Exploring Cross-Lingual Transfer Learning with Unsupervised Machine Translation

Chao Wang *
York University
Toronto, Canada
chwang@eecs.yorku.ca

Judith Gaspers
Amazon Alexa AI
Aachen, Germany
gaspers@amazon.com

Quynh Do
Amazon Alexa AI
Aachen, Germany
doquynh@amazon.com

Hui Jiang
York University
Toronto, Canada
hj@eecs.yorku.ca

Abstract

In Natural Language Understanding (NLU), to facilitate Cross-Lingual Transfer Learning (CLTL), especially CLTL between distant languages, we integrate CLTL with Machine Translation (MT), and thereby propose a novel CLTL model named Translation Aided Language Learner (TALL). TALL is constructed as a standard transformer, where the encoder is a pre-trained multilingual language model. The training of TALL includes an MT-oriented pre-training and an NLU-oriented fine-tuning. To make use of unannotated data, we implement the recently proposed Unsupervised Machine Translation (UMT) technique in the MT-oriented pre-training of TALL. The experimental results show that the application of UMT enables TALL to consistently achieve better CLTL performance than our baseline model, which is the pre-trained multilingual language model serving as the encoder of TALL, without using more annotated data, and the performance gain is relatively prominent in the case of distant languages.

1 Introduction

Virtual assistants, such as Amazon Alexa, Apple Siri, and Google Assistant, are increasingly popular due to the convenience they bring to customers. A core function of virtual assistants is Natural Language Understanding (NLU), which is a combo of slot filling and intent classification. NLU models behind virtual assistants are generally trained in a supervised manner, which requires a large amount of annotated data. Collecting annotated data is not a big deal for high-resource languages, but difficult or even impossible for low-resource languages. As a result, when ported to a low-resource language, an NLU model may suffer from the so-called "data hungriness" (van der Ploeg et al., 2014). This problem can be alleviated by conducting Cross-Lingual Transfer Learning (CLTL) (Yarowsky et al., 2001), where annotated data in a high-resource source language is used to bootstrap an NLU model aimed at a low-resource target language.

The key to CLTL is to learn a shared representation space for the given source-target language pair. A traditional way to achieve this goal is to leverage cross-lingual word embeddings, which are obtained by mapping the words in both languages to a shared word embedding space (Zhang et al., 2017; Conneau et al., 2017; Artetxe et al., 2018a; Chen et al., 2018; Chen and Cardie, 2018; Chen et al., 2019). However, most studies on this topic only consider similar languages (e.g. English-German) but ignore distant languages (e.g. English-Japanese), since it is more challenging to conduct CLTL between distant languages than between similar languages. Recently, contextualized word embeddings generated by pre-trained language models have shown significant advantages over ordinary word embeddings (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019). For the purpose of CLTL, many efforts have been made to develop multilingual variants of pre-trained language models. These efforts have in turn brought about pre-trained multilingual language models, each of which is pre-trained on a multilingual corpus so that the learned representation space is not only rich in contextual clues but also shared by all the involved languages (Mulcaire et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020). However, in this pre-training, the collection of the

*Work done during internship at Amazon Alexa AI.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2011–2020

multilingual corpus is not obviously biased to any language, thus in the learned representation space, similar languages are still similar to each other, and distant languages are still distant from each other. As a result, although pre-trained multilingual language models have greatly promoted CLTL, it is still more challenging to conduct CLTL between distant languages than between similar languages. This opinion has been verified by several empirical studies on a popular pre-trained multilingual language model named Multilingual BERT (M-BERT) (Devlin et al., 2019), where the CLTL performance of M-BERT between similar languages is decent, but that between distant languages is still far from satisfactory (Pires et al., 2019; Wu and Dredze, 2019; Karthikeyan et al., 2020).

