On Learning Language-Invariant Representations
for Universal Machine Translation

Han Zhao [1] **Junjie Hu** [1] **Andrej Risteski** [1]

Abstract

The goal of universal machine translation is to learn to translate between any pair of languages. Despite impressive empirical results and an increasing interest in massively multilingual models, theoretical analysis on translation errors made by such universal machine translation models is only nascent. In this paper, we formally prove certain impossibilities of this endeavour in general, as well as prove positive results in the presence of additional (but natural) structure of data. For the former, we derive a lower bound on the translation error in the many-to-many translation setting, which shows that any algorithm aiming to learn shared sentence representations among multiple language pairs has to make a large translation error on at least one of the translation tasks, if no assumption on the structure of the languages is made. For the latter, we show that if the paired documents in the corpus follow a natural *encoder-decoder* generative process, we can expect a natural notion of "generalization": a linear number of language pairs, rather than quadratic, suffices to learn a good representation. Our theory also explains what kinds of connection graphs between pairs of languages are better suited: ones with longer paths result in worse sample complexity. We believe our theoretical insights and implications contribute to the future algorithmic design of universal machine translation.

1. Introduction

Despite impressive improvements in neural machine translation (NMT), training a large multilingual NMT model with hundreds of millions of parameters usually requires a collection of *parallel corpora* at a large scale, on the order of millions or even billions of aligned sentences (Johnson et al., 2017; Arivazhagan et al., 2019) for supervised training. Although it is possible to automatically crawl the web (Nie et al., 1999; Resnik, 1999; Resnik & Smith, 2003) to collect parallel sentences for high-resource language pairs such as German-English and Chinese-English, it is often infeasible to manually translate large amounts of documents for low-resource language pairs, e.g., Nepali-English, Sinhala-English (Guzmán et al., 2019). Much recent progress in low-resource machine translation, has been driven by the idea of *universal machine translation* (UMT), also known as *multilingual machine translation* (Zoph & Knight, 2016; Johnson et al., 2017; Gu et al., 2018), which aims at training one single NMT to translate between multiple source and target languages. Typical UMT models leverage either a single shared encoder or language-specific encoders to map all source languages to a shared space, and translate the source sentences to a target language by a decoder. Inspired by the idea of UMT, there has been a recent trend towards learning language-invariant embeddings for multiple source languages in a shared latent space, which eases the cross-lingual generalization from high-resource languages to low-resource languages on many tasks, e.g., parallel corpus mining (Schwenk, 2018; Artetxe & Schwenk, 2019), sentence classification (Conneau et al., 2018b), cross-lingual information retrieval (Litschko et al., 2018), and dependency parsing (Kondratyuk & Straka, 2019), just to name a few.

The idea of finding an abstract "lingua franca" is very intuitive and the empirical results are impressive, yet theoretical understanding of various aspects of universal machine translation is limited. In this paper, we particularly focus on two basic questions:

1. *How can we measure the inherent tradeoff between the quality of translation and how language-invariant a representation is?*
2. *How many language pairs do we need aligned sentences for, to be able to translate between any pair of languages?*

Toward answering the first question, we show that in a completely assumption-free setup on the languages and distribution of the data, it is impossible to avoid making a large translation error on at least one pair of the translation tasks.

Informally we highlight our first theorem as follows, and provide the formal statements in Theorems 3.1 and 3.2.

Theorem 1.1 (Impossibility, Informal)**.** There exist a choice of distributions over documents from different languages, s.t. for any choice of maps from the language to a common representation, at least one of the translation pairs must incur a high cost. In addition, 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.

To answer the second question, we show that under fairly mild generative assumptions on the aligned documents, it is possible to not only do well on all of the pairwise translations, but also be able to do so after *only seeing* aligned documents of a *linear* number of languages, rather than a *quadratic* one. We summarize the second theorem as follows, and provide a formal statement in Theorem 4.1.

Theorem 1.2 (Sample complexity, Informal)**.** Under a generative model where the documents for each language are generated from a "ground-truth" encoder-decoder model, after seeing aligned documents for a *linear* number of pairs of languages, we can learn encoders/decoders that perform well on any unseen language pair.

Notation and Setup We first introduce the notation used throughout the paper and then briefly describe the problem setting of universal machine translation.

We use $\mathcal{L}$ to denote the set of all possible languages, e.g., $\{\text{English}, \text{French}, \text{German}, \text{Chinese}, \ldots\}$. For any language $L \in \mathcal{L}$, we associate with L an alphabet Σ_L that contains all the symbols from L. Note that we assume $|\Sigma_L| < \infty$, $\forall L \in \mathcal{L}$, but different languages could potentially share part of the alphabet. Given a language L, a sentence x in L is a sequence of symbols from Σ_L, and we denote Σ_L^* as the set of all sentences generated from Σ_L. Note that since in principle different languages could share the same alphabet, to avoid ambiguity, for each language L, there is a unique token $\langle L \rangle \in \Sigma_L$ and $\langle L \rangle \notin \Sigma_{L'}', \forall L' \neq L$. The goal of the unique token $\langle L \rangle$ is used to denote the source sentence, and a sentence x in L will have a unique prefix $\langle L \rangle$ to indicate that $x \in \Sigma_L^*$. Also, in this manuscript we will use sentence and string interchangeably.

Formally, let $\{L_i\}_{i \in [K]}$ be the set of K source languages and $L \notin \{L_i\}_{i \in [K]}$ be the target language we are interested in translating to. For a pair of languages L and L', we use $\mathcal{D}_{L,L'}$ to denote the joint distribution over the parallel sentence pairs from L and L'. Given this joint distribution, we also use $\mathcal{D}_{L,L'}(L)$ to mean the marginal distribution over sentences from L. Likewise we use $\mathcal{D}_{L,L'}(L')$ to denote the corresponding marginal distribution over sentences from L'. Finally, for two sets A and B, we use $A \sqcup B$ to denote the disjoint union of A and B.

2. Related Work

Multilingual Machine Translation Early studies on multilingual machine translation mostly focused on pivot methods (Och & Ney, 2001; Cohn & Lapata, 2007; De Gispert & Marino; Utiyama & Isahara, 2007) that use one pivot language to connect the translation between ultimate source and target languages, and train two separate statistical translation models (Koehn et al., 2003). Since the successful application of encoder-decoder architectures in sequential tasks (Sutskever et al., 2014), neural machine translation (Bahdanau et al., 2015; Wu et al., 2016) has made it feasible to jointly learn from parallel corpora in multiple language pairs, and perform translation to multiple languages by a single model. Existing studies have been proposed to explore different variants of encoder-decoder architectures by using separate encoders (decoders) for multiple source (target) languages (Dong et al., 2015; Firat et al., 2016a;b; Zoph & Knight, 2016; Platanios et al., 2018) or sharing the weight of a single encoder (decoder) for all source (target) languages (Ha et al., 2016; Johnson et al., 2017). Recent advances of universal neural machine translation have also been applied to improve low-resource machine translation (Neubig & Hu, 2018; Gu et al., 2018; Arivazhagan et al., 2019; Aharoni et al., 2019) and downstream NLP tasks (Artetxe & Schwenk, 2019; Schwenk & Douze, 2017). Despite the recent empirical success, theoretical understanding is only nascent. Our work takes a first step towards better understanding the limitation of existing approaches and proposes a sufficient generative assumption that guarantees the success of universal machine translation.

