On Learning Language-Invariant Representations for Universal Machine Translation

Han Zhao, Junjie Hu, Andrej Risteski
{han.zhao, junjieh, aristesk}@cs.cmu.edu

Carnegie Mellon University
Recent Success of Neural Machine Translation

Machine Translation on WMT2014 English-French

Machine Translation, ~3M parallel sentences [Cho et al. 2014; Devlin et al. 2014]
## Neural Machine Translation is Data Hungry

![Graph showing BLEU scores vs. training examples]

**Figure from [Gu et al. 18]**

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Corpora size</th>
<th>BLEU Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>French</td>
<td>~3M</td>
<td>~40</td>
</tr>
<tr>
<td>English</td>
<td>German</td>
<td>~1.92M</td>
<td>~35</td>
</tr>
<tr>
<td>Finnish</td>
<td>English</td>
<td>~1.96M</td>
<td>~34</td>
</tr>
<tr>
<td>Romanian</td>
<td>English</td>
<td>~400K</td>
<td>~30</td>
</tr>
</tbody>
</table>

WMT ’16-19, Europarl Parallel Corpus
Typical Pipeline of Multilingual Machine Translation

- **Separate MT systems**: Hard to maintain all systems
- **Pivot methods**: src-to-pivot & pivot-to-tgt translations

Machine translation by triangulation: Making effective use of multi-parallel corpora, [Cohn et al 07]
**Cross-Lingual Representations by Neural Models**

- **Language similarity**: similar words, grammar, order.
- **Shared space**: learning word/sentence representations jointly

Why Universal Machine Translation (UMT)?

- **Single model**: many-to-one, one-to-many
- **Zero-shot translation**: improve low-resource translation

Recent Advances of UMT

- **Language coverage**: 100+ languages in Google’s M4
- **Web-mined data**: 25 billion examples
- **Quality**: +5 BLEU score over all 100+ languages

Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges [Arivazhagan et al. 19]
Challenge: Theoretical Understanding of UMT

Despite the empirical success, theoretical understanding is only nascent

- **Translation Error**: Is there a performance limit even with unlimited amount of computation & data

- **Sample Complexity**: How many language pairs are required to train UMT?
Challenge: Theoretical Understanding of UMT

Despite the empirical success, theoretical understanding is only nascent

- Translation Error: Is there a performance limit even with unlimited amount of computation & data
  - Without assumption on the parallel corpus used for training, at least one translation task has to incur a large error

- Sample Complexity: How many language pairs are required to train UMT?
  - Under an encoder-decoder generative assumption of the data, a linear number of translation pairs suffice for the purpose of UMT
A Theoretical Model for UMT

Let $\mathcal{L} = \{\text{English, French, German, Chinese, …}\}$ be the set of all languages of interest.

- For each $L \in \mathcal{L}$, we associate with $L$ an alphabet $\Sigma_L$
- A sentence $x$ in $L$ is a sequence of symbols from $\Sigma_L$, i.e., $x \in \Sigma_L^*$
- For a pair of languages $L, L'$, we use $\mathcal{D}_{L,L'}$ to denote the joint distribution over the parallel sentence pairs from $L$ and $L'$
A Theoretical Model for UMT

Problem Setting:

- For each pair of languages $L, L'$, there exists a **true translator**

  $$f^*_L \rightarrow L' : \Sigma^*_L \rightarrow \Sigma^*_L,$$

- Given a translator $f$ from $L$ to $L'$, we use the 0-1 loss to measure the translation quality w.r.t. the true translator:

  $$\text{Err}^L_{L'} (f) := \mathbb{E}_D [\ell(f(X), f^*_L \rightarrow L'(X))]$$

  where $\ell(x, x') = 0$ iff $x = x'$.

There exists a perfect translator that translates input sentence from any language to a target language $L$:

$$f^*_L(x) = \sum_{L' \in \mathcal{L}} \mathbb{I}(x \in \Sigma^*_L') \cdot f^*_{L' \rightarrow L}(x)$$

Can we recover the perfect translator through UMT?
Universal Machine Translation

Universal Language Mapping:

A function mapping \( g : \bigcup_{i \in [K]} \Sigma^* \rightarrow Z \) is called universal if

\[
g_i D_i = g_j D_j, \forall i \neq j
\]

Different languages have the same distribution under representation Z.
An Impossibility Theorem

A simple warm-up (Two-to-One):

Theorem (informal): Consider a restricted setting of universal machine translation task with two source languages and one target language. If $g$ is a universal language mapping, then for any decoder $h : \mathcal{Z} \rightarrow \Sigma^*_L$, 

$$\text{Err}_{D_0}^{L_0 \rightarrow L}(h \circ g) + \text{Err}_{D_1}^{L_1 \rightarrow L}(h \circ g) \geq d_{TV}(D_{L_0,L}(L), D_{L_1,L}(L)).$$

Translation errors from $L_0, L_1$ to $L$

Distance between sentence distributions over $L$

Uncertainty Principle: UMT has to make a large error on at least one translation task
An Impossibility Theorem