From our point of view, CLTL can be analogized to the process of a human being learning a foreign language, where the prior knowledge on the native language plays an important role. Language educators believe that a foreign language learner can benefit a lot from translation, since translation not only involves all aspects of foreign language learning but also helps to enhance the correlation between the native language and the foreign language (Witte et al., 2009). According to our observation and experience, this is especially the case when the native language and the foreign language are distant from each other. Inspired by these thoughts, to facilitate CLTL, especially CLTL between distant languages, we propose a novel CLTL model named Translation Aided Language Learner (TALL), where CLTL is integrated with Machine Translation (MT). Specifically, we adopt a pre-trained multilingual language model, which is now recognized as the state of the art in CLTL, as our baseline model, and construct TALL by appending a decoder to it. On this basis, we directly fine-tune the baseline model as an NLU model to conduct CLTL, but put TALL through an MT-oriented pre-training before its NLU-oriented fine-tuning. We believe that the MT-oriented pre-training can help TALL to enhance the correlation between the given source-target language pair in its representation space, and thus can make CLTL easier to conduct in its NLU-oriented fine-tuning, especially in the case of distant languages. To make use of unannotated data, which is not only large in amount but also available for every language, we implement the recently proposed Unsupervised Machine Translation (UMT) (Artetxe et al., 2018b; Lample et al., 2018a; Yang et al., 2018; Lample et al., 2018b; Liu et al., 2020) technique in the MT-oriented pre-training of TALL.

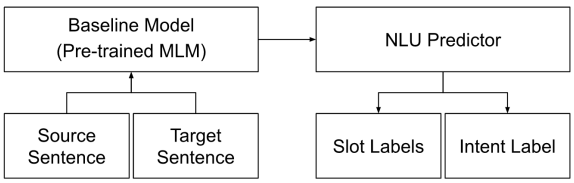

Figure 1: The NLU-oriented fine-tuning of our baseline model. MLM means multilingual language model.

To verify the effectiveness of TALL, we carry out a series of experiments to compare the CLTL performance of TALL with that of the baseline model. In these experiments, we address not only CLTL tasks between similar languages but also those between distant languages. For each given CLTL task, we separately use two popular pre-trained multilingual language models for model construction. To implement UMT, we collect unannotated sentences from Wikipedia dumps. To conduct CLTL, we separately collect annotated sentences from two multilingual NLU datasets. The experimental results show that the application of UMT enables TALL to consistently achieve better CLTL performance than the baseline model without using more annotated data, and the performance gain is relatively prominent in the case of distant languages.

2 Translation Aided Language Learner

2.1 Task Definition

NLU is a combo of slot filling and intent classification. Given a sentence x consisting of m words $\{w_1, \ldots, w_m\}$, slot filling is to predict a slot label y_i^σ for each word w_i, and intent classification is to predict an intent label y^ι for x. In this paper, NLU models are required to be trained under a zero-shot CLTL scenario, where annotated sentences in the given source language are used for model optimization, while those in the given target language are used for model evaluation.

2.2 Baseline Model

A transformer (Vaswani et al., 2017) is a sequence-to-sequence model consisting of an encoder and a decoder. A main feature of transformers is that they use multi-head self-attention and multi-head cross-attention to model dependencies in sequential data. These attention mechanisms enable transformers to extract long-term contextual clues from text. As a result, transformers have been intensively used in

transfer learning to develop pre-trained language models, which generate contextualized word embeddings. For example, some pre-trained language models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), are implemented as transformer encoders, and some other ones, such as the GPT family (Radford et al., 2018, 2019), are implemented as transformer decoders.

As a sub-field of transfer learning, CLTL has witnessed the wide application of transformers in developing pre-trained multilingual language models. Most of the existing pre-trained multilingual language models, such as M-BERT, XLM (Conneau and Lample, 2019), and XLM-RoBERTa (XLM-R) (Conneau et al., 2020), are implemented as transformer encoders. Actually, these pre-trained multilingual language models are the multilingual variants of BERT and RoBERTa, since each of them is identical to either BERT or RoBERTa except being pre-trained on a multilingual corpus. The representation space learned through this pre-training is not only rich in contextual clues but also shared by all the involved languages. Therefore, in theory, each of these pre-trained multilingual language models can be simply fine-tuned to address any CLTL task between its involved languages.

