On Learning Language-Invariant Representations for Universal Machine Translation

AUTHORS
Han Zhao and **Andrej Risteski**

AFFILIATIONS
MLD, CMU

PUBLISHED
October 23, 2020

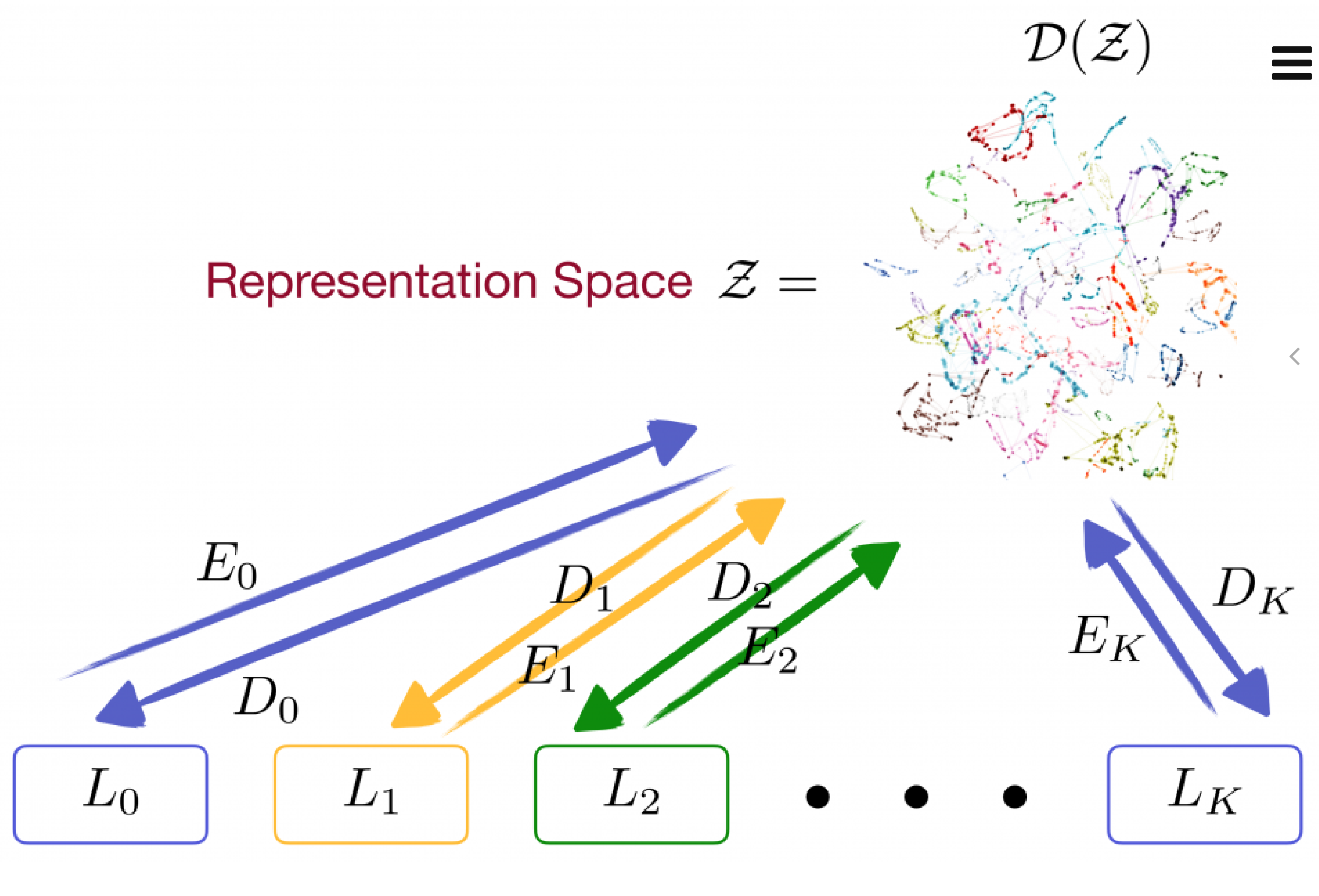
$\mathcal{D}(\mathcal{Z})$
Representation Space $\mathcal{Z} =$
E_0
D_0
D_1
E_1
D_2
E_2
D_K
E_K
L_0
L_1
L_2
L_K

Figure 1: An encoder-decoder generative model of translation pairs, which helps to circumvent the limitation discussed before. There is a global distrib h sentences of language L_i are generated via d ed via E_i to $\mathcal{Z}$.