A simple warm-up (Two-to-One):

Theorem (informal): Consider a restricted setting of universal machine translation task with two source languages and one target language. If \( g \) is a universal language mapping, then for any decoder \( h : \mathcal{Z} \rightarrow \Sigma_L^* \),

\[
\text{Err}_{\mathcal{D}_0}^{L_0 \rightarrow L}(h \circ g) + \text{Err}_{\mathcal{D}_1}^{L_1 \rightarrow L}(h \circ g) \geq d_{\text{TV}}(\mathcal{D}_{L_0,L}(L), \mathcal{D}_{L_1,L}(L)).
\]

- This is an information-theoretic lower bound, i.e., algorithm-independent
- The theorem still holds even if we use different encoders for different languages, but wouldn’t hold any more if we use target-dependent decoder!
- The lower bound gets larger whenever target data are dissimilar between different translation tasks
An Impossibility Theorem

In general (Many-to-One):

Maximum Translation Error:

$$\max_{i \in [K]} \text{Err}^{L_i \to L}_{D_i}(h \circ g) \geq \frac{1}{2} \max_{i \neq j} E(i, j)$$

Average Translation Error:

$$\frac{1}{K} \sum_{i \in [K]} \text{Err}^{L_i \to L}_{D_i}(h \circ g) \geq \frac{1}{K(K-1)} \sum_{i < j} E(i, j)$$

$E(i, j)$ measures how different two translation tasks are:

$E(i, j) := d_{TV}(D_{L_i}, L(L), D_{L_j}, L(L))$
The impossibility theorem holds in the worst case without any assumption on the data generating distribution of parallel corpus. What if we assume an encoder-decoder generative process?

We assume that $E_k \in GL_d(\mathbb{R}), D_k = E_k^{-1}, \forall k \in [K]$
A Generative Model of UMT

Why this assumption on data generative process helps?

\[ \forall i, j \in [K], \quad \text{Err}_{D_i}^{L_i \rightarrow L}(h \circ g) + \text{Err}_{D_j}^{L_j \rightarrow L}(h \circ g) \geq d_{TV}(D_{L_i}, L(L), D_{L_j}, L(L)) = 0 \]

The lower bound still holds, but it gracefully reduces to 0 under this encoder-decoder assumption on data generative process.
How Many Language Pairs Suffices?

Naively, one might think we need $\Omega(K^2)$ language pairs, one for each pair.

Our result: under some mild assumptions, only a linear number $O(K)$ of translation pairs suffices!

Translation Graph: $H$

- Each node = a language
- Two nodes are connected if we see the corresponding translation pair
- $H$ is assumed to be connected: we need to see every language at least once
- The diameter of $H$ is bounded by $K$
How Many Language Pairs Suffices?

Translation Graph: $H$
- Each node = a language
- Two nodes are connected if we see the corresponding translation pair
- $H$ is assumed to be connected: we need to see every language at least once
- The diameter of $H$ is bounded by $K$

Theorem (informal): Let $\text{diam}(H)$ be the diameter of the translation graph $H$, then for any pair of language $L_i, L_j$, the translation error has the following upper bound:

$$\varepsilon(\hat{E}_i, \hat{E}_j) \leq \rho \cdot \text{diam}(H) \cdot \epsilon^2$$
How Many Language Pairs Suffices?

Theorem (informal): Let $\text{diam}(H)$ be the diameter of the translation graph $H$, then for any pair of language $L_i, L_j$, the translation error has the following upper bound:

$$\varepsilon(\hat{E}_i, \hat{E}_j) \leq \rho \cdot \text{diam}(H) \cdot \epsilon^2$$

- $\varepsilon(\hat{E}_i, \hat{E}_j)$ is measured w.r.t. the ground truth encoder-decoder
- $\rho$ is the Lipschitz-constant of the ground truth encoder and decoder
- $\epsilon$ is the maximum error on each seen translation pair
- For a specified translation error $\epsilon$, a corpora containing $O(1/\epsilon^2)$ parallel sentences suffices

We use a epsilon-net argument to prove this result.
Summary

Without data generating assumption: An Impossibility Theorem, UMT has to incur a large error on at least one translation pair.

Theorem (informal): Consider a restricted setting of universal machine translation task with two source languages and one target language. If $g$ is a universal language mapping, then for any decoder $h : \mathcal{Z} \rightarrow \Sigma_L^*$,

$$\text{Err}_{D_0}^{L_0 \rightarrow L}(h \circ g) + \text{Err}_{D_1}^{L_1 \rightarrow L}(h \circ g) \geq d_{TV}(\mathcal{D}_{L_0,L}(L), \mathcal{D}_{L_1,L}(L)).$$

With a natural data generating assumption: Linear number of translation pairs suffices!

Theorem (informal): Let $\text{diam}(H)$ be the diameter of the translation graph $H$, then for any pair of language $L_i, L_j$, the translation error has the following upper bound:

$$\varepsilon(\hat{E}_i, \hat{E}_j) \leq \rho \cdot \text{diam}(H) \cdot \epsilon^2$$