The pre-trained multilingual language models mentioned above are now recognized as the state of the art in CLTL. To push the state of the art, we adopt one of them as our baseline model, and fine-tune it as an NLU model to conduct CLTL. As shown in Figure 1, in this NLU-oriented fine-tuning, we feed each given sentence to the baseline model, and feed the final hidden states of the baseline model to an NLU predictor. Since the baseline model is fitted with a sub-word tokenizer, a given sentence x consisting of m words $\{w_1, \ldots, w_m\}$ is tokenized into n tokens ($n \geqslant m$) such that the baseline model generates n final hidden states $\{h_1, \ldots, h_n\}$. For slot filling, the NLU predictor first performs an average pooling on the final hidden states related to each word w_i, and then uses a dense layer with a softmax normalization to map the pooling result to a slot distribution for w_i:

$$\mathrm{p}(y_i^{\sigma} \mid x) = \mathrm{softmax}\big(W^{\sigma}\mathrm{f_a}(h_{k_i}, \ldots, h_{l_i}) + b^{\sigma}\big)$$

where W^{σ} is a trainable weight, b^{σ} is a trainable bias, k_i and l_i separately represent the start position and end position of the final hidden states related to w_i, and $\mathrm{f_a}(\cdot)$ represents average pooling. For intent classification, the NLU predictor first performs an average pooling on all the final hidden states, and then uses another dense layer with another softmax normalization to map the pooling result to an intent distribution for x:

$$\mathrm{p}(y^{\iota} \mid x) = \mathrm{softmax}\big(W^{\iota}\mathrm{f_a}(h_1, \ldots, h_n) + b^{\iota}\big)$$

where W^{ι} is a trainable weight, and b^{ι} is a trainable bias. On this basis, for model optimization, we minimize the following joint loss through stochastic gradient descent on annotated sentences in the given source language:

$$\mathcal{L}_{nlu} = -\log\Big(\prod_{i=1}^{m} \mathrm{p}(y_i^{\sigma} \mid x) \cdot \mathrm{p}(y^{\iota} \mid x)\Big)$$

For model evaluation, we infer the baseline model on annotated sentences in the given target language to measure three evaluation metrics, namely Slot F1, Intent Accuracy, and Semantic Accuracy (i.e. sentence-level joint accuracy).

2.3 Proposed Model

Since the baseline model is pre-trained on a multilingual corpus, all its involved languages are correlated with each other in its representation space. Normally, the larger such correlation between languages, the easier it is to conduct CLTL. To equally treat all possible CLTL tasks, the multilingual corpus used in the pre-training of the baseline model is collected in a subtle way that is not obviously biased to any language. However, there are two side effects of doing so. On the one hand, instead of focusing on a specific CLTL task, the baseline model pays equal attention to all possible CLTL tasks. On the other hand, in the representation space of the baseline model, the correlation between languages is proportional to their linguistic similarity, or in other words, similar languages are still similar to each other, and distant languages are still distant from each other. This implies that the CLTL ability of the baseline model can be pertinently improved for each given CLTL task, and the room for improvement is relatively large when the CLTL task is between distant languages.

To pertinently improve the CLTL ability of the baseline model for each given CLTL task, we would like to transform its representation space, which is used for all possible CLTL tasks, into a specialized one, where the correlation between the given source-target language pair is expressly enhanced. This goal can be achieved by resorting to MT, since

