CS447 Literature Review: Machine Translation with Insufficient Bilingual Data

Yongda Fan,
yongdaf2@illinois.edu

August 23, 2024

Abstract

Although end-to-end neural machine translation (NMT) has huge success in language pairs with sufficient parallel corpus, it is still a big challenge to deal with language pairs with very little bilingual data available. Recently, there has been great progress in solving this issue, using techniques such as unsupervised learning, reinforcement learning or transfer learning. This paper summarizes some of the exciting work published recently that partially solves the issue of insufficient bilingual data, as well as discusses where they can be further improved.

1 Introduction

Neural Machine Translation (NMT) can achieve near-human-level performance for resource-rich language pairs (Wu et al., 2019; Hassan et al., 2018), but such a large amount of bilingual corpus may not be available for all language pairs. Additionally, training a machine translation (MT) model with limited resources can be very challenging (Koehn and Knowles, 2017). Recently, there has been some progress on how to address this data scarcity issue.

Artetxe et al. (2017) proposes a method that learns bilingual word embedding in an unsupervised or semi-supervised fashion. It can be used either to induce a bilingual dictionary or as a building block for an unsupervised NMT algorithm. Lample et al. (2018b) shows an unsupervised MT framework, which performs much better than all previous ones. Chen et al. (2017) demonstrates a reinforcement learning scheme that can be used when bilingual corpus to some intermediate pivot language is available. Zoph et al. (2016) presents a method to improve the translation accuracy on a low-resource language pair using transfer learning.

These four methods can achieve significantly better results compared with the vanilla NMT method when the training data is insufficient. Yet they each have their own constraints and use cases. In this literature review, we shall discuss these methods in the following sections.

2 Background

The idea of machine translation was first proposed by Andrew (1953). A few years later, the first machine translation demonstration was completed on a computer called APEXC at the Computational Laboratory, Birkbeck College, London (Cleave, 1957). However, its translations were rather rudimentary. In 1993, IBM published a statistical machine translation model (Brown et al., 1993), which was a big step for machine translation because it could translate sentences with complicated structures. Nowadays, the end-to-end NMT model has become more popular since it has demonstrated the ability to achieve near-human performance (Wu et al., 2019).

3 Bilingual word embedding with almost no bilingual data

Bilingual word embedding can be very useful for many cross-lingual tasks, and one such example is the machine translation task (Zou et al., 2013). However, historically, creating a bilingual word embedding usually requires a large amount of bilingual data, such as parallel corpora (Gouws et al., 2015; Luong et al., 2015a), document-aligned comparable corpora (Søgaard et al., 2015; Vulić and Moens, 2016; Mogadala and Rettinger, 2016), or bilingual dictionaries (Mikolov et al., 2013; Artetxe et al., 2016).

Artetxe et al. (2017) demonstrates a method which learns a bilingual word embedding with a seed dictionary of size as little as 25 word pairs, using a self-learning technique (see Figure 1). Such a seed dictionary should be much easier to obtain compared to other dictionary-based methods, which usually require a few thousand word pairs to work well (Artetxe et al., 2016; Xing et al., 2015; Zhang et al., 2016; Mikolov et al., 2013). Additionally, Artetxe et al. (2017)'s method can further reduce the need for bilingual data, using trivially generated seed dictionaries of numerals (i.e. 1-1, 2-2, 3-3, etc.).

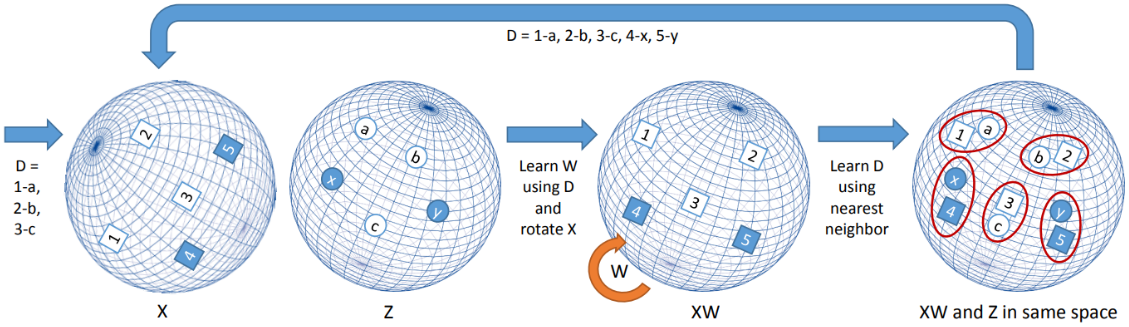


Figure 1: A general schema of the self-learning framework (Artetxe et al., 2017)