Despite the recent improv MT), training a large NMT model with hundreds of mill ection of parallel corpora at a large scale, on the order of millions or even billions of aligned sentences for supervised tra g (Arivazhagan et al.). While it might be possible to automatically crawl the web to collect p sentences for high-resource language pairs such as German-English and French-English, it is *multilingual universal machine translation*, a.k.a. *universal machine translation* (UMT), is to learn to translate between any pair of languages using a single system, given pairs of translated documents for *some* of these languages. The hope is that by learning a shared "semantic space" between multiple source and target languages, the model can leverage language-invariant structure from high-resource translation pairs to transfer to the translation between low-resource language pairs, or even enable zero-shot translation.

Indeed, training such a single massively multilingual model has gained impressive empirical results, especially in the case of low-resource language pairs (see Fig. 2). However, such success also comes with a cost. From Fig. 2 we observe that the translation quality over high-resource language pairs by using such a single UMT system is worse than the corresponding bilingual baselines.

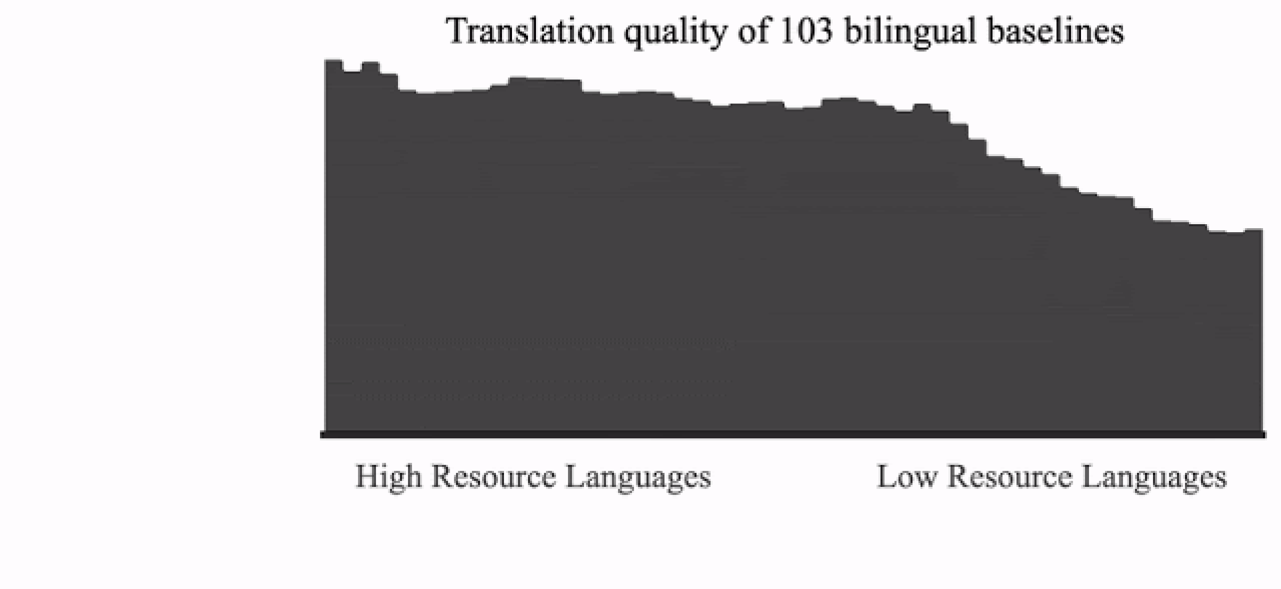

Figure 2: Translation quality by using a single massively multilingual model against bilingual baselines that are trained for each one of the 103 language pairs. While the translation performances over low resource languages increase, the performances over high resource languages decrease. Our work provides a theoretical explanation for this empirical phenomenon. Figure credit: Exploring Massively Multilingual, Massive Neural Machine Translation.

