FreeTransfer-X: Safe and Label-Free Cross-Lingual Transfer from Off-the-Shelf Models

Yinpeng Guo and **Liangyou Li** and **Xin Jiang** and **Qun Liu**

Huawei Noah's Ark Lab

{guo.yinpeng, liliangyou, jiang.xin, qun.liu}@huawei.com

Abstract

Cross-lingual transfer (CLT) is of various applications. However, labeled cross-lingual corpus is expensive or even inaccessible, especially in the fields where labels are private, such as diagnostic results of symptoms in medicine and user profiles in business. Although being lack of labels, there are off-the-shelf models in these sensitive fields. Instead of pursuing the original labels, a workaround for CLT is to transfer knowledge from the off-the-shelf models without labels. To this end, we define a novel CLT problem named *FreeTransfer-X* that aims to achieve knowledge transfer from the off-the-shelf models in rich-resource languages. To address the problem, we propose a 2-step knowledge distillation (KD, Hinton et al., 2015) framework based on multilingual pre-trained language models (mPLM)[1]. The significant improvement over strong neural machine translation (NMT) baselines demonstrates the effectiveness of the proposed method. In addition to reducing annotation cost and protecting private labels, the proposed method is compatible with different networks and easy to be deployed. Finally, a range of analyses indicate the great potential of the proposed method.

1 Introduction

Cross-lingual transfer (CLT) is a critical topic for natural language processing due to the data imbalance between languages. While models of rich-resource languages (e.g. English) have been applied on various real-world tasks, the progress on poor-resource languages lags behind. CLT researches enable the knowledge transfer from the rich-resource languages to the poor-resource languages.

Although the application of CLT is valuable, data labels are expensive or even inaccessible in private and sensitive domains, such as medicine and business. For example, the diagnostic results of a user's symptoms are private and a company's internal description of users are confidential. Since short of labels for CLT, even though there are excellent applications in rich-resource languages, it is difficult to benefit the people using poor-resource languages. Previous CLT researches have not well studied how to leverage knowledge of rich-resource languages without labels. To define and tackle this problem will benefit both the community and the industry.

In order to reduce the demand of labels, existing works mainly fall into two paradigms as follows. One paradigm focuses on learning language-agnostic representation and model parameters. CLT is realized by either aligning parameters of monolingual models or sharing parameters among different languages (Liu et al., 2019; Devlin et al., 2019b; Conneau et al., 2020; Wang et al., 2020). The objective is to build a unified representation, which is used by downstream tasks, for all the languages. In this paradigm, although the demand of labels is reduced, it still requires a certain number of labels to adapt the model to a particular language and task. Besides, models in this paradigm are usually large-scale Transformers (Vaswani et al., 2017) based on mPLMs, which limits their deployment in real-world. Another paradigm is to leverage machine translation (MT) systems to generate training or testing pseudo-corpus for a specific language (Conneau et al., 2018). For simplicity, we take English as the rich-resource languages in this paper. `Translate-train` translates annotated training corpus from English to other languages. Gold labels are directly applied to the translated data. Although labels in poor-resource languages are not required, gold labels in English are still necessary. On the contrary, `Translate-test` translates testing corpus from poor-resource languages to English. This method can directly lever-

---

[1] Source code are available at https://github.com/huawei-noah/noah-research/tree/master/NLP/FreeTransfer-X
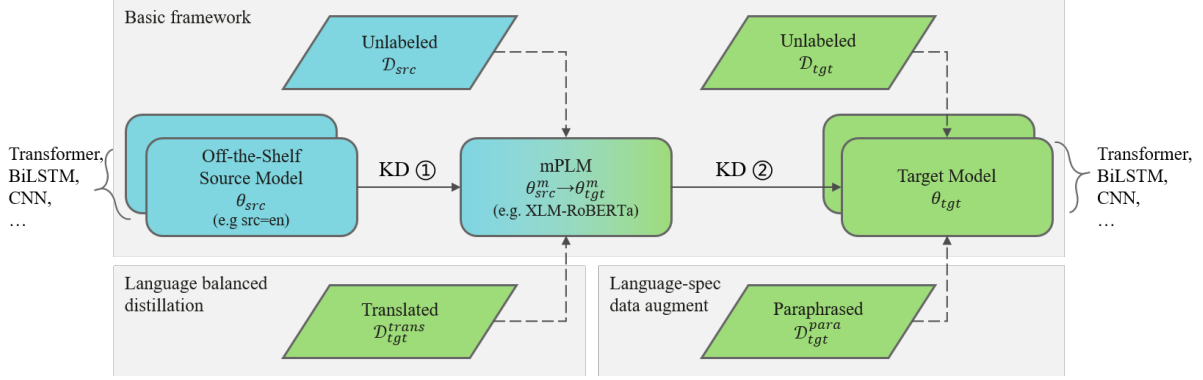
Figure 1: Overview of the proposed 2-step knowledge distillation (KD) framework. KD-(1) distills knowledge from the off-the-shelf English model to the mPLM. KD-(2) distills knowledge from the mPLM to the model in the target language. Blue modules: in the source language src, green modules: in the target language tgt.