3.1 Self-learning framework

One of the common practice applications for bilingual embedding mappings is to translate the source words that are missing the training dictionary (Zhao et al., 2015). It can also be considered as using a seed dictionary to learn a mapping, which then induces a better dictionary. Based on this observation, Artetxe et al. (2017) proposes an algorithm (see Algorithm 1) that iteratively improves the quality of the mapping and the dictionary. Such a framework can be adopted by any embedding mapping and dictionary induction techniques. However, their efficiency turns out to be important due to the fact that a dictionary is explicitly rebuilt at every iteration (Artetxe et al., 2017). In the original paper, the author adopts the Artetxe et al. (2016) method for learning the embedding and a nearest neighbour retrieval method with dot product similarity for the dictionary induction.

Algorithm 1 Self-learning framework (Artetxe et al., 2017)

```
Input: $X$ (source embedding)
Input: $Z$ (target embedding)
Input: $D$ (seed dictionary)
  repeat
      $W \leftarrow \textsc{LearnMapping}(X, Z, D)$
      $D \leftarrow \textsc{LearnDictionary}(X, Z, W)$
  until convergence
  $\textsc{EvaluateDictionary}(D)$
```

The Artetxe et al. (2016) method is chosen for learning the embedding due to its simplicity and good results (Artetxe et al., 2017). Let X and Z denote the word embedding matrices in the two languages where X_{i*} is the i-th source language word embedding and Z_{j*} is the j-th target language word embedding. Furthermore, let $D_{ij} = 1$ if i-th source language word is aligned with j-th target language word, and let $D_{ij} = 0$ otherwise, for the binary matrix D. The goal is to optimize for W^* as Equation 1.

$$W^* = \underset{W}{\arg\min} \sum_i \sum_j D_{ij} \left|\left| X_{i*}W - Z_{j*} \right|\right|^2 = \underset{W}{\arg\max} \operatorname{Tr}\left(XWZ^T D^T\right) \tag{1}$$

Since D is sparse, this optimization objective can be efficiently computed in linear time with respect to the number of dictionary entries (Artetxe et al., 2016).

As for the dictionary induction, Artetxe et al. (2017) uses a similarity measure of the dot product between source and target language embedding. This decision is because the similarity matrix XWZ^T can be efficiently vectorized and computed (Artetxe et al., 2017).

3.2 Experiments and results

Artetxe et al. (2017) evaluate the algorithm stated in Section 3.1 on bilingual lexicon induction tasks using **English-Italian** dataset from Dinu et al. (2015), **English-German** dataset from Baroni et al. (2009) and **English-Finnish** dataset from WMT 2016[1] tokenized by Stanford Tokenizer (Manning et al., 2014). The result is shown in Table 1, where the 5000 words seed dictionary, 25 words seed dictionary and trivially generated seed dictionaries of numerals are used for the evaluation.

	English-Italian			English-German			English-Finnish		
	5000	25	num	5000	25	num	5000	25	num
Mikolov et al. (2013)	34.93	0.00	0.00	35.00	0.00	0.07	25.91	0.00	0.00
Xing et al. (2015)	36.87	0.00	0.13	41.27	0.07	0.53	28.23	0.07	0.56
Zhang et al. (2016)	36.73	0.07	0.27	40.80	0.13	0.87	28.16	0.14	0.42
Artetxe et al. (2016)	39.27	0.07	0.40	41.87	0.13	0.73	30.62	0.21	0.77
Artetxe et al. (2017)	39.67	37.27	39.40	40.87	39.60	40.27	28.72	28.16	26.47

Table 1: Experiment result for bilingual lexicon induction task. (Artetxe et al., 2017)

[1] http://www.statmt.org/wmt16/translation-task.html

As shown in the comparison, Artetxe et al. (2017) yields significantly better performance for the 25 words seed dictionary and the trivially generated seed dictionaries of numerals than others.