Invariant Representations The line of work on seeking a shared multilingual embedding space started from learning cross-lingual word embeddings from parallel corpora (Gouws et al., 2015; Luong et al., 2015; Faruqui & Dyer, 2014; Artetxe et al., 2017; Conneau et al., 2018a), and later extended to learning cross-lingual contextual representations (Devlin et al., 2019; Lample & Conneau, 2019; Huang et al., 2019; Conneau et al., 2019) from monolingual corpora. The idea of learning invariant representations is not unique in machine translation. In fact, similar ideas have already been used in other contexts, including domain adaptation (Ganin et al., 2016; Zhao et al., 2018; 2019b; Combes et al., 2020), fair representations (Zemel et al., 2013; Zhang et al., 2018; Zhao & Gordon, 2019; Zhao et al., 2019a) and counterfactual reasoning in causal inference (Johansson et al., 2016; Shalit et al., 2017). Different from these existing work, our work provides the first impossibility theorem on learning language-invariant representations in terms of recovering a perfect translator under the setting of sequence to sequence learning. Furthermore, we also give a generative model assumption under which we show that generalization over unseen language pairs is possible by learning language-invariant representations.

3. An Impossibility Theorem

In this section, for the clarity of presentation, we first focus the deterministic setting where for each language pair $L \neq L'$, there exists a ground-truth translator $f_{L \to L'}^* : \Sigma_L^* \to \Sigma_{L'}^*$ that takes an input sentence x from the source language L and outputs the ground-truth translation $f_{L \to L'}^*(x) \in \Sigma_{L'}^*$. Later we shall extend the setup to allow a probabilistic extension as well. Before we proceed, we first describe some concepts that will be used in the discussion.

Given a feature map $g : \mathcal{X} \to \mathcal{Z}$ that maps instances from the input space $\mathcal{X}$ to feature space $\mathcal{Z}$, we define $g_\sharp \mathcal{D} := \mathcal{D} \circ g^{-1}$ to be the induced (pushforward) distribution of $\mathcal{D}$ under g, i.e., for any event $E' \subseteq \mathcal{Z}$, $\Pr_{g \sharp \mathcal{D}}(E') := \Pr_{\mathcal{D}}(g^{-1}(E')) = \Pr_{\mathcal{D}}(\{x \in \mathcal{X} \mid g(x) \in E'\})$. For two distribution $\mathcal{D}$ and $\mathcal{D}'$ over the same sample space, we use the total variation distance to measure the discrepancy them: $d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}') := \sup_E |\Pr_{\mathcal{D}}(E) - \Pr_{\mathcal{D}'}(E)|$, where E is taken over all the measurable events under the common sample space. We use $\mathbb{I}(E)$ to denote the indicator function which takes value 1 iff the event E is true otherwise 0.

In general, given two sentences x and x', we use $\ell(x, x')$ to denote the loss function used to measure their distance. For example, we could use a $0-1$ loss function $\ell_{0-1}(x, x') = 0$ iff $x = x'$ else 1. If both x and x' are embedded in the same Euclidean space, we could also use the squared loss $\ell_2(x, x')$ as a more refined measure. To measure the performance of a translator f on a given language pair $L \to L'$ w.r.t. the ground-truth translator $f_{L \to L'}^*$, we define the error as

$$\mathrm{Err}_{\mathcal{D}}^{L \to L'}(f) := \mathbb{E}_{\mathcal{D}} \left[\ell_{0-1}(f(X), f_{L \to L'}^*(X)) \right],$$

which is the translation error of f as compared to the ground-truth translator $f_{L \to L'}^*$. For universal machine translation, the input string of the translator can be any sentence from any language. To this end, let Σ^* be the union of all the sentences/strings from all the languages of interest: $\Sigma^* := \bigcup_{L \in \mathcal{L}} \Sigma_L^*$. Then a universal machine translator of target language $L \in \mathcal{L}$ is a mapping $f_L : \Sigma^* \to \Sigma_L^*$. In words, f_L takes as input a string (from one of the possible languages) and outputs the corresponding translation in target language L. It is not hard to see that for such task there exists a perfect translator f_L^*:

$$f_L^*(x) = \sum_{L' \in \mathcal{L}} \mathbb{I}(x \in \Sigma_{L'}^*) \cdot f_{L' \to L}^*(x). \tag{1}$$

Note that $\{\Sigma_{L'}^* \mid L' \in \mathcal{L}\}$ forms a partition of Σ^*, so exactly one of the indicator $\mathbb{I}(x \in \Sigma_{L'}^*)$ in (1) will take value 1.

Given a target language L, existing approaches for universal machine seek to find an intermediate space $\mathcal{Z}$, such that source sentences from different languages are aligned within $\mathcal{Z}$. In particular, for each source language L', the goal is to find a feature mapping $g_{L'} : \Sigma_{L'}^* \to \mathcal{Z}$ so that the induced

distributions of different languages are close in $\mathcal{Z}$. The next step is to construct a decoder $h : \mathcal{Z} \to \Sigma_L^*$ that maps feature representation in $\mathcal{Z}$ to sentence in the target language L.

One interesting question about the idea of learning language-invariant representations is that, whether such method will succeed even under the benign setting where there is a ground-truth universal translator and the learner has access to infinite amount of data with unbounded computational resources. That is, we are interested in understanding the information-theoretic limit of such methods for universal machine translation.

In this section we first present an impossibility theorem in the restricted setting of translating from two source languages L_0 and L_1 to a target language L. Then we will use this lemma to prove a lower bound of the universal translation error in the general many-to-many setting. We will mainly discuss the implications and intuition of our theoretical results and use figures to help illustrate the high-level idea of the proof. We refer readers to the appendix for all the detailed proofs.

3.1. Two-to-One Translation

Recall that for each translation task $L_i \to L$, we have a joint distribution $\mathcal{D}_{L_i, L}$ over the aligned source-target sentences. For convenience of notation, we use $\mathcal{D}_i$ to denote the marginal distribution $\mathcal{D}_{L_i, L}(L_i)$ when the target language L is clear from the context. Given a fixed constant $\epsilon > 0$, we first define the ϵ-*universal language mapping*:

Definition 3.1 (ϵ-Universal Language Mapping). A map $g : \bigcup_{i \in [K]} \Sigma_{L_i}^* \to \mathcal{Z}$ is called an ϵ-universal language mapping if $d_{\mathrm{TV}}(g_\sharp \mathcal{D}_i, g_\sharp \mathcal{D}_j) \leq \epsilon$, $\forall i \neq j$.

In particular, if $\epsilon = 0$, we call the corresponding feature mapping a universal language mapping. In other words, a universal language mapping perfectly aligns the feature representations of different languages in feature space $\mathcal{Z}$. The following lemma provides a useful tool to connect the 0-1 translation error and the TV distance between the corresponding distributions.

