr>
<td>Converted MSA, ATB</td>
<td>21.32</td>
<td>21.06</td>
<td>21.55</td>
<td>23.70</td>
<td>21.73</td>
<td>24.30</td>
</tr>
</tbody>
</table>

• Gains are low compared to the system trained using Egyptian segmentation

• Highly dialectal nature of the data
  - requires more lexical substitution than character-level changes
Data Preprocessing: Tuning Dataset Issues

• Missing markers of literal and actual translations in references

• Imbalanced length ratio (i.e. English sentence is 2x of an Arabic sentence)

• **Problem**: Imbalanced tuning sentences will result in bad tuning weights
Data Preprocessing: Cleaned Tuning Dataset

• Removing sentences with abnormal word length (4<length<25) and length ratio in either source or target

<table>
<thead>
<tr>
<th>Genre</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMS - unfiltered tune</td>
<td>23.57</td>
<td>23.56</td>
</tr>
<tr>
<td>SMS - filtered tune</td>
<td>23.35</td>
<td>24.36</td>
</tr>
<tr>
<td>CTS - unfiltered tune</td>
<td>22.07</td>
<td>25.10</td>
</tr>
<tr>
<td>CTS - filtered tune</td>
<td>22.70</td>
<td>25.22</td>
</tr>
</tbody>
</table>

+0.63 on CTS
Data Preprocessing: Ex. Noisy References

- Literal meaning is sometimes noisy
- **Solution**: We used the intended meaning only

Source

احدث الاغاني يا خارجة من باب الارشاد و واحده قرض من الدوحة، دوحة، دوحة دوحة

Reference

The latest song: [O Muslim Brotherhood who are borrowing money from Doha / O that cute girl who just took a fresh shower, with her cheeks beautifully reddish].

Actual meaning

The best songs oh who you are leaving from the door of Ershad and borrowing money from Doha, Doha, Doha
Important Modules

Data Preprocessing
- Arabizi to UTF8 conversion
- Normalization
- Speech markers removal
- Cleaning
- Intended vs. literal meaning
- Egyptian Arabic segmentation
- Egyptian Arabic to MSA conversion

System Features
- Class-based models
- Neural network joint model
- Interpolated language model
- Sparse features
- Adaptation
- Unsupervised transliteration model
- System combination
System Features: **Class-based Models**

- Map words into a coarse representation
  - Reduces data sparseness
  - Generalizes the data

- Word clusters using mkcls (k=50, 500)
  - Translation model
  - OSM model over cluster IDs

### Performance Table

<table>
<thead>
<tr>
<th></th>
<th>SMS</th>
<th>CHT</th>
<th>CTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test</td>
<td>TestG</td>
<td>Test</td>
</tr>
<tr>
<td>Baseline</td>
<td>24.22</td>
<td>24.33</td>
<td>23.02</td>
</tr>
<tr>
<td>+ class-based models</td>
<td><strong>24.63</strong></td>
<td><strong>25.16</strong></td>
<td>23.18</td>
</tr>
</tbody>
</table>
System Features: Neural Network Joint Model

• Distributed representation of words
  - Similar to class-based models
    o Reduces data sparseness
    o Generalizes the data

<table>
<thead>
<tr>
<th></th>
<th>Test</th>
<th>SMS</th>
<th>CHT</th>
<th>CTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>24.58</td>
<td>24.33</td>
<td>24.02</td>
<td>22.64</td>
</tr>
<tr>
<td>+ NNJM Model</td>
<td>25.01</td>
<td>25.72</td>
<td>24.24</td>
<td>22.68</td>
</tr>
</tbody>
</table>

+0.0 on CTS
System Features: Genre-Based Interpolated LM

- Divide the available data into groups such as target side of
  - available Egyptian data
  - available Chinese data
  - MSA News
  - MSA non-News

- Minimize the perplexity on each genre’s tuning set

<table>
<thead>
<tr>
<th></th>
<th>SMS</th>
<th></th>
<th>CHT</th>
<th></th>
<th>CTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test</td>
<td>TestG</td>
<td>Test</td>
<td>TestG</td>
<td>Test</td>
<td>TestG</td>
</tr>
<tr>
<td>Concatenated LM</td>
<td>24.19</td>
<td>24.00</td>
<td>23.34</td>
<td>25.89</td>
<td>22.75</td>
<td>25.09</td>
</tr>
<tr>
<td>Interpolated LM</td>
<td>25.20</td>
<td>25.04</td>
<td>23.48</td>
<td>26.16</td>
<td>23.01</td>
<td>25.67</td>
</tr>
</tbody>
</table>
System Features: Additional Features

