Approaching Sign Language Gloss Translation as a Low-Resource Machine Translation Task

AT4SSL @ MTSummit2021

Xuan Zhang

Kevin Duh

Department of Computer Science, Johns Hopkins University
Continuous Sign Language Recognition

Neural Machine Translation

When you were a little girl, what did you want to do when you grow up?
This work
Neural Machine Translation (NMT)

INPUT: je suis étudiant

ENCODERS

DECODERS

OUTPUT: I am a student
Neural Machine Translation (NMT)

**INPUT**

je suis étudiant

**OUTPUT**

I am a student

**ENCODERS**

**DECODERS**

**this work:**

**INPUT**

sign glosses

**OUTPUT**

text translation
Challenges of NMT in SLT

- **NMT systems are data-hungry** and usually require millions of training examples to obtain a good translation performance.

- The publicly available annotated parallel **sign gloss–text data are scarce**. The popular continuous SLT dataset, “RWTH-PHOENIX-Weather2014” contains only 7,096 gloss-text examples in training set.

**Our Solution:**
Approaching gloss-text translation as a **Low-Resource** NMT Task
Low-resource NMT Techniques

Hyperparameter Search
Hyperparameter Search is crucial especially for the low-resource scenarios.

![Figure*](Wide variance in performance of a low-resource NMT system with different hyperparameter settings.)

Back-translation
Back-translation leverages the abundant monolingual data to enhance the translation performance.

*BLEU: The most commonly used scoring tool to evaluate the machine translation performance. Higher is better.
*Figure borrowed from *Reproducible and Efficient Benchmarks for Hyperparameter Optimization of Neural Machine Translation Systems, Zhang and Duh, TACL 2020.*
Experiment Setup

**Parallel Data:** RWTH-PHOENIX-Weather 2014T
- records the weather forecast airings of the German public tv-station PHOENIX.
- a continuous SLT corpus with both gloss annotations and German translations.
- train/dev/test: 7,096/519/642 sentences
- vocabulary size: gloss 1,066; German translation 2,887

**Monolingual Data** (for back-translation): TED Talks
- German TED Talk subtitles
- contains 151,627 sentences

**NMT model:** Transformer
Experiment: Hyperparameter Search

Search space:

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPE* merge operations</td>
<td>1k, 2k</td>
</tr>
<tr>
<td>Number of layers</td>
<td>1, 2, 4</td>
</tr>
<tr>
<td>Embedding size</td>
<td>256, 512</td>
</tr>
<tr>
<td>Initial learning rate</td>
<td>5e-5, 2e-4, 5e-4</td>
</tr>
</tbody>
</table>

BPE: a word segmentation approach that combines frequent sequence of characters so that out-of-vocabulary words are handled.
Experiment: Hyperparameter Search

Search space:

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPE merge operations</td>
<td>1k, 2k</td>
</tr>
<tr>
<td>Number of layers</td>
<td>1, 2, 4</td>
</tr>
<tr>
<td>Embedding size</td>
<td>256, 512</td>
</tr>
<tr>
<td>Initial learning rate</td>
<td>5e-5, 2e-4, 5e-4</td>
</tr>
</tbody>
</table>

Compare to existing work:

<table>
<thead>
<tr>
<th>Gloss-text system</th>
<th>Best BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>This work</td>
<td>24.38</td>
</tr>
<tr>
<td>Camagoz et al. (2020)</td>
<td>24.54</td>
</tr>
<tr>
<td>Yin and Read (2020)</td>
<td>24.9</td>
</tr>
</tbody>
</table>

gloss-text

max-min: 2.65 BLEU

text-gloss

max-min: 3.39 BLEU
Experiment: Back-translation

Back-translation workflow:

1. Train PHOENIX Text and PHOENIX Gloss to create a Text-Gloss Model.
2. Inference: Use the Text-Gloss Model to generate Ted Gloss from TED Text.
3. Train Gloss Text Model using Ted Gloss, TED Text, PHOENIX Gloss, and PHOENIX Text.
One issue: PHOENIX14T and TED data are from different domains, which makes the translation task more challenging.