3.3 Reflections

The framework proposed in the Artetxe et al. (2017) is simple but effective. Many machine translation algorithm relies on bilingual word embedding, with one example being Zou et al. (2013). However, bilingual word embedding generation requires lots of bilingual data in all previous works (Gouws et al., 2015; Luong et al., 2015a; Søgaard et al., 2015; Vulić and Moens, 2016; Mogadala and Rettinger, 2016; Mikolov et al., 2013; Artetxe et al., 2016), which makes machine translation with little bilingual data particularly difficult. However, Artetxe et al. (2017) shows a method that generates surprisingly good bilingual word embedding with almost no bilingual data, which may enable the possibility for the machine translation between language pairs without a large amount of data.

There are still some limitations of this work. Although Artetxe et al. (2017) claims that many languages contain the Arabic numerals, which are essential for generating the seed dictionary when we don't have any knowledge of the language pairs, many languages do not use it, such as Pirahã and Jarawara (Everett, 2012). Therefore, creating bilingual word embedding for them is still challenging.

4 Unsupervised machine translation

Lample et al. (2018b) identifies the common principles of unsupervised machine translation based on Lample et al. (2018a) and Artetxe et al. (2018). Additionally, by applying these principles, two new models are constructed and evaluated, one based on NMT and another based on PBSMT. Both models achieve better performances (Lample et al., 2018b).

4.1 Principles of unsupervised machine translation

Machine Translation with only monolingual data is an ill-posed task, since the association between source sentences and target sentences may not be unique (Lample et al., 2018b). However, lots of exciting progress has been made in recent years. Lample et al. (2018b) attempts to abstract away from the specific assumptions made by previous works and to identify the common principles. Specifically, unsupervised machine learning can be accomplished by three components as shown in Figure 2: (i) suitable initialization of the translation models, (ii) language modelling and (iii) iterative back-translation (Lample et al., 2018b).

Model initialization: This step expresses a natural prior over the space of solutions we expect to reach (Lample et al., 2018b). One may jump-start the process by leveraging the approximated translation of words, phrases or even sub-word units (Sennrich et al., 2016). For example, Klementiev et al. (2012) uses a provided bilingual dictionary whereas Lample et al. (2018a) and Artetxe et al. (2018) use dictionaries inferred from monolingual data. Although such initialization may not be entirely accurate, some of the original semantics are still preserved (Lample et al., 2018b).

Language modelling: The language model can be trained using monolingual data only on both source and target languages, expressing a data-driven prior about how sentences should be read in

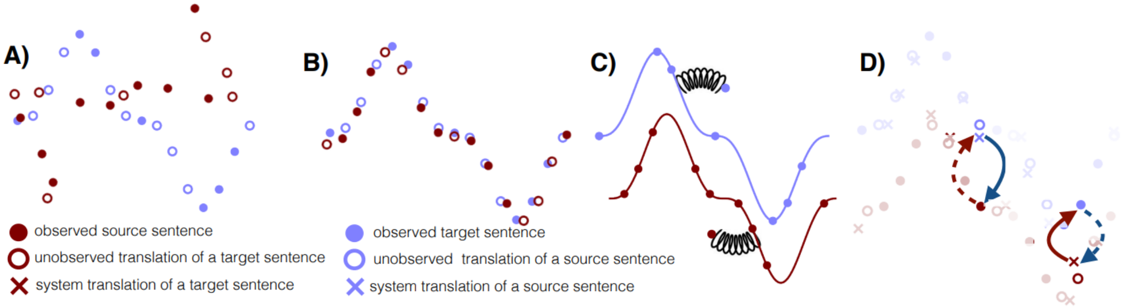

A)
B)
C)
D)
● observed source sentence
○ unobserved translation of a target sentence
× system translation of a target sentence
● observed target sentence
○ unobserved translation of a source sentence
× system translation of a source sentence

Figure 2: The illustration of 3 components: A) two monolingual datasets; B) initialization; C) language modelling; D) back-translation (Lample et al., 2018b).

each language. This step improves the quality of translation by forming local substitutions and word reordering (Lample et al., 2018b).

