Cross-Lingual Unsupervised Sentiment Classification with Multi-View Transfer Learning

Hongliang Fe, Ping li
ACL 2020
Sentiment Classification

- Standard supervised classification problem
  - Categorize a document/sentence into one out of a set of discrete classes
  - Sentiment class types
    - Binary
      - Positive/negative
    - Multi-class
      - Positive, neutral, negative
      - 5-star ratings
Unsupervised Cross-lingual

• Motivation
  – Scarcity of data
    • High-resource vs low-resource languages
    • Scarcity of labeled data
  – Cross-lingual text classification
Usual limitations

- Reliance on parallel resources
  - Bilingual dictionary
  - Parallel text, Wikipedia articles

- “Zero resource” models
  - No parallel corpus needed
  - Reliance on *Bilingual Word Embeddings (BWE)*
    - Poor performance on distant language pairs
Transformer-based approaches

- Cross-lingual Language Models
  - Trained from raw Wikipedia texts
    - Multilingual BERT
    - XLM
  - Fine-tuned alongside a classifier layer
  - No requirement for target language labeled data
- Issues
  - May not capture semantic similarity
Proposed approach

- Multi-View Encoder-Classifier (MVEC)
  - Monolingual corpora from two languages \( \{ D_{src}, D_{tgt} \} \)
  - Labeled data in source language \( \{ D_{src}^L, y_{src}^L \} \) \( D_{src}^L \subset D_{src} \)
  - Key differences
    - Unsupervised Machine Translation (UMT)
    - Language discriminator
    - Multi-view learning
      - Simultaneously learn multiple representations
      - Complementary information
      - Exploit consistency
Proposed approach

- Refined language-invariant latent space
  - More effective for reconstruction by the decoder
  - More robust to language shift
  - Better model interpretation

- Evaluation
  - 5 language pairs and 11 tasks
  - SOTA performance in 8/11
Model
Encoder-Decoder

- **Within-domain**

\[
R_{wd}(\theta_{ed}, l) = \mathbb{E}_{x \sim \mathcal{D}_l, \hat{x} \sim d(e(G(x)))}[\Delta(x, \hat{x})]
\]

\[
\theta_{ed} = [\theta_{enc}, \theta_{dec}], \quad l \in \{src, tgt\}
\]

- **Cross-domain**

\[
R_{cd}(\theta_{ed}, l_1, l_2) = \mathbb{E}_{x \sim \mathcal{D}_{l_1}, \hat{x} \sim d(e(T(x)))}[\Delta(x, \hat{x})]
\]

\[
(l_1, l_2) \in \{(src, tgt), (tgt, src)\}
\]
Language Discriminator

- Discriminator loss

\[
L_D(\theta_D|\theta_{enc}) = -\mathbb{E}_{(l,x(l))}[\log P_D(l|e(x(l)))]
\]

- Adversarial loss

\[
L_{adv}(\theta_{enc}|\theta_D) = -\mathbb{E}_{x(l_i)\sim D_l,}[\log P_D(l_j|e(x(l_i)))]
\]

\[l_j = l_1 \text{ if } l_i = l_2, \text{ and vice versa.}\]
Multi-view Classifier

- 2 views
  - Encoded sentences in the source language
  - Encoded translations of source sentences in target language

\[
L_C(\theta_C, \theta_{ed}) = \mathbb{E}_{(x,y)}[\Delta(y, P_{\theta_C}(e(x))) + D_{KL}(P_{\theta_C}(e(x)) \| P_{\theta_C}(e(T(x))))]
\]

Two views' consensus
Training Algorithm

\[ L_{all} = L_C + \lambda_{wd} \times (R_{wd\_src} + R_{wd\_tgt}) + \lambda_{cd} \times (R_{cd\_src} + R_{cd\_tgt}) + \lambda_{adv} \times L_{adv} \]

**Algorithm 1** The proposed MVEC algorithm.

1. **procedure** TRAINING($D_{src}$, $D_{tgt}$, $y_{src}^L$)
   \(D_{src}\) and \(D_{tgt}\): monolingual datasets, \(y_{src}^L\): labels in the source language.
2. \(T^{(0)} \leftarrow\) pretrain a transformer based UMT using (Conneau and Lample, 2019);
3. **for** \(t = 0, \ldots, max\_epoch\) **do**
4. \hspace{1em} Using \(T^{(t)}\) to translate each document in a batch;
5. \hspace{2em} \(\theta_D \leftarrow\) argmin \(L_D\) in Eq. (3) while fixing \(\theta_C, \theta_{ed}\);
6. \hspace{2em} \(\theta_C, \theta_{ed} \leftarrow\) argmin \(L_{all}\) in Eq. (6) while fixing \(\theta_D\);
7. \hspace{2em} Update \(T^{(t+1)} \leftarrow\) \{\(e(t), d(t)\}\};
8. **return** \(\theta_C, \theta_{enc}\)
9. **End procedure**
## Results