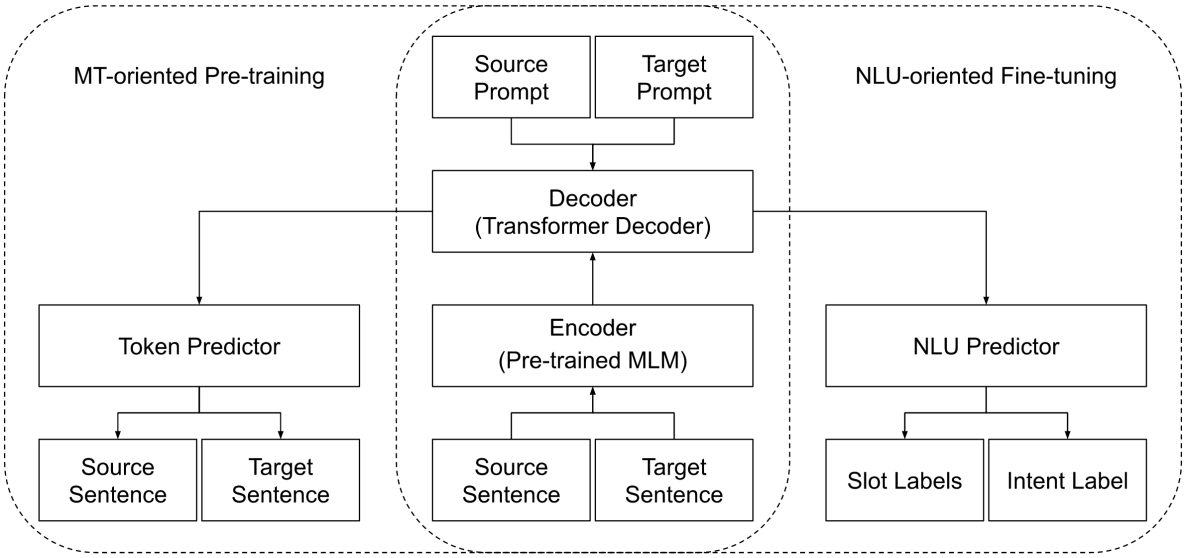

Figure 2: The MT-oriented pre-training and NLU-oriented fine-tuning of our proposed Translation Aided Language Learner (TALL).

translation is the most direct way to correlate languages with each other. As shown in Figure 2, for MT to be workable, we treat the baseline model as an encoder and append a decoder to it. Considering that the encoder is implemented as a transformer encoder, we implement the decoder as a transformer decoder to keep the model architecture consistent. Besides, as in Vaswani et al. (2017), we also share the token embeddings between the encoder and the decoder. The resulting new model can be seen as a standard transformer, where the encoder is a pre-trained multilingual language model. We expect this model to learn the correlation between the given source-target language pair by addressing a two-way MT task, and thus name it Translation Aided Language Learner (TALL).

Before conducting CLTL with TALL, we need to pre-train it as a two-way MT model that translates between the given source-target language pair. As shown in Figure 2, in this MT-oriented pre-training, we feed each given sentence to the encoder, feed a prompt for the translated sentence to the decoder, and feed the final hidden states of the decoder to a token predictor. Given a sentence x and a prompt x' for the translated sentence, suppose the decoder generates a final hidden state h'_i for the i-th token in x', then h'_i can be seen as a memory of both x and the first i tokens in x'. The token predictor uses a dense layer with a softmax normalization to map this memory to a token distribution for the position i in the translated sentence:

$$\mathrm{p}(y_i^\tau \mid x, x') = \mathrm{softmax}(W^\tau h'_i + b^\tau)$$

where W^τ is a trainable weight tied to the token embeddings, and b^τ is a trainable bias. Since two-way MT requires the translated sentence to be in either the source language or the target language, which depends on the current direction, we extend the token vocabulary with two language identifiers, which separately represent the two languages, and thereby inform the decoder about the currently required language by setting the first token of x' to the corresponding language identifier. By the way, since the token vocabulary is highly multilingual, most probabilities in the above token distribution are for the tokens beyond the given source-target language pair and thus make no sense. Therefore, we ignore these probabilities when inferring TALL to generate translated sentences.

By convention, the training of MT models is supervised and thus requires parallel corpora. However, parallel corpora are generally expensive to collect, which makes them scarce or even unavailable for many source-target language pairs. Since TALL is designed to be a general-purpose CLTL model, a supervised training on parallel corpora is not applicable to its MT-oriented pre-training. Recently, an unsupervised training technique for MT models, which is named Unsupervised Machine Translation (UMT), has been proposed. Instead of relying on parallel corpora, UMT relies on monolingual corpora of unannotated sentences. This is attractive to us, since a large amount of unannotated sentences are always available for every language. Therefore, we implement the UMT training recipe proposed

by Lample et al. (2018b) in the MT-oriented pre-training of TALL. Specifically, for model optimization, we collect a source-language corpus S and a target-language corpus T, each of which is a set of unannotated sentences. On this basis, we measure the following two losses:

- **Denoising auto-encoding loss.** As in Lample et al. (2018a), we implement a noise injector $\mathrm{f_n}(\cdot)$, which injects noise to each given sentence by randomly dropping and swapping its tokens. For each source-language sentence $s \in S$, we first run the noise injector to obtain a noise-injected sentence $\mathrm{f_n}(s)$, which can be seen as a sentence in a different language, and then use TALL to translate $\mathrm{f_n}(s)$ to the source language, the expected result of which is s. Besides, we also perform this process on each target-language sentence $t \in T$. This is the so-called "denoising auto-encoding", whose loss is defined as the cross-entropy loss on recovering the original sentences from the noise-injected sentences:

$$\begin{aligned}\mathcal{L}_{dae} =& \mathbb{E}_{s\in S}\big[-\log \mathrm{p}\big(s \mid \mathrm{f_n}(s)\big)\big] + \\ & \mathbb{E}_{t\in T}\big[-\log \mathrm{p}\big(t \mid \mathrm{f_n}(t)\big)\big]\end{aligned}$$

- **Back-translation loss.** Let us use $\mathrm{f_m}(\cdot)$ to represent the inference of TALL, which translates each given sentence to its opposite language in the given source-target language pair. For each source-language sentence $s \in S$, we first infer TALL to obtain a TALL-translated sentence $\mathrm{f_m}(s)$, which is in the target language, and then use TALL to translate $\mathrm{f_m}(s)$ to the source language, the expected result of which is s. Besides, we also perform this process on each target-language sentence $t \in T$. This is the so-called "back-translation", whose loss is defined as the cross-entropy loss on recovering the original sentences from the TALL-translated sentences:

$$\begin{aligned}\mathcal{L}_{bt} =& \mathbb{E}_{s\in S}\big[-\log \mathrm{p}\big(s \mid \mathrm{f_m}(s)\big)\big] + \\ & \mathbb{E}_{t\in T}\big[-\log \mathrm{p}\big(t \mid \mathrm{f_m}(t)\big)\big]\end{aligned}$$

We sum up the above two losses to obtain a joint loss, and thereby minimize the joint loss through stochastic gradient descent. For model evaluation, we collect another source-language corpus and another target-language corpus, each of which is also a set of unannotated sentences. On this basis, we implement the round-trip translation trick proposed by Lample et al. (2018a), where we first translate each given sentence to its opposite language in the current source-target language pair, and then translate the resulting sentence to the original language. By inferring TALL, we perform this process on all the sentences in the above two corpora to obtain two reconstructed corpora. Thereby, we measure the BLEU score between the two original corpora and the two reconstructed corpora to evaluate the translation performance of TALL.

The above MT-oriented pre-training guarantees that TALL can learn a representation space, where the given source-target language pair are expressly correlated with each other. As a result, it will be easier to conduct CLTL with the pre-trained TALL than with the baseline model. This is especially the case when the given source-target language pair are distant from each other, since translating between distant languages reveals more knowledge than translating between similar languages. However, considering that the representation space of TALL is co-carried by the encoder and the decoder, we have to fine-tune them together as an NLU model when we conduct CLTL with the pre-trained TALL. To this end, we implement the fine-tuning approach of BART (Lewis et al., 2020) in the NLU-oriented fine-tuning of TALL. Specifically, as shown in Figure 2, we feed each given sentence to the encoder, feed this sentence again as a prompt to the decoder with the corresponding language identifier prefixed to it, and feed the final hidden states of the decoder except the last one to the NLU predictor. On this basis, both the model optimization and the model evaluation remain the same as in the NLU-oriented fine-tuning of the baseline model.

3 Related Works

Cross-lingual word embeddings. A traditional way to conduct CLTL is to leverage cross-lingual word embeddings, which are usually learned in an unsupervised manner. For example, Zhang et al. (2017) formulate the learning of cross-lingual word embeddings as an adversarial game, and explore several adversarial training methods to implement it. Conneau et al. (2017) first use adversarial training to learn a linear mapping from the word embeddings of a source language to those of a target language, and then use a Procrustes solution to refine it. Artetxe et al. (2018a) first use an unsupervised initialization scheme to create an initial

mapping, and then use a self-learning procedure to iteratively improve it. Chen et al. (2018) propose a language-adversarial training method, and use it to address cross-lingual sentiment classification. Besides, there are also several studies on multilingual word embeddings. For example, Chen and Cardie (2018) propose an unsupervised approach to learning multilingual word embeddings, which directly exploits the relations between all the involved languages. On this basis, Chen et al. (2019) propose a multi-source CLTL model, which not only uses adversarial training to learn language-invariant features, but also uses a mixture-of-experts method to dynamically exploit the similarity between a target language and multiple source languages.