Back-translation: It is one of the most effective ways to leverage monolingual data in a semi-supervised setting (Sennrich et al., 2016; Lample et al., 2018b). It couples the source-to-target translation system with a model that translates backward from target to source language. Such a model can turn the unsupervised learning problem into a supervised learning task, though noisy. As the original model improves, the back-translation model also gets better, which gives a coupled system trained by an iterative algorithm (Lample et al., 2018b).

4.2 Unsupervised NMT

Lample et al. (2018b) proposes an unsupervised NMT method, which is based on Artetxe et al. (2018) and Lample et al. (2018a).

Model initialization: In the contrast of two prior work (Artetxe et al., 2018; Lample et al., 2018a), which relies on bilingual dictionaries, Lample et al. (2018b) is based on byte-pair encoding (BPE) (Sennrich et al., 2016). The BPE has two advantages over bilingual dictionaries, namely, reducing the dictionary size and eliminating the presence of unknown words. Additionally, instead of learning an explicit mapping between BPE in the source and target languages, Lample et al. (2018b) defines BPE tokens by jointly processing both of them, which naturally share many BPE tokens if they are related (Lample et al., 2018b).

Language modelling: The language modelling is accomplished by minimizing the Equation 2, where C is a noise model with some words dropped and swapped as in Lample et al. (2018a). $P_{s\rightarrow s}$ and $P_{t\rightarrow t}$ are the composition of the encoder and decoder both operating on the source and target sides, respectively (Lample et al., 2018b).

$$\mathcal{L}^{lm} = \mathbb{E}_{x\sim S}\left[-\log P_{s\rightarrow s}\left(x|C(x)\right)\right] + \mathbb{E}_{y\sim T}\left[-\log P_{t\rightarrow t}\left(y|C(y)\right)\right] \quad (2)$$

Back-translation: Let $u^*(y)$ denotes the sentence in the source language inferred from $y \in \mathscr{T}$ such that $u^*(y) \arg\max P_{t\to s}(u|y)$. Similarly, let $v^*(x)$ denote the sentence in the target language inferred from $x \in \mathscr{S}$ such that $v^*(x) = \arg\max P_{s\to t}(u|x)$. Using back-translation, the pair $(u^*(y), y)$ and $(x, v^*(x))$ can be used to train the two MT models by minimizing the loss function of Equation 3

$$\mathscr{L}^{back} = \mathbb{E}_{y\sim\mathscr{T}}[-\log P_{s\to t}(y|u^*(y))] + \mathbb{E}_{x\sim\mathscr{S}}[-\log P_{t\to s}(x|v^*(x))] \tag{3}$$

Sharing latent representations: A shared encoder representation can be viewed as an interlingua. This ensures that the benefits of language modelling can be transferred from noisy sources to translations, which will eventually help the accuracy of the NMT model. Sharing the encoder representations can be achieved by sharing all its parameters across two languages (Lample et al., 2018b).

4.3 Unsupervised PBSMT

Lample et al. (2018b) also attempts to perform unsupervised machine learning on Phrase-Based Statistical Machine Translation (PBSMT) system (Koehn et al., 2003). PBSMT is known to perform well on low-resource language pairs and therefore has the potential to surpass the NMT in an unsupervised learning setting (Lample et al., 2018b).

Model initialization: The initial phrase tables can be inferred from a bilingual dictionary built by Conneau et al. (2018), which only required monolingual corpora. The phrase table can be defined by Equation 4, where t_j is the j-th word in the target vocabulary, s_i is the i-th word in the source vocabulary, T is a hyper-parameter used to tune the peakiness of the distribution, and W is the rotation matrix mapping source embedding into target embedding (Conneau et al., 2018).

$$p(t_j|s_j) = \frac{e^{\frac{1}{T}\cos(e(t_j), We(s_i))}}{\sum_k e^{\frac{1}{T}\cos(e(t_k), We(s_i))}} \tag{4}$$

Language modelling: A smoothed n-gram model from Heafield (2011) is used for both the source and target languages.

Back-translation: First, construct a seed PBSMT using the unsupervised phrase tables and a language model on the target side. Then, this model can translate the source monolingual corpus into the target language. Once the training result is available, a new PBSMT can be trained reversely. Repeating these steps gives us the back-translation step (Lample et al., 2018b).