Domain adaptation leverages out-of-domain data to improve the domain-specific translation. We adopt two domain adaptation methods to aid back-translation.
One issue: PHOENIX14T and TED data are from different domains, which makes the translation task more challenging.

**Domain adaptation** leverages out-of-domain data to improve the domain-specific translation. We adopt two domain adaptation methods to aid back-translation.

- **Data Selection**
- **Fine-tuning**
**Experiment: Back-translation**

**Domain Adaptation – Data Selection**

\[ I: \text{in-domain data (PHEONIX14T)} \]
\[ N: \text{out-of-domain data (TED Talks)} \]

**Goal:** select top \( n \) training examples from \( N \) which are most similar to \( I \).
Experiment: Back-translation
Domain Adaptation – Data Selection

$I$: in-domain data (PHEONIX14T)
$N$: out-of-domain data (TED Talks)

**Goal:** select top $n$ training examples from $N$ which are most similar to $I$.

Each sentence $s$ in $N$ is assigned a score:

$$H_I(s) - H_N(s)$$

$H_I(s)$: the per-word cross-entropy of $s$ according to a language model trained on $I$. 
Experiment: Back-translation
Domain Adaptation – Fine-tuning

1. First, train on the mix of out-of-domain and in-domain data till convergence.

2. Then, continue training or fine-tuning on the in-domain data.
Experiment: Back-translation

Back-translation workflow review:

1. **Train**
   - PHOENIX Text
   - PHOENIX Gloss
   - Text-Gloss Model

2. **Inference**
   - TED Text
   - Text-Gloss Model
   - Ted Gloss

3. **Train**
   - TED Gloss + PHOENIX Gloss
   - TED Text + PHOENIX Text
   - Gloss Text Model
Experiment: Back-translation

Back-translation + data selection workflow:
Experiment: Back-translation

Back-translation + data selection + fine-tuning workflow:

1. Train
   - PHOENIX Text
   - PHOENIX Gloss
   - Text-Gloss Model

2. Inference
   - TED Text
   - Data selection
   - Text-Gloss Model
   - Ted Gloss

3. Train
   - Selected TED Gloss + PHOENIX Gloss
   - Selected TED Text + PHOENIX Text
   - Gloss-Text Model

4. Fine-tuning
   - PHOENIX Text
   - PHOENIX Gloss
   - Fine-tuned Gloss-Text Model
Figure: Performance of NMT systems on gloss-text translation. Systems vary in whether fine-tuned on PHOENIX14T (bt vs. ft), the size of selected TED data (10k, 50k, 100k) and the number of copies of PHOENIX14T added into the training data (0, 1, 2, 5, 10).
Experiment: Back-translation

**issue:** The best BLEU score of back-translation + domain adaptation is slightly lower than that of hyperparameter search (24.31 vs. 24.38).

**possible cause:** the domain adaptation techniques fail to overcome the side-effect of introducing the out-of-domain data into training.
issue: The best BLEU score of back-translation + domain adaptation is slightly lower than that of hyperparameter search (24.31 vs. 24.38).

possible cause: the domain adaptation techniques fail to overcome the side-effect of introducing the out-of-domain data into training.

We thus consider a simpler situation in order to evaluate back-translation: incorporating in-domain monolingual data.

- Divide PHOENIX14T into two sets, with the first half acting as parallel data, the second as monolingual data containing only the German text side of the data.
Experiment: Back-translation  
incorporating in-domain monolingual data

<table>
<thead>
<tr>
<th>NMT Systems</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o back-translation</td>
<td>19.13</td>
</tr>
<tr>
<td>w/ back-translation</td>
<td>21.57</td>
</tr>
</tbody>
</table>

Back-translation has a good potential to improve the translation performance when in-domain monolingual data are available.
The translation between sign language glosses and written languages is a challenging task in SLT.

The obstacle lies in the sparsity of parallel data.

Approaching the translation task with low-resource MT techniques like hyperparameter search and back-translation is promising.

Back-translation is more likely to contribute when abundant in-domain monolingual data are available.

We urge the sign language processing community to put in extra efforts in creating more annotated parallel data.
Approaching Sign Language Gloss Translation as a Low-Resource Machine Translation Task

Q & A