Pre-trained multilingual language models. The currently dominant way to conduct CLTL is to fine-tune pre-trained multilingual language models, which are multilingual variants of pre-trained language models, and are each pre-trained on a multilingual corpus. For example, Mulcaire et al. (2019) propose Rosita as a multilingual variant of ELMo, and pre-train it on a multilingual corpus covering 3 languages. Devlin et al. (2019) propose M-BERT as a multilingual variant of BERT, and pre-train it on a multilingual corpus covering 104 languages. Conneau and Lample (2019) propose XLM as a multilingual variant of BERT, and pre-train it on a multilingual corpus covering 15 languages. Conneau et al. (2020) propose XLM-R as a multilingual variant of RoBERTa, and pre-train it on a multilingual corpus covering 100 languages. Besides, there are also several empirical studies on M-BERT. For example, Pires et al. (2019) carry out a large number of probing experiments to verify and interpret the zero-shot CLTL performance of M-BERT. Wu and Dredze (2019) explore the zero-shot CLTL potential of M-BERT on 5 downstream tasks covering 39 languages. Karthikeyan et al. (2020) provide a comprehensive study on the contribution of each component of M-BERT to its CLTL ability, which focuses on the impact of linguistic properties of the languages, the architecture of the model, and the learning objectives.

UMT technique. The UMT technique is aimed at reducing the reliance of MT models on parallel corpora. For example, Artetxe et al. (2018b) construct an MT model consisting of a language-invariant encoder and two language-specific decoders, and train it on a non-parallel corpus through denoising auto-encoding and back-translation. Lample et al. (2018a) construct an MT model consisting of a language-invariant pair of encoder and decoder, and train it on a non-parallel corpus not only through denoising auto-encoding and back-translation but also through adversarial training. Yang et al. (2018) construct an MT model consisting of two pairs of encoder and decoder, which partially share their parameters, and train it on a non-parallel corpus not only through denoising auto-encoding and back-translation but also through adversarial training. Lample et al. (2018b) propose a simple but effective approach based on the above works, where the constructed MT model only consists of a language-invariant pair of encoder and decoder, and its training on a non-parallel corpus only requires denoising auto-encoding and back-translation. Liu et al. (2020) first pre-train BART on a non-parallel multilingual corpus through denoising auto-encoding, and then fine-tune the pre-trained BART for downstream MT tasks.

4 Verification Experiments

4.1 Experimental Settings

CLTL tasks. For generality, we address not only CLTL tasks between distant languages but also those between similar languages. Specifically, we separately conduct CLTL between three source-target language pairs, which include two distant language pairs, namely English-Japanese and German-Japanese, and one similar language pair, namely English-German.

Pre-trained multilingual language models. For compatibility, we use different pre-trained multilingual language models for model construction. Specifically, for each given CLTL task, we separately use two popular pre-trained multilingual language models, namely M-BERT (base and cased) and XLM-R (base), to construct both the baseline model and TALL.

Training data. For practicality, we adopt large-scale corpora and real-world datasets as training data. Specifically, to implement UMT, we collect a source-language corpus of 1M unannotated sentences and a target-language corpus of 1M unannotated sentences from Wikipedia dumps for model optimization, and also collect a source-language corpus of 10K unannotated sentences and a target-language corpus of 10K unannotated sentences from Wikipedia dumps for model evaluation. To conduct CLTL, we collect annotated sentences in real-world domains from two multilingual NLU