4.4 Experiments and results

Lample et al. (2018b) compare its method with 3 previous works (Artetxe et al., 2018; Lample et al., 2018a; Yang et al., 2018) using **English-French** and **English-German** datasets obtained from WMT monolingual News Crawl from years 2007 through 2017. The result is shown in Table 2.

There are two variants of the NMT model, based on LSTM (Hochreiter and Schmidhuber, 1997) and Transformer (Vaswani et al., 2017) respectively. There are also two variants of the PBSMT model, one without any back-translation (Iter. 0), and the other with back-translations (Iter. n). A combination of NMT and PBSMT is also attempted, one using NMT with PBSMT as back-translation (NMT + PBSMT) and the other using PBSMT with NMT as back-translation (PBSMT + NMT) (Lample et al., 2018b).

	en-fr	fr-en	en-de	de-en
Artetxe et al. (2018)	15.1	15.6	-	-
Lample et al. (2018a)	15.0	14.3	9.6	13.3
NMT (LSTM)	24.5	23.7	14.7	19.6
NMT (Transformer)	25.1	24.2	17.2	21.0
PBSMT (Iter. 0)	16.2	17.5	11.0	15.6
PBSMT (Iter. n)	28.1	27.2	17.9	22.9
NMT + PBSMT	27.1	26.3	17.5	22.1
PBSMT + NMT	27.6	27.7	20.2	25.2

Table 2: BLEU score for different models (Lample et al., 2018b).

4.5 Reflections

Although Lample et al. (2018b) is not the first one to propose the unsupervised MT, it greatly simplifies and refines the assumptions of all previous works into 3 basic principles, namely, model initialization, language modelling, and back-translation. Furthermore, combining these principles with PBSMT and NMT, the BLEU score has greatly improved over all past works. Lample et al. (2018b) provides a good method for unsupervised MT and creates a solid foundation for future work by proposing these basic principles.

However, compared to a supervised learning one (Wu et al., 2019), which has a BLEU score of 38.95 on **English-French** task, there is still lots of room for improving the translation accuracy of the current work.

5 Zero resource machine translation using teacher-student method

When there are no parallel corpora directly available between the source and target languages, there are roughly two methods have been proposed, *multilingual* (Firat et al., 2016; Johnson et al., 2017; Ha et al., 2016) and *pivot-based* (Cheng et al., 2016).

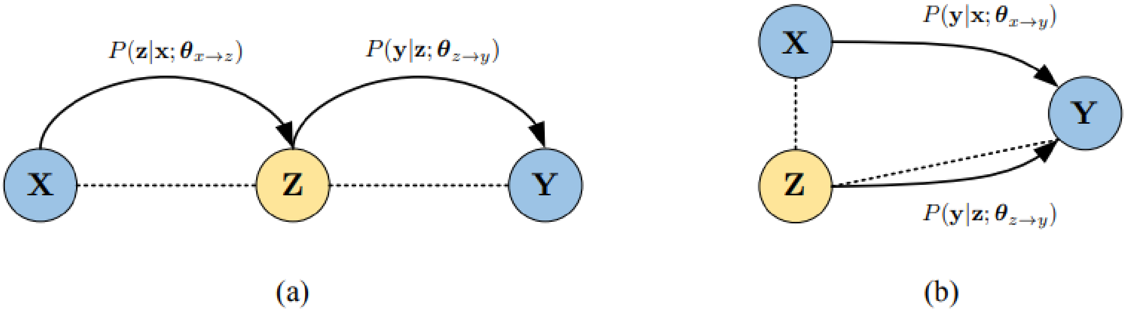
Chen et al. (2017) proposes a new method for this situation using a teacher-student approach, as shown in Figure 3. That is, to train a source-to-target NMT (“student”), an existing pivot-to-target NMT (“teacher”) can be leveraged by using a source-to-pivot parallel corpus. Compared with a pivot-based approach (Cheng et al., 2016), this method improves both translation accuracy and decoding efficiency (Chen et al., 2017).

5.1 Assumptions

Chen et al. (2017) makes two assumptions for this teacher-student approach.