Is this empirical phenomenon by coincidence? If not, why does it happen? Furthermore, what kind of structural assumptions about languages could help us get over this detrimental effect? In this blog post, based on our recent ICML paper, we take the first step towards understanding universal machine translation by providing answers to the above questions. The key takeaways of this blog post could be summ

HOME
SUBMISSIONS
ABOUT

- In a completely *assu* shared representation, no matter what decode es, it is impossible to avoid making a large transla lation pairs.
- Under a natural *generative model* assumption for the data, after seeing aligned sentenc r a **linear** number of language pairs (**instead of quadratic!**), we can learn encoder/dec

An Impossibility Theorem on UMT via Language-Invariant Representations

Suppose we have an unlimited amount of parallel sentences for each pair of languages, with unbounded computational resources. Could we train a single model that performs well on all pairs of translation tasks based on a common representation space? Put it in other words, is there any information-theoretic limit of such systems for the task of UMT? In this paragraph we will show that there is an inherent tradeoff between the translation quality and the degree of representation invariance w.r.t.\ languages: the better the language invariance, the higher the cost on at least one of the translation pairs. At a high-level, this result holds due to the general data-processing principle: if a representation is invariant to multiple source languages, then any decoder based on this representation will have to generate the same language model on the target language. But on the other hand, the parallel corpora we use to train such a system could have drastically different sentence distributions on the target language, thus leading to a discrepancy (error) between the generated sentence distribution and the ground-truth sentence distribution over the target language.

To keep our discussions simple and transparent, let's start with a basic Two-to-One setup where there are only two source languages L_0 and L_1 and one target language L. Furthermore, for each source language $L_i, i \in \{0, 1\}$, let's assume that there *is* a perfect translator $f^*_{L_i \to L}$ that tak sentence (or string, sequence) from L_i and outputs the corresponding translation in L. Under setup, it is easy to see that there exists a perfect translator f^*_L in this Two-to-One task:

$$f^*_L(x) = \sum_{i \in \{0,1\}} \mathbb{I}(x \in L_i) \cdot f^*_{L_i \to L}(x)$$

In words: upon receiving a sentence x, f^*_L simply checks which source language x comes from and then call the corresponding ground-truth translator.

To make the idea of *language-invariant representations* formal, let $g : \Sigma^* \to \mathcal{Z}$ be an encoder that takes a sentence (string) from alphabet Σ to a representation in a vector space $\mathcal{Z}$. We call g an ϵ-*universal language mapping* if the distributions of sentence representations from different languages L_0 and L_1 are ϵ-close to each other. In words, $d(g_\sharp \mathcal{D}_0, g_\sharp \mathcal{D}_1) \leq \epsilon$ for some divergence measure d, where $g_\sharp \mathcal{D}_i$ is the induced distribution of sentence (from L_i) representations in the shared space $\mathcal{Z}$. Subsequen der h that takes a sentence representation z and outpu language L. The hope here is that z encodes the langu the input sentence (either from L_0 or from L_1) based c e L.

So far so good, but could we recover the perfect translator f^*_L by learning a common, sh d representation Z, i.e., ϵ is small? Unfortunately, the answer here is negative if we don't hav

> *Theorem (informal): Let $g : \Sigma^* \to \mathcal{Z}$ be an ϵ-universal language mapping. Then for any decoder $h : \mathcal{Z} \to \Sigma_L^*$, the following lower bound holds:*
>
> $$\mathrm{Err}_{\mathcal{D}_0}^{L_0 \to L}(h \circ g) + \mathrm{Err}_{\mathcal{D}_1}^{L_1 \to L}(h \circ g) \geq d(\mathcal{D}_0(L), \mathcal{D}_1(L)) - \epsilon.$$

