Contrastive Learning for Many-to-many Multilingual Neural Machine Translation

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li
Outline

• Motivation and Goal
• mRASP2 Methodology
• Experiments and Analysis
  – Supervised / Unsupervised / Zero-shot
  – Better alignment
• Summary and Take-away
The Ultimate Quest of Machine Translation

• # of human languages: >6900.
• How to build a universal MT system that is capable of translating any source language into a target one?
Why Training Multilingual MT Jointly?

- Data scarcity for low/zero resource languages.
Many-to-many Multilingual NMT

Training only w/ En-X Corpus

Many-to-many MNMT
Existing Multilingual NMT (1)

- **Supervised**
  - ✔️ En-Zh, En-Fr, En-De

- **Unsupervised**
  - ✗ Fr-Zh, Fr-De, De-Zh

- **Zero-shot**
  - En-Pt (Assume only have monolingual data of Pt)

Severe degradation on zero-shot translation
- M Johnson, 2017
- N Arivazhagan, 2019
Existing Multilingual NMT(2)

Degradation on high-resource directions

Arivazhagan et al. 2019
Existing Multilingual NMT(3)

- Parallel: Only use parallel data
- Monolingual: Not allowed
We want ….

- Supervised
- Unsupervised
- Zero-shot

Enabling unsupervised / zero-shot translation

- Parallel
- Monolingual

Leveraging both parallel & monolingual data

Comparable / better performance on high-resource directions

Arivazhagan et al. 2019
Goal of mRASP2

• Build a universal NMT model that is both
  • A unified multilingual NMT model that support complete many-to-many translation.
  • A ready-to-use model from which we can derive any NMT model for specific translation direction
Intuition of mRASP2: Bring Representation Closer

- Sentences with the same semantics across different languages should have similar representations.
Outline

• Motivation and Goal
• mRASP2 Methodology
• Experiments and Analysis
  – Supervised / Unsupervised / Zero-shot
  – Better alignment
• Summary and Take-away
**mRASP**

Pre-training

<table>
<thead>
<tr>
<th>Orig tok</th>
<th>I</th>
<th>like</th>
<th>singing</th>
<th>and</th>
<th>dancing</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RAS tok</th>
<th>I</th>
<th>like</th>
<th>chanter</th>
<th>and</th>
<th>danser</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Encoder

Decoder

<table>
<thead>
<tr>
<th>FR id</th>
<th>J'adore</th>
<th>chanter</th>
<th>et</th>
<th>danser</th>
<th>&lt;EOS&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Random Aligned Substitution (RAS)

- **En**: I love you.
- **Fr**: Je t'aime.
- **De**: Ich liebe dich.
- **Es**: Te quiero.
- **It**: ti amo.

Z Lin · 2020
Seq2seq Training with Aligned Augmentation

• Parallel text

• Monolingual text
mRASP2 Training

Contrastive Loss: $L_{ctr}$

Cross Entropy Loss: $L_{ce}$

\[
\mathcal{L} = L_{ce} + \lambda |s| L_{ctr}
\]

Encoder

Decoder

Negative

Positive

Anchor

<Fr> Je t’aime.

<En> It’s sunny.

<Fr> C’est la vie.

......

<En> I love you.

<Fr> Je t’aime.

<Zh> 你是谁
Contrastive Learning

\[ \mathcal{L}_{\text{ctr}} = - \sum_{x_i, x_j \in \mathcal{D}} \log \frac{e^{\text{sim}^+(\mathcal{R}(x^i), \mathcal{R}(x^j))/\tau}}{\sum_{y_j} e^{\text{sim}^-(\mathcal{R}(x^i), \mathcal{R}(y^j))/\tau}} \]

\[ \mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda |s| \mathcal{L}_{\text{ctr}} \]
Outline

• Motivation and Goal
• mRASP Methodology
• Experiments and Analysis
  – Supervised / Unsupervised / Zero-shot
  – Better alignment
• Summary and Take-away
Two Main Questions

- Does mRASP2 work on supervised / unsupervised / zero-shot scenarios?
- Why mRASP2 works?
Datasets

- Parallel Dataset: **PC32** (32 language pairs)
  - 32 English-centric language pairs, resulting in 64 directed translation pairs in total
  - Contains a total size of 110.4M public parallel sentence pairs
• Monolingual Dataset: **MC24** (24 languages)
  – 21 languages that also appear in **PC32**
  – 3 additional languages: NL, PL, PT
  – Temperature sampling: $T=5$
Overall Results

Overall Results in all scenarios: 56 directions

- **Averaged (ALL)**
  - m-Transformer: 13.41
  - mRASP2: 21.03
Comparable or Better Performance on Supervised Directions

Tokenized BLEU on supervised directions
Effectiveness on Zero-shot Directions

Averaged De-tokenized BLEU on zero-shot directions

- Pivot (m-Transformer)
- m-Transformer
- mRASP2

mRASP2 effectively improves zero-shot translation

Fr->Zh: 6.5 —> 42.3
Effectiveness on Unsupervised Directions

Averaged Tokenized BLEU on En-X
unsupervised directions

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg(En-&gt;Ni)</th>
<th>Avg(Ni-&gt;Pt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>m-Transformer</td>
<td>4.42</td>
<td>5.4</td>
</tr>
<tr>
<td>mRASP (w/o ft)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRASP2</td>
<td>18.55</td>
<td>9.3</td>
</tr>
</tbody>
</table>

mRASP2 also works on fully unsupervised directions
Contrastive Learning effectively improves zero-shot translation without hurting supervised translation performance.
Contrastive Learning effectively improves zero-shot translation without hurting supervised translation performance.
Monolingual Corpus mainly contributes to unsupervised translation

- m-Transformer
- mRASP (w/o ft)
- mRASP2 w/o AA
- mRASP2 w/o MC24

Legend:
- Supervised
- Unsupervised
- Zero-shot

Bar chart comparing different models across supervised, unsupervised, and zero-shot scenarios.
Better Semantic Alignment Across Languages: Improved Sentence Retrieval

Contrastive Learning and Aligned Augmentation both contribute to the improvement on sentence retrieval.

15-way parallel test set (Ted-M): 2284 samples
Better Semantic Alignment: Visualization of Sentence Repr

Better Alignment of En, Ja, De Representations !!
Outline

• Motivation and Goal
• mRASP2 Methodology
• Experiments and Analysis
  – Supervised / Unsupervised / Zero-shot
  – Better alignment
• Summary and Take-away
We propose mRASP2

- A universal Multilingual MT model
- Leverages monolingual data along with parallel data in a unified framework
- Bridges the representation gap of utterances in different languages with the same semantics.
Take Home Messages

• Closer representation —> Improved multilingual MT performance
• Leverage both parallel and monolingual corpora !!
• Contrastive Learning and Aligned Augmentation are effective in bridging representation
Thanks!

• Code and models available at:
  – https://github.com/PANXiao1994/mRASP2

• Also MT in ACL21:
  – Green vocabulary learning: VOLT [Xu et al. 2021]
  – Language-specific subnets for MNMT: LaSS [Lin et al. 2021]
  – Language Tag Matters [Wu et al. 2021]
  – Glancing Transformer [Qian et al. 2021]

• Other tools:
  – Transformer fast training and inference: https://github.com/bytedance/lightseq
  – Speech & MT toolkit: https://github.com/bytedance/neurst