CLTL Task	Pre-trained MLM	BLEU Score	Slot F1			Intent Accuracy			Semantic Accuracy		
			BSL	TALL	Gain	BSL	TALL	Gain	BSL	TALL	Gain
EN-JA (distant)	M-BERT	41.19	56.78	60.87	7.20%	80.37	83.21	3.53%	14.56	16.39	12.57%
	XLM-R	39.83	58.21	63.19	8.56%	81.19	83.92	3.36%	16.58	18.47	11.40%
DE-JA (distant)	M-BERT	38.54	51.28	54.56	6.40%	79.08	81.54	3.11%	11.71	13.24	13.07%
	XLM-R	35.11	50.36	53.68	6.59%	78.43	81.12	3.43%	12.76	14.61	14.50%
EN-DE (similar)	M-BERT	71.21	70.42	72.39	2.80%	89.39	91.16	1.98%	36.53	38.64	5.78%
	XLM-R	72.91	75.29	77.14	2.46%	92.82	94.33	1.63%	44.86	47.25	5.33%

Table 1: The translation performance of TALL on Wikipedia and the CLTL performance of both the baseline model and TALL on MultiATIS++. EN means English. JA means Japanese. DE means German. BSL means the baseline model. Gain means the CLTL performance gain of TALL over the baseline model. The gain numbers are in percentage and calculated as $(\text{TALL} - \text{BSL}) \div \text{BSL}$.

CLTL Task	Pre-trained MLM	Slot F1 Gain	Intent Accuracy Gain	Semantic Accuracy Gain
EN-JA (distant)	M-BERT	53.61%	36.75%	59.45%
	XLM-R	50.96%	31.60%	71.86%
DE-JA (distant)	M-BERT	47.75%	31.46%	55.55%
	XLM-R	58.49%	34.45%	69.19%
EN-DE (similar)	M-BERT	10.42%	9.69%	18.00%
	XLM-R	12.75%	7.81%	23.87%

Table 2: The CLTL performance gain of TALL over the baseline model on the multi-domain multilingual NLU dataset.

datasets. The first multilingual NLU dataset is MultiATIS++ (Xu et al., 2020), which is an extension to Multilingual ATIS (Upadhyay et al., 2018). It provides 5K annotated sentences for each language, which are all in the domain of airline travel. The second multilingual NLU dataset is a multi-domain dataset collected from a virtual assistant. It provides 100K annotated sentences for each language, which are evenly distributed in five domains, namely music, notifications, smart home, weather, and books. By the way, in the above two multilingual NLU datasets, each word is annotated with a slot label in the B-I-O format, and each sentence is annotated with an intent label.

4.2 Implementation details.

We use WikiExtractor (Attardi, 2015) to extract paragraphs from Wikipedia dumps, use Stanza (Qi et al., 2020) to split paragraphs into sentences, use HuggingFace's Transformers (Wolf et al., 2019) to tokenize sentences into tokens and load pre-trained multilingual language models, and use PyTorch (Paszke et al., 2019) to implement both the baseline model and TALL. For model optimization, we apply an AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of 0.0001, a weight decay factor of 0.01, and a batch size of 64 in the MT-oriented pre-training of TALL, and apply another AdamW optimizer with an initial learning rate of 0.00005, a weight decay factor of 0.01, and a batch size of 256 in the NLU-oriented fine-tuning of both the baseline model and TALL. After each epoch, we evaluate the validation performance, which refers to BLEU score in the MT-oriented pre-training of TALL and Semantic Accuracy in the NLU-oriented fine-tuning of both the baseline model and TALL. If the obtained performance number is improved, we save the model, otherwise we cancel the finished epoch by restoring the model to the last saved version. We decay the learning rate by 0.5 after each cancelled epoch, and terminate the model optimization after the 5th cancelled epoch. For model evaluation, we use NLTK (Loper and Bird, 2004) to measure BLEU score, and use the evaluation script for the CoNLL-2000 shared task to measure Slot F1.

4.3 Experimental Results

As shown in Table 1, we carry out a series of experiments on the unannotated sentences collected from Wikipedia and the annotated sentences collected from MultiATIS++. Each of these experiments is aimed at a different combination of CLTL task and pre-trained multilingual language model, and includes the corresponding MT-oriented pre-training of TALL and the corresponding NLU-oriented fine-tuning of both the baseline model and TALL. On this basis, we first evaluate the translation performance of TALL in its MT-oriented pre-training, then evaluate the CLTL performance of both the baseline model and TALL in their NLU-oriented fine-tuning, and finally calculate the CLTL performance gain of TALL over the baseline model in percentage. Besides, as shown in Table 2, we

also repeat the NLU-oriented fine-tuning of both the baseline model and TALL on the annotated sentences collected from the multi-domain multilingual NLU dataset, and thereby obtain another CLTL performance gain of TALL over the baseline model. The experimental results show that due to the application of UMT in the MT-oriented pre-training, TALL consistently achieves better CLTL performance than the baseline model in the NLU-oriented fine-tuning without using more annotated data, and the performance gain is relatively prominent in the case of distant languages.