Lemma 3.1. Let $\Sigma := \bigcup_{L \in \mathcal{L}} \Sigma_L$ and $\mathcal{D}_\Sigma$ be a language model over Σ^*. For any two string-to-string maps $f, f' : \Sigma^* \to \Sigma^*$, let $f_\sharp \mathcal{D}_\Sigma$ and $f'_\sharp \mathcal{D}_\Sigma$ be the corresponding pushforward distributions. Then $d_{\mathrm{TV}}(f_\sharp \mathcal{D}_\Sigma, f'_\sharp \mathcal{D}_\Sigma) \leq \Pr_{\mathcal{D}_\Sigma}(f(X) \neq f'(X))$ where $X \sim \mathcal{D}_\Sigma$.

The next lemma follows from the data-processing inequality for total variation and it shows that if languages are close in a feature space, then any decoder cannot increase the corresponding discrepancy in the output space.

Lemma 3.2. (Data-processing inequality) Let $\mathcal{D}$ and $\mathcal{D}'$ be any distributions over $\mathcal{Z}$, then for any decoder $h : \mathcal{Z} \to \Sigma_L^*$, $d_{\mathrm{TV}}(h_\sharp \mathcal{D}, h_\sharp \mathcal{D}') \leq d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}')$.
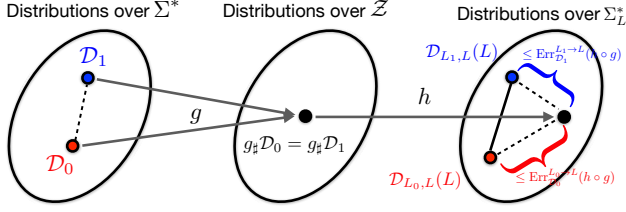
Figure 1. Proof by picture: Language-invariant representation g induces the same feature distribution over $\mathcal{Z}$, which leads to the same output distribution over the target language Σ_L^*. However, the parallel corpora of the two translation tasks in general have different marginal distributions over the target language, hence a triangle inequality over the output distributions gives the desired lower bound.

As a direct corollary, this implies that any distributions induced by a decoder over ϵ-universal language mapping must also be close in the output space:

Corollary 3.1. If $g : \Sigma^* \to \mathcal{Z}$ is an ϵ-universal language mapping, then for any decoder $h : \mathcal{Z} \to \Sigma_L^*$, $d_{\mathrm{TV}}((h \circ g)_\sharp \mathcal{D}_0, (h \circ g)_\sharp \mathcal{D}_1) \leq \epsilon$.

With the above tools, we can state the following theorem that characterizes the translation error in a two-to-one setting:

Theorem 3.1. (Lower bound, Two-to-One) Consider a setting of universal machine translation task with two source languages where $\Sigma^* = \Sigma_{L_0}^* \bigcup \Sigma_{L_1}^*$ and the target language is L. Let $g : \Sigma^* \to \mathcal{Z}$ be an ϵ-universal language mapping, then for any decoder $h : \mathcal{Z} \to \Sigma_L^*$, we have

$$\mathrm{Err}_{\mathcal{D}_0}^{L_0 \to L}(h \circ g) + \mathrm{Err}_{\mathcal{D}_1}^{L_1 \to L}(h \circ g)$$
$$\geq d_{\mathrm{TV}}(\mathcal{D}_{L_0,L}(L), \mathcal{D}_{L_1,L}(L)) - \epsilon. \quad (2)$$

Remark Recall that under our setting, there exists a perfect translator $f_L^* : \Sigma^* \to \Sigma_L^*$ in (1) that achieves zero translation error on both translation tasks. Nevertheless, the lower bound in Theorem 3.1 shows that one cannot hope to simultaneously minimize the joint translation error on both tasks through universal language mapping. Second, the lower bound is algorithm-independent and it holds even with unbounded computation and data. Third, the lower bound also holds even if all the data are perfect, in the sense that all the data are sampled from the perfect translator on each task. Hence, the above result could be interpreted as a kind of *uncertainty principle* in the context of universal machine translation, which says that any decoder based on language-invariant representations has to achieve a large translation error on at least one pair of translation task. We provide a proof-by-picture in Fig. 1 to illustrate the main idea underlying the proof of Theorem 3.1 in the special case where $\epsilon = 0$.

The lower bound is large whenever the distribution over target sentences differ between these two translation tasks.

This often happens in practical scenarios where the parallel corpus of high-resource language pair contains texts over a diverse domain whereas as a comparison, parallel corpus of low-resource language pair only contains target translations from a specific domain, e.g., sports, news, product reviews, etc. Such negative impact on translation quality due to domain mismatch between source and target sentences has also recently been observed and confirmed in practical universal machine translation systems, see Shen et al. (2019) and Pires et al. (2019) for more empirical corroborations.

3.2. Many-to-Many Translation

Theorem 3.1 presents a negative result in the setting where we have two source languages and one target language for translation. Nevertheless universal machine translation systems often involve multiple input and output languages simultaneously (Wu et al., 2016; Ranzato et al., 2019; Artetxe & Schwenk, 2019; Johnson et al., 2017). In this section we shall extend the previous lower bound in the simple two-to-one setting to the more general translation task of many-to-many setting.

To enable such extension, i.e., to be able to make use of multilingual data within a single system, we need to modify the input sentence to introduce the language token $\langle L \rangle$ at the beginning of the input sentence to indicate the target language L the model should translate to. This simple modification has already been used in practical MT systems (Johnson et al., 2017, Section 3). As an example, consider the following English sentence to be translated to French,

$\langle English \rangle$ *Hello, how are you?*

It will be modified to:

$\langle French \rangle \langle English \rangle$ *Hello, how are you?*

Note that the first token is used to indicate the target language to translate to while the second one is used to indicate the source language to avoid the ambiguity due to the potential overlapping alphabets between different languages.

Recall in Definition 3.1 we define a language map g to be ϵ-universal iff $d_{\mathrm{TV}}(g_\sharp \mathcal{D}_i, g_\sharp \mathcal{D}_j) \leq \epsilon, \forall i, j$. This definition is too stringent in the many-to-many translation setting since this will imply that the feature representations lose the information about which target language to translate to. In what follows we shall first provide a relaxed definition of ϵ-universal language mapping in the many-to-many setting and then show that even under this relaxed definition, learning universal machine translator via language-invariant representations is impossible in the worst case.

Definition 3.2 (ϵ-Universal Language Mapping, Many–to-Many). Let $\mathcal{D}_{L_i, L_k}$, $i, k \in [K]$ be the joint distribution of sentences (parallel corpus) in translating from L_i to L_k. A map $g : \bigcup_{i \in [K]} \Sigma_{L_i}^* \to \mathcal{Z}$ is called an

ϵ-universal language mapping if there exists a partition of $\mathcal{Z} = \sqcup_{k \in [K]} \mathcal{Z}_k$ such that $\forall k \in [K]$ and $\forall i \neq j$, $g_\sharp \mathcal{D}_{L_i, L_k}(L_i)$ and $g_\sharp \mathcal{D}_{L_j, L_k}(L_j)$ are supported on $\mathcal{Z}_k$ and $d_{\mathrm{TV}}(g_\sharp \mathcal{D}_{L_i, L_k}(L_i), g_\sharp \mathcal{D}_{L_j, L_k}(L_j)) \leq \epsilon$.