age off-the-shelf English models, but it runs a 2-pass inference which highly limits its efficiency. Both the two CLT paradigms mentioned above require language-specific and task-specific labels, except for the 2-pass `Translate-test`. The demand of labels highly limits the reuse of the English knowledge in private and sensitive domains. Then a question comes up: *Is it possible to perform CLT totally without labels?*

In this paper, we define a novel problem: *safe and label-free cross-lingual transfer from off-the-shelf models (FreeTransfer-X)*. The FreeTransfer-X asks researchers to achieve CLT only with off-the-shelf English models but any labels, as formally defined in Section 2.1. To the best knowledge of the authors, it's the first time that the FreeTransfer-X is clearly defined.

To address the FreeTransfer-X, we propose a 2-step knowledge distillation (KD, Hinton et al., 2015) framework based on mPLM, as shown in Figure 1. Given an off-the-shelf model θ_{src} in the source language (e.g. English), first we take θ_{src} as the teacher and an mPLM model θ_{src}^m as the student, then train θ_{src}^m on unlabeled corpus $\mathcal{D}_{src}$. Second, we take θ_{tgt}^m as the teacher and train a student θ_{tgt} on unlabeled corpus $\mathcal{D}_{tgt}$. This cross-lingual transfer framework is label-free and applicable for any model architecture. Experimental results demonstrate the effectiveness of the proposed framework on both sentence classification and sequence tagging.

In short, the major contributions of this work include:

- A novel cross-lingual transfer problem FreeTransfer-X is defined. The FreeTransfer-X asks researchers to achieve CLT from off-

the-shelf models without using labels. It reduces the labeling cost and protects the labels in private domains such as medicine and business.

- We propose a 2-step knowledge distillation framework based on mPLMs, e.g. XLM-RoBERTa (Conneau et al., 2020), to address the FreeTransfer-X. It significantly outperforms the NMT baselines on sentence classification and sequence tagging tasks. Besides, it's compatible with heterogeneous networks.

- Further analysis indicates abundant research potentials of the proposed framework. To improve the two distillation steps and the mPLM may benefit the framework.

2 Methodology

2.1 Problem Definition

Denote the source language and the target language as src and tgt respectively. Given an off-the-shelf model θ_{src} (e.g. English intent classifier), unlabeled in-domain corpus $\mathcal{D}_{src}$ and unlabeled in-domain corpus $\mathcal{D}_{tgt}$, the objective is to output a model θ_{tgt} in the target language tgt. For simplicity in this paper, we constrain the target model θ_{tgt} to be of the same network architecture to the off-the-shelf source model θ_{src}.

2.2 Basic Framework

We propose to adopt knowledge distillation (KD, Hinton et al., 2015) to address the FreeTransfer-X, since it can transfer knowledge from teacher models without knowing original labels. In addition, knowledge distillation is free from network architectures and can be applied between heterogeneous

230

networks, which benefits the deployment in various environment.

2.2.1 Two-Step Knowledge Distillation

For a specific natural language processing (NLP) task, given a model θ_{src} and the unlabeled data $\mathcal{D}_{src}$ in the source language src and the unlabeled data $\mathcal{D}_{tgt}$ in the target language tgt. As shown in Figure 1, we propose to train a model θ_{tgt} in the target language tgt via 2 KD steps:

1. Leverage the NLP capability of the off-the-shelf model θ_{src}, e.g. an English sentence classifier $\theta_{en,cls}$. We distill knowledge from the teacher θ_{src} to the student mPLM θ_{src}^m on data $\mathcal{D}_{src}$.

2. Due to the zero-shot cross-lingual transfer capability of the mPLMs, θ_{src}^m implicitly achieve the NLP capability on the target language θ_{tgt}^m. Then similar to the step 1, we distill knowledge from the teacher θ_{tgt}^m to the student θ_{tgt} in the target language tgt on data $\mathcal{D}_{tgt}$.

The proposed framework works for arbitrary network including but not limited to Transformers (Vaswani et al., 2017), BiLSTM (Schuster and Paliwal, 1997) and CNN (Kim, 2014).