Assumption 1 (sentence-level assumption): If a source sentence x is a translation of a pivot sentence z, then the probability of generating a target sentence y from x should be close to that from its counterpart z (Chen et al., 2017).

Assumption 2 (word-level assumption): If a source sentence x is a translation of a pivot sentence z, then the probability of generating a target word y from x should be close to that from its counter


$P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_{x\to z})$
$P(\mathbf{y}|\mathbf{z};\boldsymbol{\theta}_{z\to y})$
X
Z
Y
$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})$
X
Y
Z
$P(\mathbf{y}|\mathbf{z};\boldsymbol{\theta}_{z\to y})$

(a)

(b)

Figure 3: a) pivot-based approach; b) teacher-student approach (Chen et al., 2017).

z, given already obtained partial translation $y_{<j}$ (Chen et al., 2017).

These two assumptions have been empirically verified with a trilingual Europarl corpus. The sentence level and the word level KL divergence from the source-to-target model decrease over iteration when trained, indicating they have a small KL divergence (Chen et al., 2017).

5.2 Sentence-level teaching

Based on Assumption 1, the training objective of a source-pivot parallel corpus $D_{x,z}$ can be described as Equation 5 (Chen et al., 2017).

$$\mathscr{T}_{\text{SENT}}(\theta_{x\to y}) = \sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} KL\left(P(\mathbf{y}|\mathbf{z};\hat{\theta}_{z\to y}) \middle|\middle| P(\mathbf{y}|\mathbf{x};\theta_{x\to y})\right) \tag{5}$$

The goal is to minimize the training objective with the model parameter $\hat{\theta}_{z\to y}$, as described in Equation 6 (Chen et al., 2017).

$$\hat{\theta}_{x\to y} = \underset{\theta_{x\to y}}{\arg\min}\{\mathscr{T}_{\text{SENT}}(\theta_{x\to y})\} \tag{6}$$

5.3 Word-level teaching

Based on Assumption 2, the training objective can be further defined at the word level, as shown in Equation 7, which can be learned using stochastic gradient descent (Chen et al., 2017)

$$\mathscr{T}_{\text{WORD}}(\theta_{x\to y}) = \sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathbb{E}_{\mathbf{y}|\mathbf{z};\hat{\theta}_{z\to y}}\left[\sum_{j=1}^{|y|} KL\left(P(y|\mathbf{z},\mathbf{y}_{<j};\hat{\theta}_{z\to y}) \middle|\middle| P(y|\mathbf{x},\mathbf{y}_{<j};\theta_{x\to y})\right)\right] \tag{7}$$

The goal is to minimize the training objective as described in Equation 8.

$$\hat{\theta}_{x\to y} = \underset{\theta_{x\to y}}{\arg\min}\{\mathscr{T}_{\text{WORD}}(\theta_{x\to y})\} \tag{8}$$

5.4 Experiments and results

The experiments use the WMT corpus of Spanish, English and French, where English acts as the pivot. Chen et al. (2017) uses the sentence-level teaching method described in Section 5.3. The comparison is shown in Table 3. The new method proposed by Chen et al. (2017) demonstrates a significant improvement over the previous works.

	method	training			evaluation	
		es→en	en→fr	es→fr	WMT 2012	WMT 2013
Cheng et al. (2016)	pivot	6.78M	9.29M	-	24.60	-
Cheng et al. (2016)	likelihood	6.78M	9.29M	100K	25.78	-
Firat et al. (2016)	one-to-one	34.71M	65.77M	-	17.59	17.61
Firat et al. (2016)	many-to-one	34.71M	65.77M	-	21.33	21.19
Chen et al. (2017)	word-sampling	6.78M	9.29M	-	28.06	27.03

Table 3: BLEU score for different models over the WMT corpus (Chen et al., 2017).

5.5 Reflections

The method proposed by Chen et al. (2017) is ground-breaking since it is different from all methods in the previous works. Additionally, it outperforms other methods, such as the multilingual model or the pivot-based model, in terms of BLEU score, as demonstrated in Section 5.4.

On the other hand, although this method does not require any bilingual data between the source and target languages, lots of source-to-pivot and pivot-to-target bilingual data is still necessary for the training. Additionally, with these additional data used, the BLEU score is on par with the Lample et al. (2018b) method, which is an unsupervised learning framework.