<table>
<thead>
<tr>
<th>Approach</th>
<th>German (2)</th>
<th>French (2)</th>
<th>Japanese (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>books</td>
<td>DVD</td>
<td>music</td>
</tr>
<tr>
<td><strong>With cross-lingual resources</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR+MT</td>
<td>79.68</td>
<td>77.92</td>
<td>77.22</td>
</tr>
<tr>
<td>CR-RL</td>
<td>79.89</td>
<td>77.14</td>
<td>77.27</td>
</tr>
<tr>
<td>Bi-PV</td>
<td>79.51</td>
<td>78.60</td>
<td>82.45</td>
</tr>
<tr>
<td>CLDFA</td>
<td>83.95</td>
<td>83.14</td>
<td>79.02</td>
</tr>
<tr>
<td><strong>With implicit cross-lingual resources</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UMM</td>
<td>81.65</td>
<td>81.27</td>
<td>81.32</td>
</tr>
<tr>
<td>PBLM</td>
<td>78.65</td>
<td>79.90</td>
<td>80.10</td>
</tr>
<tr>
<td><strong>Without cross-lingual resources</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BWE (1-to-1)</td>
<td>76.00</td>
<td>76.30</td>
<td>73.50</td>
</tr>
<tr>
<td>BWE (3-to-1)</td>
<td>78.35</td>
<td>77.45</td>
<td>76.70</td>
</tr>
<tr>
<td>MAN-MoE</td>
<td>82.40</td>
<td>78.80</td>
<td>77.15</td>
</tr>
<tr>
<td>MBERT</td>
<td>84.35</td>
<td>82.85</td>
<td>83.85</td>
</tr>
<tr>
<td>XLM-FT</td>
<td>86.85</td>
<td>84.20</td>
<td>85.90</td>
</tr>
<tr>
<td>MVEC (Ours)</td>
<td><strong>88.41</strong></td>
<td><strong>87.32</strong></td>
<td><strong>89.97</strong></td>
</tr>
</tbody>
</table>
### Results/Ablation

<table>
<thead>
<tr>
<th>Approach</th>
<th>Chinese (5)</th>
<th>Arabic (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR+MT</td>
<td>34.01</td>
<td>51.67</td>
</tr>
<tr>
<td>DAN</td>
<td>29.11</td>
<td>48.00</td>
</tr>
<tr>
<td>mSDA</td>
<td>31.44</td>
<td>48.33</td>
</tr>
<tr>
<td>ADAN</td>
<td>42.49</td>
<td><strong>52.54</strong></td>
</tr>
<tr>
<td>MBERT</td>
<td>38.85</td>
<td>50.40</td>
</tr>
<tr>
<td>XLM-FT</td>
<td>42.22</td>
<td>49.50</td>
</tr>
<tr>
<td>MVEC (Ours)</td>
<td><strong>43.36</strong></td>
<td>49.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>German</th>
<th>French</th>
<th>Japanese</th>
<th>Chinese</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model:</td>
<td><strong>88.61</strong></td>
<td><strong>88.62</strong></td>
<td><strong>78.67</strong></td>
<td><strong>43.36</strong></td>
<td><strong>49.70</strong></td>
</tr>
<tr>
<td>w/o cross-domain loss:</td>
<td>83.22</td>
<td>82.40</td>
<td>72.05</td>
<td>35.74</td>
<td>42.80</td>
</tr>
<tr>
<td>w/o within-domain loss:</td>
<td>82.90</td>
<td>82.15</td>
<td>71.27</td>
<td>37.21</td>
<td>41.60</td>
</tr>
<tr>
<td>w/o adversarial training:</td>
<td>84.85</td>
<td>84.58</td>
<td>73.75</td>
<td>39.36</td>
<td>46.37</td>
</tr>
<tr>
<td>w/o two-views consensus:</td>
<td>86.21</td>
<td>86.18</td>
<td>75.25</td>
<td>40.95</td>
<td>46.77</td>
</tr>
</tbody>
</table>