First of all, it is clear that when there is only one target language, then Definition 3.2 reduces to Definition 3.1. Next, the partition of the feature space $\mathcal{Z} = \sqcup_{k \in [K]} \mathcal{Z}_k$ essentially serves as a way to determine the target language L the model should translate to. Note that it is important here to enforce the partitioning condition of the feature space $\mathcal{Z}$, otherwise there will be ambiguity in determining the target language to translate to.

With the above extensions, now we are ready to present the following theorem which gives a lower bound for both the maximum error as well as the average error in the many-to-many universal translation setting.

Theorem 3.2. (Lower bound, Many-to-Many) Consider a universal machine translation task where $\Sigma^* = \bigcup_{i \in [K]} \Sigma_{L_i}^*$. Let $\mathcal{D}_{L_i, L_k}, i, k \in [K]$ be the joint distribution of sentences (parallel corpus) in translating from L_i to L_k. If $g : \Sigma^* \to \mathcal{Z}$ be an ϵ-universal language mapping, then for any decoder $h : \mathcal{Z} \to \Sigma^*$, we have

$$\max_{i,k \in [K]} \mathrm{Err}_{\mathcal{D}_{L_i, L_k}}^{L_i \to L_k}(h \circ g) \geq$$
$$\frac{1}{2} \max_{k \in [K]} \max_{i \neq j} d_{\mathrm{TV}}(\mathcal{D}_{L_i, L_k}(L_k), \mathcal{D}_{L_j, L_k}(L_k)) - \frac{\epsilon}{2},$$
$$\frac{1}{K^2} \sum_{i,k \in [K]} \mathrm{Err}_{\mathcal{D}_{L_i, L_k}}^{L_i \to L_k}(h \circ g) \geq$$
$$\frac{1}{K^2(K-1)} \sum_{k \in [K]} \sum_{i < j} d_{\mathrm{TV}}(\mathcal{D}_{L_i, L_k}(L_k), \mathcal{D}_{L_j, L_k}(L_k)) - \frac{\epsilon}{2}.$$

It is clear that both lower bounds in Theorem 3.2 include the many-to-one setting as a special case. The proof of Theorem 3.2 essentially applies the lower bound in Theorem 3.1 iteratively. Again, the underlying reason for such negative result to hold in the worst case is due to the mismatch of distributions of the target language in different pairs of translation tasks. It should also be noted that the results in Theorem 3.2 hold even if language-dependent encoders are used, as long as they induce invariant feature representations for the source languages.

How to Bypass this Limitation? There are various ways to get around the limitations pointed out by the theorems in this section.

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. 1, which is an essential ingredient in the proof of Theorem 3.1 and Theorem 3.2. 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 $\mathcal{D}_{L_i, L_j}$, i.e., by assuming some natural language generation process for the parallel corpora that are used for training (Cf. Section 4). 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 section 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.

4. Sample Complexity under a Generative Model

The results from the prior sections showed that absent additional assumptions on the distributions of the sentences in the corpus, there is a fundamental limitation on learning language-invariant representations for universal machine translation. Note that our negative result also holds in the setting where there exists a ground-truth universal machine translator – it's just that learning language-invariant representations cannot recover this ground-truth translator.

In this section we show that with additional natural structure on the distribution of the corpora we can resolve this issue. The structure is a natural underlying generative model from which sentences from different languages are generated, which "models" a common encoder-decoder structure that has been frequently used in practice (Cho et al., 2014; Sutskever et al., 2014; Ha et al., 2016). Under this setting, we show that it is not only possible to learn the optimal translator, but it is possible to do so only seeing documents from only a small subset of all the possible language pairs.

Moreover, we will formalize a notion of "*sample complexity*" in terms of number of pairs of languages for which parallel corpora are necessary, and how it depends on the structure of the connection graph between language pairs.

We first describe our generative model for languages and briefly talk about why such generative model could help to overcome the negative result in Theorem 3.2.

4.1. Language Generation Process and Setup

Language Generative Process The language generation process is illustrated in Fig. 2. Formally, we assume the existence of a shared "semantic space" $\mathcal{Z}$. Furthermore, for every language $L \in \mathcal{L}$, we have a "ground truth" pair
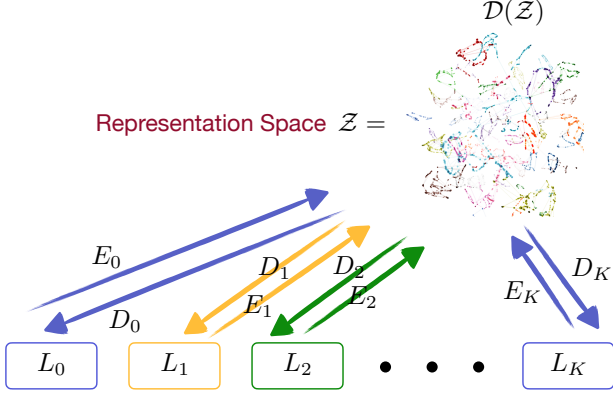
Figure 2. An encoder-decoder generative model of translation pairs. There is a global distribution $\mathcal{D}$ over representation space $\mathcal{Z}$, from which sentences of language L_i are generated via decoder D_i. Similarly, sentences could also be encoded via E_i to $\mathcal{Z}$.

of encoder and decoder $(\mathbf{E}_L, \mathbf{D}_L)$, where $\mathbf{E}_L : \mathbb{R}^d \to \mathbb{R}^d$, $\mathbf{E}_L \in \mathcal{F}$ is *bijective* and $\mathbf{D}_L = \mathbf{E}_L^{-1}$. We assume that $\mathcal{F}$ has a group structure under function composition: namely, for $\forall f_1, f_2 \in \mathcal{F}$, we have that $f_1^{-1}, f_2^{-1} \in \mathcal{F}$ and $f_1 \circ f_2^{-1}, f_2 \circ f_1^{-1} \in \mathcal{F}$ (e.g., a typical example of such group is the general linear group $\mathcal{F} = GL_d(\mathbb{R})$).

To generate a pair of aligned sentences for two languages L, L', we first sample a $z \sim \mathcal{D}$, and subsequently generate

$$x = \mathbf{D}_L(z), \quad x' = \mathbf{D}_{L'}(z), \tag{3}$$

where the x is a vector encoding of the appropriate sentence in L (e.g., a typical encoding is a frequency count of the words in the sentence, or a sentence embedding using various neural network models (Zhao et al., 2015; Kiros et al., 2015; Wang et al., 2017)). Similarly, x' is the corresponding sentence in L'. Reciprocally, given a sentence x from language L, the encoder $\mathbf{E}_L$ maps the sentence x into its corresponding latent vector in $\mathcal{Z}$: $z = \mathbf{E}_L(x)$.

We note that we assume this deterministic map between z and x for simplicity of exposition—in Section 4.4 we will extend the results to the setting where x has a conditional distribution given z of a parametric form.