However, considering this is the first attempt at this method, there may still be lots of room for improvement, both in terms of the amount of data used and translation accuracy.

6 Transfer learning for machine translation

Zoph et al. (2016) proposed a method to improve the performance of NMT using transfer learning when the training data is scarce. The idea is to train a neural network using a high-resource language pair (the *parent model*) first; then fix certain parameters and fine-tune the rest by a low-resource language pair (the *child model*) (Zoph et al., 2016). This method could significantly improve the BLEU score, compared with training by a low-resource language pair directly (Zoph et al., 2016).

6.1 Transfer Learning

Zoph et al. (2016) uses the following steps to accomplish the transfer learning. First, an NMT model on a large corpus of parallel data (e.g. French-English) is trained, which is the *parent model*. Next, create a new NMT model that has the same architecture as the *parent model*, which is the *child model*. Finally, the *child model* will be trained on a very small parallel corpus (e.g. Uzbek-English). The NMT model architecture is shown in Figure 4

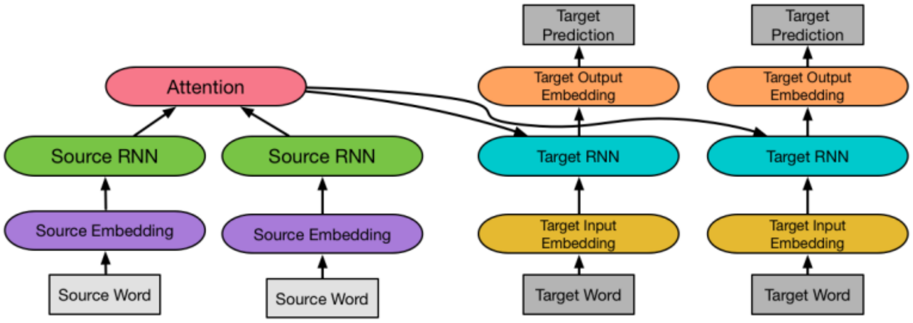


Figure 4: the NMT model architecture used for transfer learning (Zoph et al., 2016).

When the *child model* is trained, the initial weights are copied from the *parent model*. In the training process, the parameters which are likely to be useful across the languages are fixed, while other parameters get fine-tuned. In French-to-English to Uzbek-to-English example, the English word embedding is kept, while the French (or Uzbek) word embedding is free to adjust (Zoph et al., 2016).

6.2 Similar Parent Language

In order to test if a similar parent language will positively affect the transfer learning result, Zoph et al. (2016) creates a new language called FRENCH'. FRENCH' is exactly like French, but its vocabularies are randomly shuffled.

Choosing French-English as the parent language pair, the BLEU score improvement of different child language pairs can be compared to determine if the similarity between parent-child language pairs has effects on translation accuracy. The Uzbek-English and the FRENCH'-English are selected as the child language pairs for this purpose. If the BLEU score improvement of the FRENCH'-English case is higher than the Uzbek-English case, it suggests similar language pairs can yield a better transfer learning result (Zoph et al., 2016).

6.3 Experiments and results

Table 4 shows a comparison of different methods for creating MT models of some low-resource languages to English. The "NMT" row is the result of a model without any transfer learning. The "Xfer" row is the result of a model with transfer learning, and its parent model (French-English) is trained on WMT 2015 (Bojar et al., 2015). A significant improvement can be observed when a transfer learning method is used. The "Final" row is the result after ensembles of eight models and the usage of an unknown word replacement (Luong et al., 2015b), which can improve the BLEU score further (Zoph et al., 2016).

Additionally, to verify the assumption proposed in Section 6.2, another experiment is conducted, and the result is shown in Table 5. Since the BLEU score improvement of "FRENCH'-English" row (+6.7) is higher than the "Uzbek-English" row (+4.3), the hypothesis made in Section 6.2 is verified.

method	Hausa	Turkish	Uzbek	Urdu
NMT	16.8	11.4	10.7	5.2
Xfer	21.3	17.0	14.4	13.8
Final	24.0	18.7	16.8	14.5

Table 4: BLEU score for translating low-resource languages into English (Zoph et al., 2016).

language pair	parent	train size	BLEU
Uzbek-English	None	1.8m	10.7
	French-English	1.8m	15.0 (+4.3)
FRENCH'-English	None	1.8m	13.3
	French-English	1.8m	20.0 (+6.7)

Table 5: Experiment to see if similar parent languages can produce better results (Zoph et al., 2016).