Here the error term $\mathrm{Err}_{\mathcal{D}_i}^{L_i \to L}(h \circ g)$ measures the $0-1$ translation performance given by the encoder-decoder pair $h \circ g$ from L_i to L over distribution $\mathcal{D}_i$. The first term $d(\mathcal{D}_0(L), \mathcal{D}_1(L))$ in the lower bound measures the *difference of distributions over sentences from the target language in the two parallel corpora, i.e., $L_0 - L$ and $L_1 - L$.* For example, in many practical scenarios, it may happen that the parallel corpus of high-resource language pair, e.g., German-English, contains sentences over a diverse domain whereas as a comparison, the parallel corpus of low-resource language pair, e.g., Sinhala-English, only contains target translations from a specific domain, e.g., sports, news, product reviews, etc. In this case, despite the fact that the target is the same language L, the corresponding sentence distributions from English are quite different between different corpora, leading to a large lower bound. As a result, our theorem, which could be interpreted as a kind of *uncertainty principle in UMT*, says that no matter what kind of decoder we are going to use, it has to incur a large error on at least one of the translation pairs. It is also w pointing out that our lower bound is algorithm-independent and it holds even with unbour computation and data. As a final note, realize that for fixed distributions $\mathcal{D}_i, i \in \{0, 1\}$, the smaller the ϵ (hence the better the language-invariant representations), the larger the lower bound, demonstrating an inherent tradeoff between language-invariance and translation performance in general.

Proof Sketch: Here we provide a proof-by-picture (Fig. 3) in the special case of perfectly language-invariant representations, i.e., $\epsilon = 0$, to highlight the main idea in our proof of the above impossibility theorem. Please refer to our paper for more detailed proof as well as an extension of the above impossibility theorem in the more general many-to-many translation setting.

[Home]
[Submissions]
[About]

tasks in general have different marginal distribu get language, hence a triangle inequality over the output distributions gi desired lower bound.

How can we Bypass this Limitation?

One way is to allow the decoder h to have access to the input sentences (besides the language-invariant representations) during the decoding process — e.g. via an attention mechanism on the input level. Technically, such information flow from input sentences during decoding would break the Markov structure of "input-representation-output" in Fig. 3, which is an essential ingredient in the proof of our theorem. Intuitively, in this case both language-invariant (hence language-independent) and language-dependent information would be used.

Another way would be to assume extra structure on the distributions of our corpora $\mathcal{D}_i$, i.e., by assuming some natural generative process capturing the distribution of the parallel corpora that are used for training. Since languages share a lot of semantic and syntactic characteristics, this would make a lot of sense — and intuitively, this is what universal translation approaches are banking on. In the next paragraph, we will do exactly this — we will show that under a suitable generative model, not only will there be a language-invariant representation, but it will be learnable using corpora from a very small (linear) number of pairs of language.

A Generative Model for UMT: A Linear Number Translation Pairs Suffices!

In this section we will discuss a generative model, under which not only will there be a language-invariant representation, but it will be learnable using corpora from a very small (linear) number of pairs of language. Note that there are a quadratic number of translation pairs in our universe, hence our result shows that under this generative model zero-shot translation is actually possible.

To start with, what kind of generative model is suitable for the task of UMT? Ideally, we would like to have a feature space where vectors correspond to the semantic encoding of sentences from different languages. One could also understand it as a sort of "meaning" space. Then, language-dependent decoders would take these semantic vectors and decode them as the observable sentences. Figure 1 illustrates the generative process of our model, where we assume there is a common distribution $\mathcal{D}$ ove sentences are sampled and generated.