We will assume the existence of a graph H capturing the pairs of languages for which we have aligned corpora – we can think of these as the "high-resource" pairs of languages. For each edge in this graph, we will have a corpus $S = \{(x_i, x_i')\}_{i=1}^n$ of aligned sentences.[1] The goal will be to learn encoder/decoders that perform well on the potentially unseen pairs of languages. To this end, we will be providing a sample complexity analysis for the number of

---

[1]In general each edge can have different number of aligned sentences. We use the same number of aligned sentences n just for the ease of presentation.

paired sentences for each pair of languages with an edge in the graph, so we will need a measure of the complexity of $\mathcal{F}$. We will use the covering number, though our proofs are flexible, and similar results would hold for Rademacher complexity, VC dimension, or any of the usual complexity measures.

Definition 4.1 (Covering number). For any $\epsilon > 0$, the *covering number* $\mathcal{N}(\mathcal{F}, \epsilon)$ of the function class $\mathcal{F}$ under the ℓ_∞ norm is the minimum number $k \in \mathbb{N}$ such that $\mathcal{F}$ could be covered with k (ℓ_∞) balls of radius ϵ, i.e., there exists $\{f_1, \ldots, f_k\} \subseteq \mathcal{F}$ such that, for all $f \in \mathcal{F}$, there exists $i \in [k]$ with $\|f - f_i\|_\infty = \max_{x \in \mathbb{R}^d} \|f(x) - f_i(x)\|_2 \le \epsilon$.

Finally, we will assume that the functions in $\mathcal{F}$ are bounded and Lipschitz:

Assumption 4.1 (Smoothness and Boundedness). $\mathcal{F}$ is bounded under the $\|\cdot\|_\infty$ norm, i.e., there exists $M > 0$, such that $\forall f \in \mathcal{F}$, $\|f\|_\infty \le M$. Furthermore, there exists $0 \le \rho < \infty$, such that for $\forall x, x' \in \mathbb{R}^d$, $\forall f \in \mathcal{F}$, $\|f(x) - f(x')\|_2 \le \rho \cdot \|x - x'\|_2$.

Training Procedure Turning to the training procedure, we will be learning encoders $E_L \in \mathcal{F}$ for each language L. The decoder for that language will be E_L^{-1}, which is well defined since $\mathcal{F}$ has a group structure. Since we are working with a vector space, rather than using the (crude) 0-1 distance, we will work with a more refined loss metric for a translation task $L \to L'$:

$$\varepsilon(E_L, E_{L'}) := \|E_{L'}^{-1} \circ E_L - \mathbf{E}_{L'}^{-1} \circ \mathbf{E}_L\|_{\ell_2(\mathbf{D}_{L\sharp}\mathcal{D})}^2. \tag{4}$$

Note that the ℓ_2 loss is taken over the distribution of the input samples $\mathbf{D}_{L\sharp}\mathcal{D} = \mathbf{E}_{L\ \sharp}^{-1}\mathcal{D}$, which is the natural one under our generative process. Again, the above error measures the discrepancy between the predicted translation w.r.t. the one give by the ground-truth translator, i.e., the composition of encoder $\mathbf{E}_L$ and decoder $\mathbf{D}_{L'}$. Straightforwardly, the empirical error over a corpus $S = \{(x_i, x_i')\}_{i=1}^n$ of aligned sentences for a pair of languages (L, L') is defined by

$$\widehat{\varepsilon}_S(E_L, E_{L'}) := \frac{1}{n} \sum_{i \in [n]} \|E_{L'}^{-1} \circ E_L(x_i) - x_i'\|_2^2, \tag{5}$$

where S is generated by the generation process. Following the paradigm of empirical risk minimization, the loss to train the encoders will be the obvious one:

$$\min_{\{E_L, L \in \mathcal{L}\}} \sum_{(L,L') \in H} \widehat{\varepsilon}_S(E_L, E_{L'}). \tag{6}$$

Remarks Before we proceed, one natural question to ask here is that, how does this generative model assumption circumvent the lower bound in Theorem 3.2? To answer this question, note the following easy proposition:

Proposition 4.1. Under the encoder-decoder generative assumption, $\forall i, j \in [K]$, $d_{\text{TV}}(\mathcal{D}_{L_i,L}(L), \mathcal{D}_{L_j,L}(L)) = 0$.

Proposition 4.1 holds because the marginal distribution of the target language L under any pair of translation task equals the pushforward of $\mathcal{D}(Z)$ under $\mathbf{D}_L$: $\forall i \in [K], \mathcal{D}_{L_i,L}(L) = \mathbf{D}_{L\sharp}\mathcal{D}(\mathcal{Z})$. Hence the lower bounds gracefully reduce to 0 under our encoder-decoder generative process, meaning that there is no loss of translation accuracy using universal language mapping.

4.2. Main Result: Translation between Arbitrary Pairs of Languages

The main theorem we prove is that if the graph H capturing the pairs of languages for which we have aligned corpora is connected, given sufficiently many sentences for each pair, we will learn encoder/decoders that perform well on the unseen pairs. Moreover, we can characterize how good the translation will be based on the distance of the languages in the graph. Concretely:

Theorem 4.1 (Sample complexity under generative model). Suppose H is connected. Furthermore, suppose the trained $\{E_L\}_{L \in \mathcal{L}}$ satisfy

$$\forall L, L' \in H : \widehat{\varepsilon}_S(E_L, E_{L'}) \le \epsilon_{L,L'},$$

for $\epsilon_{L,L'} > 0$. Furthermore, for $0 < \delta < 1$ suppose the number of sentences for each aligned corpora for each training pair (L, L') is $\Omega\left(\frac{1}{\epsilon_{L,L'}^2} \cdot \left(\log \mathcal{N}(\mathcal{F}, \frac{\epsilon_{L,L'}}{16M}) + \log(K/\delta)\right)\right)$. Then, with probability $1 - \delta$, for any pair of languages $(L, L') \in \mathcal{L} \times \mathcal{L}$ and $L = L_1, L_2, \ldots, L_m = L'$ a path between L and L' in H, we have $\varepsilon(E_L, E_{L'}) \le 2\rho^2 \sum_{k=1}^{m-1} \epsilon_{L_k, L_{k+1}}$.

Remark We make several remarks about the theorem statement. Note that the guarantee is in terms of translation error rather than parameter recovery. In fact, due to the identifiability issue, we cannot hope to recover the ground truth encoders $\{\mathbf{E}_L\}_{L \in \mathcal{L}}$: it is easy to see that composing all the encoders with an invertible mapping $f \in \mathcal{F}$ and composing all the decoders with $f^{-1} \in \mathcal{F}$ produces exactly the same outputs.