6.4 Reflections

Zoph et al. (2016) applies the transfer learning technique in the field of MT and improves the performance of the NMT model of low-resource language pairs. This can be extremely helpful for languages which do not have sufficient data. Zoph et al. (2016) has also suggested that similar parent-child pairs will yield better results for transfer learning, which can be a clue on how to improve the effectiveness of transfer learning in MT further.

However, Zoph et al. (2016) does not conduct an experiment when the parent and the child model has different source and target languages, which could lead to a deeper understanding of the transfer learning in MT. Additionally, although Zoph et al. (2016) suggests similar language pairs for transfer learning will yield a better result, it fails to define what exactly is the similarity between the child language pair and the parent language pair. Some additional work on this may help us better pick the parent model.

7 Discussion

The four papers we discussed above suggest three different pathways to train an MT model with limited parallel corpus, a) turn supervised learning tasks into unsupervised or semi-supervised learning tasks (Artetxe et al., 2017; Lample et al., 2018b); b) use a reinforcement learning technique (Chen et al., 2017); c) use a transfer learning technique (Zoph et al., 2016).

Lample et al. (2018b)'s method is the most universal one because it does not require any training data to build an MT model, but its performance heavily relies on the similarity between the source and target languages (Artetxe et al., 2017). Additionally, other methods may further boost the performance. For example, instead of using a randomly initialized model, one may choose to use an MT model of parent language pair and possibly fine-tuned by some limited amount of bilingual corpus (Chen et al., 2017). One could also use the teacher-student method to improve the performance further with a trained model, if some bilingual data are available to some pivot languages (Zoph et al., 2016).

However, we shall notice that the translation accuracy of all existing methods discussed above is significantly worse than the supervised learning approach (Wu et al., 2019). Therefore it is an open question if a machine translation model with limited bilingual data could ever achieve the similar performance to a supervised learning method with sufficient bilingual data.

8 Conclusion

In order to answer the question, “How can we train a machine translation model without sufficient bilingual data”, this literature review discusses four related works, each with their unique methods. Section 3 demonstrates a semi-supervised learning method to learn a bilingual embedding; Section 4 shows an unsupervised learning method for training an MT model; Section 5 gives a teacher-student method for creating an MT model without bilingual data between the source and target languages; and Section 6 suggests a transfer learning technique to improve MT model performance.

All these works indicate significant progress has been made in how to train an MT model without sufficient data. However, we should also realize that the translation accuracy of these methods are still far behind a human translator. Therefore, further research is required in this topic.

References

Booth Andrew. 1953. Machine translation. *Computers and Automation*, 2:6–8.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. *International Conference on Learning Representation (ICLR)*.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer, et al. 1993. The mathematics of statistical machine translation: Parameter estimation.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada. Association for Computational Linguistics.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016. Neural machine translation with pivot languages. *arXiv preprint arXiv:1611.04928*.

J. P. Cleave. 1957. A type of program for mechanical translation. *Mechanical Translation*, 4(3):54–58.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

G Dinu, A Lazaridou, and M Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. iclr 2015. In *Workshop Track*.

Caleb Everett. 2012. A closer look at a supposedly anumeric language. *International Journal of American Linguistics*, 78(4):575–590.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille, France. PMLR.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–140, Avignon, France. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. *International COnference on Learning Representations (ICLR)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015b. Addressing the rare word problem in neural machine 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)*, pages 11–19, Beijing, China. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation.

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–702, San Diego, California. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual NLP. 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)*, pages 1713–1722, Beijing, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2019. Google's neural machine translation system: Bridging the gap between human and machine translation. arxiv 2016. *Transcations of the Association for Computational Linguistics*, pages 339–351.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55, Melbourne, Australia. Association for Computational Linguistics.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317, San Diego, California. Association for Computational Linguistics.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1527–1536, Denver, Colorado. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA. Association for Computational Linguistics.