Multi-Domain Neural Machine Translation with Word Level Adaptive Layer-wise Domain Mixing

Haoming Jiang*, Chen Liang*, Chong Wang*, Tuo Zhao*

*Georgia Tech *ByteDance (Seattle)

**Objective**

Neural Machine Translation (NMT) Encoder Decoder

- **Motivation**: Handle data from multiple domains by sharing knowledge (Haddow et al. 2012).
- **Challenges**: Enforcing knowledge sharing lacks adaptivity to each individual domain.
- **Example**: Failure to handle word-level ambiguity across domains: The word "articles" has different meanings in laws and media domains.

**Background**

- **Recurrent Network based Encoder-Decoder**
  1. Computationally Expensive (Recursive Nature)
  2. Fail to Capture Long-term Dependency
  3. Various Training Issues (e.g. Gradient Exploding/Vanishing)
- **Transformer Models** (Vaswani et al. 2017)
  - Feedforward Network based Encoder-Decoder
  - Transformer Models

**Proposed Method**

- **Word-Level Domain Proportion**
  \[ D(x) = \frac{1 - \epsilon}{\text{softmax}(Rx) + \epsilon/k} \]
  - \( \epsilon \in (0, 1) \) : Smoothing parameter.
  - \( k \) : Number of domains.
  - \( x \) : Word vector.
  - \( R \) : Weight matrix of the softmax layer.
- **Word Level Adaptive Layer-wise Domain Mixing**

**Experiment – Domain Proportion**

- **Domain Proportion Visualization:**
  T2D Domain (white) vs. Medical Domain (black)

- **Word-Level Analysis:**
  Top Layer: The phrase is well understood and finding has little need to borrow domain knowledge.
  Bottom Layer: The phrase is domain specific.

- **Histograms of the Domain Proportions**
  Within each histogram, 0 means Medical domain, and 1 means T2D domain.

**Experiment – Translation**

- **Training Perplexity:**
  - **Testing BLEU Scores:**
    - **English to German**
      - Direct Training
        - News 40.60 55.88
        - TED 40.43 55.85
        - News + TED 40.52 55.82
    - **English to French**
      - Direct Training
        - News 40.60 54.39
        - TED 40.60 54.30
  - **Chinese to English**
    - Direct Training
      - News 53.98 3.80
      - TED 54.93 3.80
      - Speech 53.88 3.80
      - Thaw 53.80 3.80
      - Mixed 48.97 3.80
    - **Chinese to French**
      - Direct Training
      - News 40.43 54.14
      - TED 40.43 54.14
      - Speech 40.40 54.14
      - Thaw 40.38 54.14
      - Mixed 40.35 54.14

- **Benefits of Transformer Models**
  1. High Efficiency (Parallel and Feedforward Structures)
  2. Capture Long-term Dependency (Attention Module)
  3. Enable Deeper Representation Learning

**Diagram**

- Encoder Output
- Decoder Output
- Domain Proportion
- Word Level Mixing
- Multi-head Attention
- Forward/Backward
- Attention
- Softmax
- Value Key
- Linear Combination
- Point-wise Linear
- Word vector
- Input
- Layer-1 2 3 4 5 6
- Epoches 0 10 20 30 40 50
- Perplexity News+TED Thes MTL AdvL PAdvL WDC w/ WL
- Testing BLEU Scores English to German English to French Chinese to English

**Table**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Training BLEU Scores</th>
<th>Chinese to French</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>(40.60, 55.88)</td>
<td>(40.43, 54.14)</td>
</tr>
<tr>
<td>TED</td>
<td>(40.43, 55.85)</td>
<td>(40.43, 54.14)</td>
</tr>
<tr>
<td>News + TED</td>
<td>(40.52, 55.82)</td>
<td>(40.38, 54.14)</td>
</tr>
<tr>
<td>Speech</td>
<td>(53.88, 3.80)</td>
<td>(53.88, 3.80)</td>
</tr>
<tr>
<td>Thaw</td>
<td>(53.80, 3.80)</td>
<td>(53.80, 3.80)</td>
</tr>
<tr>
<td>Mixed</td>
<td>(48.97, 3.80)</td>
<td>(48.97, 3.80)</td>
</tr>
<tr>
<td>News</td>
<td>(53.98, 3.80)</td>
<td>(53.98, 3.80)</td>
</tr>
<tr>
<td>TED</td>
<td>(54.93, 3.80)</td>
<td>(54.93, 3.80)</td>
</tr>
<tr>
<td>Speech</td>
<td>(53.88, 3.80)</td>
<td>(53.88, 3.80)</td>
</tr>
<tr>
<td>Thaw</td>
<td>(53.80, 3.80)</td>
<td>(53.80, 3.80)</td>
</tr>
<tr>
<td>Mixed</td>
<td>(48.97, 3.80)</td>
<td>(48.97, 3.80)</td>
</tr>
</tbody>
</table>