Furthermore, the upper bound is adaptive, in the sense that for any language pair (L_i, L_j), the error depends on the sum of the errors connecting (L_i, L_j) in the translation graph H. One can think naturally as the low-error edges as resource-rich pairs: if the function class $\mathcal{F}$ is parametrized by finite-dimensional parameter space with dimension p, then using standard result on the covering number of finite-dimensional vector space (Anthony & Bartlett, 2009), we know that $\log \mathcal{N}(\mathcal{F}, \frac{\epsilon}{16M}) = \Theta(p \log(1/\epsilon))$; as a consequence, the number of documents needed for a pair scales as $\log(1/\epsilon_{L,L'})/\epsilon_{L,L'}^2$.
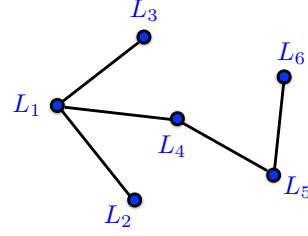

Figure 3. A translation graph H over $K = 6$ languages. The existence of an edge between a pair of nodes L_i and L_j means that the learner has been trained on the corresponding language pair. In this example the diameter of the graph $\text{diam}(H) = 4$: L_3, L_1, L_4, L_5, L_6.

Furthermore, as an immediate corollary of the theorem, if we assume $\epsilon_{L,L'} \le \epsilon$ for all $(L, L') \in H$, we have $\varepsilon(E_L, E_{L'}) \le 2\rho^2 d_{L,L'} \cdot \epsilon$, where $d_{L,L'}$ is the length of the shortest path connecting L and L' in H. It also immediately follows that for any pair of languages L, L', we have $\varepsilon(E_L, E_{L'}) \le 2\rho^2 \text{diam}(H) \cdot \epsilon$ where $\text{diam}(H)$ is the diameter of H – thus the intuitive conclusion that graphs that do not have long paths are preferable.

The upper bound in Theorem 4.1 also provides a counterpoint to the lower-bound, showing that under a generative model for the data, it is possible to learn a pair of encoder/decoder for each language pair after seeing aligned corpora only for a linear number of pairs of languages (and not quadratic!), corresponding to those captured by the edges of the translation graph H. As a final note, we would like to point out that an analogous bound can be proved easily for other losses like the 0-1 loss or the general ℓ_p loss as well.

4.3. Proof Sketch of the Theorem

Before we provide the proof for the theorem, we first state several useful lemmas that will be used during our analysis.

Concentration Bounds The first step is to prove a concentration bound for the translation loss metric on each pair of languages. In this case, it will be easier to write the losses in terms of one single function: namely notice that in fact $\varepsilon(E_L, E'_L)$ only depends on $E_{L'}^{-1} \circ E_L$, and due to the group structure, $\mathcal{F} \ni f := E_{L'}^{-1} \circ E_L$. To that end, we will abuse the notation somewhat and denote $\varepsilon(f) := \varepsilon(E_L, E'_L)$. The following lemma is adapted from Bartlett (1998) where the bound is given in terms of binary classification error while here we present a bound using ℓ_2 loss. At a high level, the bound uses covering number to concentrate the empirical loss metric to its corresponding population counterpart.

Lemma 4.1. If $S = \{(x_i, x'_i)\}_{i=1}^n$ is sampled i.i.d. according to the encoder-decoder generative process, the following

bound holds:

$$\Pr_{S \sim \mathcal{D}^n} \left(\sup_{f \in \mathcal{F}} |\varepsilon(f) - \widehat{\varepsilon}_S(f)| \geq \epsilon \right)$$
$$\leq 2\mathcal{N}(\mathcal{F}, \frac{\epsilon}{16M}) \cdot \exp \left(\frac{-n\epsilon^2}{16M^4} \right).$$

This lemma can be proved using a ϵ-net argument with covering number. With this lemma, we can bound the error given by an empirical risk minimization algorithm:

Theorem 4.2. (Generalization, single task) Let S be a sample of size n according to our generative process. Then for any $0 < \delta < 1$, for any $f \in \mathcal{F}$, w.p. at least $1 - \delta$, the following bound holds:

$$\varepsilon(f) \leq \widehat{\varepsilon}_S(f) + O \left(\sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \frac{\epsilon}{16M}) + \log(1/\delta)}{n}} \right). \tag{7}$$

Theorem 4.2 is a finite sample bound for generalization on a single pair of languages. This bound gives us an error measure on an edge in the translation graph in Fig. 3. Now, with an upper bound on the translation error of each *seen* language pair, we are ready to prove the main theorem (Theorem 4.1) which bounds the translation error for all possible pairs of translation tasks:

Proof of Theorem 4.1. First, under the assumption of Theorem 4.1, for any pair of language (L, L'), we know that the corpus contains at least $\Omega \left(\frac{1}{\epsilon_{L,L'}^2} \cdot \left(\log \mathcal{N}(\mathcal{F}, \frac{\epsilon_{L,L'}}{16M}) + \log(K/\delta) \right) \right)$ parallel sentences. Then by Theorem 4.2, with probability $1 - \delta$, for any L, L' connected by an edge in H, we have

$$\varepsilon(E_L, E_{L'}) \leq \widehat{\varepsilon}(E_L, E_{L'}) + \epsilon_{L,L'} \leq \epsilon_{L,L'} + \epsilon_{L,L'} = 2\epsilon_{L,L'},$$

where the last inequality is due to the assumption that $\widehat{\varepsilon}(E_L, E_{L'}) \leq \epsilon_{L,L'}$. Now consider any $L, L' \in \mathcal{L} \times \mathcal{L}$, connected by a path

$$L' = L_1, L_2, L_3, \ldots, L_m = L$$

of length at most m. We will bound the error

$$\varepsilon(E_L, E_{L'}) = \|E_{L'}^{-1} \circ E_L - \mathbf{E}_{L'}^{-1} \circ \mathbf{E}_L\|_{\ell_2(\mathbf{D}_{L\sharp}\mathcal{D})}^2$$

by a judicious use of the triangle inequality. Namely, let's denote

$$I_1 := \mathbf{E}_{L_1}^{-1} \circ \mathbf{E}_{L_m},$$
$$I_k := E_{L_1}^{-1} \circ E_{L_k} \circ \mathbf{E}_{L_k}^{-1} \circ \mathbf{E}_{L_m}, \qquad 2 \leq k \leq m-1,$$
$$I_m := E_{L_1}^{-1} \circ E_{L_m}.$$

Then, we can write

$$\|E_{L'}^{-1} \circ E_L - \mathbf{E}_{L'}^{-1} \circ \mathbf{E}_L\|_{\ell_2(\mathbf{D}_{L\sharp}\mathcal{D})}$$
$$= \| \sum_{k=1}^{m-1} I_k - I_{k+1} \|_{\ell_2(\mathbf{D}_{L\sharp}\mathcal{D})}$$
$$\leq \sum_{k=1}^{m-1} \|I_k - I_{k+1}\|_{\ell_2(\mathbf{D}_{L\sharp}\mathcal{D})}. \tag{8}$$

Furthermore, notice that we can rewrite $I_k - I_{k+1}$ as

$$E_{L_1}^{-1} \circ E_{L_k} \left(\mathbf{E}_{L_k}^{-1} \circ \mathbf{E}_{L_{k+1}} - E_{L_k}^{-1} \circ E_{L_{k+1}} \right) \mathbf{E}_{L_{k+1}}^{-1} \circ \mathbf{E}_{L_m}.$$

