Background

• Community forums where user post questions and answers are becoming increasingly popular
• Arabic community question answering (cQA) received little attention in the past
• The SemEval-2016 Task 3 offered a cQA Arabic dataset in the medical domain, where:
  o Given a question and 30 related question-answer pairs, rank the pairs with respect to the original question
  o QA pairs annotated as Direct, Relevant, or Irrelevant
• Challenges: long texts, user-generated content, medical terminology, mixed standard and colloquial language

Preprocessing with Keyword Extraction

• Text preprocessing in order to deal with several challenges:
  o Long texts: average question/answer length is 50/120 words
  o Rich morphology: multiple surface forms per lemma
  o Latin terminology in the medical domain
• Keyword extraction with TextRank
  o Treat every thread as a document
  o Form a graph where nodes are word types and edges represent co-occurrence in N-sized window
  o Compute importance weight iteratively and keep top P% of words
• Lemmatization: we apply MADA for finding lemmas and part-of-speech tags
• Stop-word removal: we keep only content words, Latin words, and words with no morphological analysis.

Feature Representation

• Given a question q’ and a related question-answer pair q-a, compute features between the pairs q’-q and q’-a
• Text-based features
  o Various text-similarity metrics such as Longest Common Substring, Longest Common Subsequence, Greedy String Tiling, etc. (Belinkov et al. 2015)
• Vector-based features
  o Vector representations of closest pairs of words or sentences in q’-q and q’-a
  o Word vectors computed from Arabic Gigaword and medical domain raw data using Word2Vec
  o Sentence representation is average of word vectors
• Machine translation evaluation features
  o BLEU, TER, Meteor

Tree Kernels

Syntactic Tree Kernels

\[ K((t_1, t_2), (u_1, u_2)) = TK(t_1, u_1) + TK(t_2, u_2) \]

Constituency Trees with relational labels

Experiments

• Preprocessing settings
  i. No preprocessing
  ii. Only keeping content lemmas
  iii. Only content lemmas and keyword extraction with TextRank params N=3, P=5
  iv. Same, with TextRank params N=4, P=1
• Tree kernels settings
  a. ConvKN-contrastive1: only basic features
  b. ConvKN-contrastive2: MT features
  c. ConvKN-primary: basic features + tree kernels

<table>
<thead>
<tr>
<th>MAP</th>
<th>AvgRec</th>
<th>MRR</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand</td>
<td>29.79</td>
<td>31.00</td>
<td>19.53</td>
<td>20.66</td>
<td>20.08</td>
<td>68.35</td>
</tr>
<tr>
<td>a</td>
<td>38.33</td>
<td>42.09</td>
<td>43.75</td>
<td>20.38</td>
<td>96.95</td>
<td>33.68</td>
</tr>
<tr>
<td>b</td>
<td>39.98</td>
<td>43.68</td>
<td>46.41</td>
<td>26.26</td>
<td>68.39</td>
<td>37.95</td>
</tr>
<tr>
<td>c</td>
<td>45.50</td>
<td>50.13</td>
<td>52.55</td>
<td>28.55</td>
<td>64.53</td>
<td>39.58</td>
</tr>
<tr>
<td>i</td>
<td>44.94</td>
<td>49.72</td>
<td>51.58</td>
<td>26.26</td>
<td>82.96</td>
<td>24.0</td>
</tr>
<tr>
<td>iii</td>
<td>42.95</td>
<td>47.61</td>
<td>49.55</td>
<td>27.20</td>
<td>74.40</td>
<td>39.84</td>
</tr>
<tr>
<td>i-iv</td>
<td>45.83</td>
<td>51.01</td>
<td>53.66</td>
<td>34.45</td>
<td>52.33</td>
<td>41.55</td>
</tr>
</tbody>
</table>

Future Work

• Combine keyword extraction with tree kernels
• How do deal with grammatical structure after keyword extraction?
• Automatically detecting the most important sentences to be matched with the tree kernels

References

• ConvKL at SemEval-2016 Task 3: Answer and Question Selection for Question Answering on Arabic and English Fora