Domain Attentive Fusion for End-to-end Dialect Identification with Unknown Target domain
Suwon Shon¹, Ahmed Ali², James Glass¹
MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA, USA¹
Qatar Computing Research Institute, HBKU, Doha, Qatar²

Motivation

- Domain mismatch issue on dialect identification task
  - one of the challenges in real-world spoken language processing
- Ideally we have domain matched result if we have enough data on multiple domains and know about the input data
- Proposed domain attentive fusion network to deal with the unknown target domain

Dataset Collection

- Identifying 5 Dialects
  - Modern Standard Arabic, Egyptian Levantine, Gulf, North African
- Multiple Domains
  - MGB-3: Collected from broadcast system
  - 53.6 hours (recorded), 10.0 hours [high qual.]
- VarDial 2018: Collected from YouTube
  - 1000 hours

Experimental results

<table>
<thead>
<tr>
<th>Training data</th>
<th>MGR-3 Test</th>
<th>VarDial 2018 Test</th>
<th>Averaged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc.</td>
<td>EER</td>
<td>Carg</td>
</tr>
<tr>
<td>MGR-3 Train + MGR-3 Dev (.d)</td>
<td>68.63</td>
<td>19.07</td>
<td>19.01</td>
</tr>
<tr>
<td>VarDial 2018 Train (β)</td>
<td>51.27</td>
<td>28.37</td>
<td>27.41</td>
</tr>
<tr>
<td>MGR-3 Train + MGR-3 Dev + VarDial 2018 Train (4+β)</td>
<td>61.86</td>
<td>22.92</td>
<td>21.41</td>
</tr>
<tr>
<td>Logistic regression fusion of A and B (optimized for A)</td>
<td>68.63</td>
<td>19.07</td>
<td>19.01</td>
</tr>
<tr>
<td>Logistic regression fusion of A and B (optimized for B)</td>
<td>57.84</td>
<td>24.30</td>
<td>23.34</td>
</tr>
<tr>
<td>Using fusion layer on A and B (Figure 1)</td>
<td>67.69</td>
<td>19.30</td>
<td>18.39</td>
</tr>
<tr>
<td>Domain Attentive fusion of A and B (Figure 2 (a))</td>
<td>67.49</td>
<td>18.52</td>
<td>18.01</td>
</tr>
<tr>
<td>Domain Attentive fusion of A and B (Figure 2 (b))</td>
<td>68.23</td>
<td>18.30</td>
<td>17.69</td>
</tr>
</tbody>
</table>

End-to-end Dialect ID

- CNN based End-to-end neural network structure
  - 2 FC layers (1500-600 neurons)
  - 4 CNN layers
  - Filter size: 40x5-50x67
  - Stride: 1-2-1
  - No. of filters: 500-500-500-3000

- Performance based on the training data
  - Using both domains for training performs generally better for both domain
  - Logistic regression based fusion with an individually trained network on a single domain gives better performance than training with multiple domains
  - However, fusion gives the best performance when the input domain is known a priori

Domain Attentive Fusion

- Self attention based fusion
- Learning scalar score $e_d$ for the language ID output $b_d$ where $d \in \{D_1, D_2\}$ as $e_d = f(\theta_d)$.
- Scoring function $f(\theta_d) = \sqrt{d} \tan(h(W_d0_d + b_d))$
- Normalized weights $a_d$ is $a_d = \frac{exp(e_d)}{exp(e_{d_1}) + exp(e_{d_2})}$
- Domain attentive output is $o = (D_1 * 0_d + D_2 * 0_d + 0_d)$
- Fusion structures
  1. Logistic regression
  2. Adding fully connected layer
  3. Domain attentive fusion using output
  4. Domain attentive fusion using hidden layer activation

Performance evaluation on multiple domains

- The proposed approach was verified in conditions where the test set domain was seen and unseen when training a network.
- The traditional approach shows reasonable performance only if the input domain is known a priori.
- Neural network based fusion approaches learn how to fuse networks from the training set and automatically calculate the weight of the network to contribute optimally for the random test input.
- The proposed approach performs consistently better on inputs from unknown domains compared to the traditional fusion approach.
  - Improved 16% in EER for seen domain input
  - Improved 4% in EER for unseen domain input

Conclusion

- The traditional fusion strategy performs optimally when the input domain is known.
- Training single network with multiple domain does not show best result on each domain.
- We propose a neural network based self-attention fusion approach.
- A domain attentive layer automatically decides the weights of multiple dialect ID systems by looking at the output or the embedding layer of each system.
- The performance on each test set from different domains is even better than the single domain optimized fusion approach.

Acknowledgements: This research was carried out in collaboration between the HBKU Qatar Computing Research Institute (QCR) and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).