Given that $E_{L_1}^{-1}$ and E_{L_k} are ρ-Lipschitz we have

$$\|I_k - I_{k+1}\|_{\ell_2(\mathbf{D}_{L\sharp}\mathcal{D})}$$
$$= \|E_{L_1}^{-1} \circ E_{L_k} \left(E_{L_k}^{-1} \circ E_{L_{k+1}} - \mathbf{E}_{L_k}^{-1} \circ \mathbf{E}_{L_{k+1}} \right)\|_{\ell_2(\mathbf{D}_{L_{k+1}\sharp}\mathcal{D})}$$
$$\leq \rho^2 \|\left(E_{L_k}^{-1} \circ E_{L_{k+1}} - \mathbf{E}_{L_k}^{-1} \circ \mathbf{E}_{L_{k+1}} \right)\|_{\ell_2(\mathbf{D}_{L_{k+1}\sharp}\mathcal{D})}$$
$$\leq 2\rho^2 \epsilon_{L_k, L_{k+1}},$$

where the first line is from the definition of pushforward distribution, the second line is due to the Lipschitzness of $\mathcal{F}$ and the last line follows since all (L_k, L_{k+1}) are edges in H. Plugging this into (8), we have

$$\|E_{L'}^{-1} \circ E_L - \mathbf{E}_{L'}^{-1} \circ \mathbf{E}_L\|_{\ell_2(\mathbf{D}_{L\sharp}\mathcal{D})} \leq 2\rho^2 \sum_{k=1}^{m} \epsilon_{L_k, L_{k+1}}.$$

To complete the proof, realize that we need the events $|\varepsilon(L_k, L_{k+1}) - \widehat{\varepsilon}(L_k, L_{k+1})| \leq \epsilon_{L_k, L_{k+1}}$ to hold simultaneously for all the edges in the graph H. Hence it suffices if we can use a union bound to bound the failing probability. To this end, for each edge, we amplify the success probability by choosing the failure probability to be δ/K^2, and we can then bound the overall failure probability as:

Pr (At least one edge in the graph H fails to satisfy (7))
$$\leq \sum_{(i,j) \in H} \Pr \left(|\varepsilon(L_i, L_j) - \widehat{\varepsilon}(L_i, L_j)| > \epsilon_{L_i, L_j} \right)$$
$$\leq \sum_{(i,j) \in H} \delta/K^2$$
$$\leq \frac{K(K-1)}{2} \cdot \frac{\delta}{K^2}$$
$$\leq \delta.$$

The first inequality above is due to the union bound, and the second one is from Theorem 4.2 by choosing the failing probability to be δ/K^2. ∎

4.4. Extension to Randomized Encoders and Decoders

Our discussions so far on the sample complexity under the encoder-decoder generative process assume that the ground-truth encoders and decoders are *deterministic* and *bijective*.

This might seem to be a quite restrictive assumption, but nevertheless our underlying proof strategy using transitions on the translation graph still works in more general settings. In this section we shall provide an extension of the previous deterministic encoder-decoder generative process to allow randomness in the generation process. Note that this extension simultaneously relaxes both the deterministic and bijective assumptions before.

As a first step of the extension, since there is not a notion of inverse function anymore in the randomized setting, we first define the ground-truth encoder-decoder pair $(\mathbf{E}_L, \mathbf{D}_L)$ for a language $L \in \mathcal{L}$.

Definition 4.2. Let $\mathcal{D}_r$ and $\mathcal{D}_{r'}$ be two distributions over random seeds r and r' respectively. A randomized decoder $\mathbf{D}_{L_i}$ is a deterministic function that maps a feature z along with a random seed r to a sentences in language L_i. Similarly, a randomized encoder $\mathbf{E}_{L_i}$ maps a sentence $x \in \Sigma_{L_i}^*$ and a random seed r' to a representation in $\mathcal{Z}$. $(\mathbf{E}_{L_i}, \mathbf{D}_{L_i})$ is called an encoder-decoder pair if it keeps the distribution $\mathcal{D}$ over $\mathcal{Z}$ invariant under the randomness of $\mathcal{D}_r$ and $\mathcal{D}_{r'}$:

$$\mathbf{E}_{L_i \sharp}(\mathbf{D}_{L_i \sharp}(\mathcal{D} \times \mathcal{D}_r) \times \mathcal{D}_{r'}) = \mathcal{D}, \qquad (9)$$

where we use $\mathcal{D} \times \mathcal{D}'$ to denote the product measure of distributions $\mathcal{D}$ and $\mathcal{D}'$.

Just like the deterministic setting, here we still assume that $\mathbf{E}_{L_i}, \mathbf{D}_{L_i} \in \mathcal{F}$ where $\mathcal{F}$ is closed under function composition. Furthermore, in order to satisfy Definition 4.2, we assume that $\forall \ \mathbf{D}_{L_i} \in \mathcal{F}$, there exists a corresponding $\mathbf{E}_{L_i} \in \mathcal{F}$, such that $(\mathbf{E}_{L_i}, \mathbf{D}_{L_i})$ is an encoder-decoder pair that verifies Definition 4.2. It is clear that the deterministic encoder-decoder pair in Section 4.1 is a special case of that in Definition 4.2: in that case $\mathbf{D}_{L_i} = \mathbf{E}_{L_i}^{-1}$ so that $\mathbf{E}_{L_i} \circ \mathbf{D}_{L_i} = \mathrm{id}_{\mathcal{Z}}$, the identity map over feature space $\mathcal{Z}$. Furthermore there is no randomness from r and r', hence the invariant criterion becomes $\mathbf{E}_{L_i \sharp} \mathbf{D}_{L_i \sharp} \mathcal{D} = (\mathbf{E}_{L_i} \circ \mathbf{D}_{L_i})_\sharp \mathcal{D} = \mathrm{id}_{\mathcal{Z} \sharp} \mathcal{D} = \mathcal{D}$, which trivially holds.

The randomness mechanism in Definition 4.2 has several practical implementations in practice. For example, the denoising autoencoder (Vincent et al., 2008), the encoder part of the conditional generative adversarial network (Mirza & Osindero, 2014), etc. Again, in the randomized setting we still need to have an assumption on the structure of $\mathcal{F}$, but this time a relaxed one:

Assumption 4.2 (Smoothness and Boundedness). $\mathcal{F}$ is bounded under the $\| \cdot \|_\infty$ norm, i.e., there exists $M > 0$, such that $\forall f \in \mathcal{F}, \|f\|_\infty \leq M$. Furthermore, there exists $0 \leq \rho < \infty$, such that for $\forall x, x' \in \mathbb{R}^d, \forall f \in \mathcal{F}$, $\|\mathbb{E}_{\mathcal{D}_r}[f(x, r) - f(x', r)]\|_2 \leq \rho \cdot \|x - x'\|_2$.