- Domain indicator features
- Source and target word deletion features

<table>
<thead>
<tr>
<th></th>
<th>SMS</th>
<th>CHT</th>
<th>CTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test</td>
<td>Test</td>
<td>Test</td>
</tr>
<tr>
<td>Baseline</td>
<td>24.58</td>
<td>23.36</td>
<td>22.64</td>
</tr>
<tr>
<td>+ sparse features</td>
<td>24.54</td>
<td>24.02</td>
<td>21.61</td>
</tr>
</tbody>
</table>

Mixed results
System Features: Adaptation

- Egyptian data with three genres
- MSA data

Techniques
- Concatenation
- Phrase table merging
- Back-off phrase tables
System Features: Adaptation

- Various combination of available Egyptian data for training
- Testing on SMS genre

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
<th>TestG</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMS</td>
<td>21.30</td>
<td>21.99</td>
</tr>
<tr>
<td>CAT(SMS, CHT, CTS)</td>
<td>23.78</td>
<td>23.20</td>
</tr>
<tr>
<td>SMS, Backoff(CHT,CTS)</td>
<td>22.55</td>
<td>23.00</td>
</tr>
<tr>
<td>CAT(SMS,CHT), Backoff(CTS)</td>
<td>22.54</td>
<td>23.20</td>
</tr>
<tr>
<td>MergePT(CAT(SMS,CHT),CTS)</td>
<td>23.69</td>
<td>24.40</td>
</tr>
</tbody>
</table>

Concatenation works the best!
System Features: Adaptation

- MSA phrase tables – Backoff and Merging
- Helps to translate OOV words which would also help in human evaluation

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Test</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAT(SMS, CHT, CTS)</td>
<td>23.78</td>
<td>23.20</td>
<td></td>
</tr>
<tr>
<td>CAT(SMS, CHT,CTS), Backoff(MSA)</td>
<td>23.70</td>
<td>23.64</td>
<td></td>
</tr>
<tr>
<td>MergePT(CAT(SMS,CHT,CTS),MSA)</td>
<td>23.83</td>
<td>23.60</td>
<td></td>
</tr>
</tbody>
</table>
System Features: Unsupervised Transliteration Mining

- Used unsupervised transliteration mining module (implemented in Moses) to transliterate OOV words
  - extracts a list of candidates from parallel training sentences
  - mines transliteration pairs
  - builds a phrase table of transliteration options
  - Post-processes the machine translation output

- Most of the OOVs are non-named entities
  - Require translation rather than transliteration
System Combination

- Combine machine translation output of various systems.

<table>
<thead>
<tr>
<th></th>
<th>SMS Test</th>
<th>SMS TestG</th>
<th>CHT Test</th>
<th>CHT TestG</th>
<th>Test</th>
<th>TestG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egyptian D3</td>
<td>25.28</td>
<td>26.05</td>
<td>23.87</td>
<td>27.07</td>
<td><strong>23.34</strong></td>
<td><strong>26.05</strong></td>
</tr>
<tr>
<td>Egyptian S2</td>
<td>24.93</td>
<td>25.61</td>
<td>24.09</td>
<td>27.01</td>
<td>22.11</td>
<td>24.50</td>
</tr>
<tr>
<td>Egyptian ATB</td>
<td>25.13</td>
<td>25.80</td>
<td>24.24</td>
<td>27.41</td>
<td>22.83</td>
<td>25.56</td>
</tr>
<tr>
<td>Egyptian ATB + MSA backoff</td>
<td>25.20</td>
<td>25.04</td>
<td>23.48</td>
<td>26.16</td>
<td>23.01</td>
<td>25.67</td>
</tr>
<tr>
<td>Output combination</td>
<td><strong>26.13</strong></td>
<td><strong>26.79</strong></td>
<td><strong>24.86</strong></td>
<td><strong>27.95</strong></td>
<td>22.89</td>
<td>25.88</td>
</tr>
</tbody>
</table>
Summary

- Data preprocessing is one of the major challenges in this translation task
- Normalization such as handling emoticons, fixing repetitions and cleaning helps to achieve better alignment
- Improvements of each module vary by genre
- Consistent improvements
  - Egyptian Arabic segmentation (up to +3 points)
  - Genre-based interpolated LM (up to +1 points)
  - Class-based models (up to +0.6 points)
Thank you