4.4 Ablation Study

Denoising auto-encoding. In the MT-oriented pre-training of TALL, we try to discard the denoising auto-encoding loss and only minimize the back-translation loss in the UMT training. As a result, we observe that TALL achieves a very poor translation performance and a very poor CLTL performance. This implies that TALL learns little cross-lingual knowledge through the UMT training without denoising auto-encoding.

Back-translation. In the MT-oriented pre-training of TALL, we also try to discard the back-translation loss and only minimize the denoising auto-encoding loss in the UMT training. As a result, we observe that TALL achieves an almost perfect translation performance but a very poor CLTL performance. This is because the UMT training without back-translation makes TALL a copying model instead of an MT model, and a copying model can work perfectly in the model evaluation based on round-trip translation.

BART-style fine-tuning. In the NLU-oriented fine-tuning of TALL, instead of following the fine-tuning approach of BART, we try to discard the decoder and only fine-tune the encoder following the way we fine-tune the baseline model. As a result, we observe a very poor CLTL performance. This implies that the decoder of TALL is necessary for its NLU-oriented fine-tuning.

5 Further Discussion

Is a startup supervision necessary for the back-translation? In several existing UMT training recipes, the back-translation is supervised during its startup stage, where the supervision is provided by replacing the inference of TALL with a bilingual dictionary (Lample et al., 2018a; Artetxe et al., 2018b). This startup supervision is aimed at initializing a shared representation space for the given source-target language pair. However, since the encoder of TALL is a pre-trained multilingual language model, TALL already possesses a properly initialized representation space, which is shared by all the involved languages, and thus does not need a startup supervision. Actually, we tried to use a parallel corpus generated by a naive MT model to provide a startup supervision, which is equivalent to using a bilingual dictionary, but did not observe any translation performance gain.

How does the UMT training affect the CLTL performance? The UMT training uses the denoising auto-encoding and the back-translation to enhance the correlation between the given source-target language pair in the representation space of TALL. Since the encoder of TALL is a pre-trained multilingual language model, the representation space of TALL can be seen as an extension to that of the pre-trained multilingual language model. In the representation space of the pre-trained multilingual language model, similar languages have been more correlated with each other than distant languages. That is to say, in the representation space of TALL, there is more potential to enhanced the correlation between distant languages than between similar languages. As a result, although the CLTL performance between similar languages is better than that between distant languages, the CLTL performance gain between distant languages is larger than that between similar languages.

6 Conclusion

The contribution of this paper is three-fold. First of all, we construct a novel CLTL model TALL based on a pre-trained multilingual language model. In the next place, we train TALL to conduct CLTL through an MT-oriented pre-training and an NLU-oriented fine-tuning. Last but not least, we implement UMT in the MT-oriented pre-training of TALL to make use of unannotated data. Compared with the baseline model, which is the pre-trained multilingual language model used to construct TALL, TALL consistently achieves better CLTL performance without using more annotated data, and the performance gain is relatively prominent in the case of distant languages. In the future, we will collect unannotated corpora that are linguistically compatible with the downstream NLU tasks for the UMT training, which we believe can further boost the CLTL performance of TALL.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Giusepppe Attardi. 2015. Wikiextractor.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

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

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*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Edward Loper and Steven Bird. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Phoebe Mulcaire, Jungo Kasai, and Noah A Smith. 2019. Polyglot contextual representations improve crosslingual transfer. 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)*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tjeerd van der Ploeg, Peter C Austin, and Ewout W Steyerberg. 2014. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC medical research methodology*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*.

Arnd Witte, Theo Harden, and Alessandra Ramos de Oliveira Harden. 2009. *Translation in Second Language Learning and Teaching*, volume 3. Peter Lang.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. 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)*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu. *arXiv preprint arXiv:2004.14353*.

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)*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.