Correspondingly, we also need to slightly extend our loss metric under the randomized setting to the following:

$$\varepsilon(E_L, D_{L'}) := \mathbb{E}_{r,r'} \|D_{L'} \circ E_L - \mathbf{D}_{L'} \circ \mathbf{E}_L\|_{\ell_2(\mathbf{D}_{L \sharp}(\mathcal{D} \times \mathcal{D}_r))}^2,$$

where the expectation is taken over the distributions over random seeds r and r'. The empirical error could be extended in a similar way by replacing the population expectation with the empirical expectation. With the above extended definitions, now we are ready to state the following generalization theorem under randomized setting:

Theorem 4.3. (Sample complexity under generative model, randomized setting) Suppose H is connected and the trained $\{E_L\}_{L \in \mathcal{L}}$ satisfy

$$\forall L, L' \in H : \widehat{\varepsilon}_S(E_L, D_{L'}) \leq \epsilon_{L,L'},$$

for $\epsilon_{L,L'} > 0$. Furthermore, for $0 < \delta < 1$ suppose the number of sentences for each aligned corpora for each training pair (L, L') is $\Omega\left(\frac{1}{\epsilon_{L,L'}^2} \cdot \left(\log \mathcal{N}(\mathcal{F}, \frac{\epsilon_{L,L'}}{16M}) + \log(K/\delta))\right)\right)$. Then, with probability $1 - \delta$, for any pair of languages $(L, L') \in \mathcal{L} \times \mathcal{L}$ and $L = L_1, L_2, \ldots, L_m = L'$ a path between L and L' in H, we have $\varepsilon(E_L, D_{L'}) \leq 2\rho^2 \sum_{k=1}^{m-1} \epsilon_{L_k, L_{k+1}}$.

We comment that Theorem 4.3 is completely parallel to Theorem 4.1, except that we use generalized definitions under the randomized setting instead. Hence all the discussions before on Theorem 4.1 also apply here.

5. Discussion and Conclusion

In this paper we provided the first theoretical study on using language-invariant representations for universal machine translation. Our results are two-fold. First, we showed that without appropriate assumption on the generative structure of languages, there is an inherent tradeoff between learning language-invariant representations versus achieving good translation performance jointly in general. In particular, our results show that if the distributions (language models) of the target language differ between different translation pairs, then any machine translation method based on learning language-invariant representations is bound to achieve a large error on at least one of the translation tasks, even with unbounded computational resources.

On the positive side, we also show that, under appropriate generative model assumption of languages, e.g., a typical encoder-decoder model, it is not only possible to recover the ground-truth translator between any pair of languages that appear in the parallel corpora, but also we can hope to achieve a small translation error on sentences from unseen pair of languages, as long as they are connected in the so-called translation graph. This result holds in both deterministic and randomized settings. In addition, our result also characterizes how the relationship (distance) between these two languages in the graph affects the quality of translation in an intuitive manner: a graph with long connections results in a poorer translation.

Acknowledgements

References

Aharoni, R., Johnson, M., and Firat, O. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388.

Anthony, M. and Bartlett, P. L. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

Artetxe, M. and Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

Artetxe, M., Labaka, G., and Agirre, E. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL 2017*, pp. 451–462, 2017.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

Bartlett, P. Thesamplecomplexityofp atternclassification withneuralnetworks: Thesizeoftheweightsismo reimportantthan thesizeofthenetwork. *IEEETrans. Inf. Theory*, 44 (2), 1998.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Cohn, T. and Lapata, M. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 728–735, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Combes, R. T. d., Zhao, H., Wang, Y.-X., and Gordon, G. Domain adaptation with conditional distribution matching and generalized label shift. *arXiv preprint arXiv:2003.04475*, 2020.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, 2018a.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

De Gispert, A. and Marino, J. B. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. Citeseer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1723–1732, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1166.

Faruqui, M. and Dyer, C. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter*

of the Association for Computational Linguistics, pp. 462–471, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1049.

Firat, O., Cho, K., and Bengio, Y. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 866–875, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1101.

Firat, O., Sankaran, B., Al-onaizan, Y., Yarman Vural, F. T., and Cho, K. Zero-resource translation with multilingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 268–277, Austin, Texas, November 2016b. Association for Computational Linguistics. doi: 10.18653/v1/D16-1026.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Gouws, S., Bengio, Y., and Corrado, G. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of ICML 2015*, pp. 748–756, 2015.

Gu, J., Hassan, H., Devlin, J., and Li, V. O. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 344–354, 2018.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6098–6111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1632.

Ha, T.-L., Niehues, J., and Waibel, A. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*, 2016.

Huang, H., Liang, Y., Duan, N., Gong, M., Shou, L., Jiang, D., and Zhou, M. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2485–2494, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1252.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029, 2016.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. Skip-thought vectors. In *Advances in neural information processing systems*, pp. 3294–3302, 2015.

Koehn, P., Och, F. J., and Marcu, D. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133, 2003.

Kondratyuk, D. and Straka, M. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2779–2795, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1279.

Lample, G. and Conneau, A. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Litschko, R., Glavaš, G., Ponzetto, S. P., and Vulić, I. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1253–1256, 2018.

Luong, T., Pham, H., and Manning, C. D. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 151–159, 2015.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Neubig, G. and Hu, J. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 875–880, Brussels, Belgium,

October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1103.

Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 74–81, 1999.

Och, F. J. and Ney, H. Statistical multi-source translation. pp. 253–258, 2001.

Pires, T., Schlinger, E., and Garrette, D. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*, 2019.

Platanios, E. A., Sachan, M., Neubig, G., and Mitchell, T. Contextual parameter generation for universal neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, November 2018.

Ranzato, M., Guzmán, P., Chaudhary, V., Chen, P.-J., Shen, J., and Li, X. Recent advances in low-resource machine translation. 2019.

Resnik, P. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 527–534, College Park, Maryland, USA, June 1999. Association for Computational Linguistics. doi: 10.3115/1034678.1034757.

Resnik, P. and Smith, N. A. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, 2003. doi: 10.1162/089120103322711578.

Schwenk, H. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 228–234, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2037.

Schwenk, H. and Douze, M. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 157–167, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2619.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3076–3085. JMLR. org, 2017.

Shen, J., Chen, P.-J., Le, M., He, J., Gu, J., Ott, M., Auli, M., and Ranzato, M. The source-target domain mismatch problem in machine translation. *arXiv preprint arXiv:1909.13151*, 2019.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. Curran Associates, Inc., 2014.

Utiyama, M. and Isahara, H. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 484–491, Rochester, New York, April 2007. Association for Computational Linguistics.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Wang, R., Finch, A., Utiyama, M., and Sumita, E. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 560–566, 2017.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Zhao, H. and Gordon, G. Inherent tradeoffs in learning fair representations. In *Advances in neural information processing systems*, pp. 15649–15659, 2019.

Zhao, H., Lu, Z., and Poupart, P. Self-adaptive hierarchical sentence model. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

Zhao, H., Zhang, S., Wu, G., Moura, J. M., Costeira, J. P., and Gordon, G. J. Adversarial multiple source domain adaptation. In *Advances in neural information processing systems*, pp. 8559–8570, 2018.

Zhao, H., Coston, A., Adel, T., and Gordon, G. J. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162*, 2019a.

Zhao, H., Des Combes, R. T., Zhang, K., and Gordon, G. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532, 2019b.

Zoph, B. and Knight, K. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 30–34, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1004.