HOME
SUBMISSIONS
ABOUT

For ease of presentation, le er pair (E_i, D_i) consists of deterministic mappings (see ed encoders/decoders). The first question to ask is: how does this generative model assumption circumvent our previous lower bound in the last paragraph? We can easily observe that under the encoder-de generative assumption in Figure 1, the first term in our lower bound, $d(\mathcal{D}_0(L), \mathcal{D}_1(L))$, What's more interesting is that, under proper assumptions on the structure of $\mathcal{F}$, the class of encoders and decoders we learn from, by using the traditional empirical risk minimization (ERM) framework to learn the language-dependent encoders and decoders on a small number of language pairs, we could expect the learned encoders/decoders to well generalize on unseen language pairs as well! Informally,

> *Theorem (informal): Let H be a connected graph where each node L_i corresponds to a language and each edge (L_i, L_j) means that the learner has been trained on language pair L_i and L_j, with empirical translation error $\epsilon_{i,j}$ and corpus of size $\Omega(1/\epsilon_{i,j}^2 \cdot \log C(\mathcal{F}))$. Then with high probability, for any pair of language L and L' that are connected by a path $L = L_0, L_1, \ldots, L_m = L'$ in H, its population level translation error is upper bounded by $O(\sum_{k=0}^{m-1} \epsilon_{k,k+1})$.*

In the theorem above, $C(\mathcal{F})$ is some complexity measure of the class $\mathcal{F}$. If we slightly simplify the theorem above by defining $\epsilon := \max_{(L_i,L_j)\in H} \epsilon_{i,j}$ and realizing that the path length m is upper bounded by the diameter of the graph H, $\mathrm{diam}(H)$, we immediately obtain the follo intuitive result:

> *For any pair of languages L, L' (the parallel corpus between L and L' may not necessarily appear in our training corpora), the translation error between L and L' is upper bounded by $O(\mathrm{diam}(H) \cdot \epsilon)$.*

The above corollary says that graphs H that do not have long paths are preferable. For example, H could be a star graph, where a central (high-resource) language acts as a pivot node. The proof of the theorem above essentially boils down to two steps: first, we use an epsilon-net argument to show that the learned encoders/decoders generalize on a pair of language that appears in our training corpora, and then b ve apply a chain of triangle-like inequalities to bound th of languages.

Some Concludi

The prospect of building a single system for universal machine translation is appealing. Com with building a quadratic number of bilingual translators, such a single system is easier to train,

[Home]
[Submissions]
[About]

However, such promise often comes with a price, which calls for proper assumptions on the generative process of the parallel corpora used for training. Our paper takes a first step towards better understanding the tradeoff in this regard and proposes a simple setup that allows for zero-shot translation. On the other hand, there are still some gaps between theory and practice. For example, it would be interesting to see whether the BLEU score, a metric used in the empirical evaluation of translation quality, bears a similar kind of lower bound. Also, could we further extend our generative modeling of sentences so that there are more hierarchical structures in the semantic space $\mathcal{Z}$? Empirically, it would be interesting to implement the above generative model on synthetic data to see the actual performance of zero-shot translation under the model assumption. These challenging problems (and more) will require collaborative efforts from a wide range of research communities and we hope our initial efforts could inspire more efforts in bridging the gap.

Reference

1. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges, Arivazhagan et al., https://arxiv.org/abs/1907.05019.
2. Investigating Multilingual NMT Representations at Scale, Kudugunta et al., EMNLP 2019, https://arxiv.org/abs/1909.02197.
3. On Learning Language-Invariant Representations for Universal Machine Translation, Z et al., ICML 2020, https://arxiv.org/abs/2008.04510.
4. The Source-Target Domain Mismatch Problem in Machine Translation, Shen et al., https://arxiv.org/abs/1909.13151.
5. How multilingual is Multilingual BERT? Pires et al., ACL 2019, https://arxiv.org/abs/1906.01502.

DISCLAIMER: All opinions expressed in this post are those of the author and do not represent the views of CMU.

♡ 127 👁 9319

HOME
SUBMISSIONS
ABOUT

FACT Diagnostic: How g Group Fairness

Experiments with the I

[Home]
[Submissions]
[About]

An Inferential Perspective on Federated Learning

FEBRUARY 19, 2021

Learning DAGs with Continuous Optimization

APRIL 10, 2020

In defense of weight-sharing for neural architecture search: an optimization perspective

JULY 17, 2020

NO COMMENTS

LEAVE A REPLY

Name *

Email *

POST COMMENT

HOME

SUBMISSIONS

ABOUT