2.2.2 Training Objectives

The training is purely based on KD that no other training objectives is included. We only apply KD between the classification distribution $P_T(\cdot)$ and $P_S(\cdot)$ of the teacher and the student respectively, which is compatible to arbitrary model architecture. Freezing the parameters of the teacher, we train the student by minimizing the Kullback-Leibler Divergence (Div_{KL}, Joyce, 2011) between them. Denote the prediction category as $\mathcal{C} = [c_0, c_1, ..., c_k]$, then the Div_{KL} can be formalized as,

$$Div_{KL}(P_T(\mathcal{C}|\cdot)\|P_S(\mathcal{C}|\cdot))$$
$$= \sum_{c_i \in \mathcal{C}} P_T(c_i|\cdot) \log\left(\frac{P_T(c_i|\cdot)}{P_S(c_i|\cdot)}\right) \quad (1)$$

However, KD objectives of different NLU tasks varies a lot. We classify NLU tasks into two categories: 1) sentence-level tasks like sentence classification, 2) word-level tasks like sequence tagging. Given an input example $\mathcal{X} \in \mathcal{D}$ as a sequence of words $\mathcal{X} = [x_0, x_1, ..., x_n]$. For sentence-level tasks, $\mathcal{X}$'s sentence-level category is $\mathcal{C}_\mathcal{X}$. The teacher model and student model respectively output sentence-level prediction distribution

$P_T(\mathcal{C}_\mathcal{X}|\mathcal{X})$ and $P_S(\mathcal{C}_\mathcal{X}|\mathcal{X})$. For word-level tasks, $\mathcal{X}$'s word-level category is $\mathcal{C}_{x_i}, i \in [0, n]$. Then the KD objective can be written as,

$$\mathcal{L} = Div_{KL}(P_T(\mathcal{C}|\mathcal{X})\|P_S(\mathcal{C}|\mathcal{X}))$$
$$\text{where } \mathcal{C} = \begin{cases} \mathcal{C}_\mathcal{X} & \text{, sentence-level} \\ \mathcal{C}_{x_i} & \text{, word-level} \end{cases} \quad (2)$$

It's worth noting that word-level Div_{KL} cannot be directly applied for heterogeneous teacher and student models since their tokenizations are different. In order to align the predictions of teacher and student, we only adopt the prediction on the first sub-word of each word.

2.3 Enhanced Cross-Lingual Distillation

To explore the potentials of improving the two KD steps, we propose to enhance them with machine translation (MT) and paraphrase generation (PG).

2.3.1 Language Balanced Distillation

During the first KD step that training the mPLM from an English (i.e. the source language) classifier, to leverage the cross-lingual transferarability of mPLM, the conventional method is to train the mPLM only on the English corpus. However, in our preliminary experiments, we notice that the mPLM's accuracy gap between English and the target languages are very huge. It's over 5% between the English target model (94.0) and the average of all target models (88.4), as reported by *2-step KD* in Table 7, Appendix A.
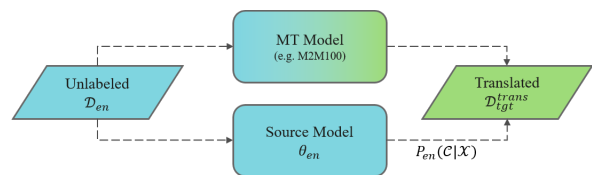

Figure 2: Language balanced distillation. Leverage the MT model to translate unlabeled English $\mathcal{D}_{en}$ into target languages $\mathcal{D}_{tgt}^{trans}$. Perform KD on the translated $\mathcal{D}_{tgt}^{trans}$ with θ_{en}'s predicted distribution $P_{en}(\mathcal{C}|\mathcal{X})$.

Hence, we propose to translate the unlabeled English corpus $\mathcal{D}_{en}$ to target languages $\mathcal{D}_{tgt}^{trans}$, as depicted by Figure 2. Since $\mathcal{D}_{en}$ and $\mathcal{D}_{tgt}^{trans}$ are aligned, source English model's predicted distribution $P_{en}(\mathcal{C}|\mathcal{X})$ of $\mathcal{D}_{en}$ can be directly applied to $\mathcal{D}_{tgt}^{trans}$. In this way, KD is able to be performed on not only the source language but also the target languages.

As shown in the lower left of Figure 1, the translated $\mathcal{D}_{tgt}^{trans}$ is incorporated in the training of KD step one.

2.3.2 Language-Specific Data Augmentation

Inspired by data augmentation for KD (Jiao et al., 2020) and multilingual paraphrase generation (Guo et al., 2019), we augment the unlabeled target corpus $\mathcal{D}_{tgt}$ via paraphrasing.
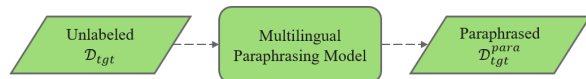

Figure 3: Language-specific data augmentation. We paraphrase the target corpus $\mathcal{D}_{tgt}$ into $\mathcal{D}_{tgt}^{para}$ as the augmented training data. KD is then performed on the mixture of $\mathcal{D}_{tgt}$ and $\mathcal{D}_{tgt}^{para}$.

3 Experiments

3.1 Datasets and Preprocessing

MultiATIS++ (Xu et al., 2020) extends the Multilingual ATIS corpus (Upadhyay et al., 2018) to 9 languages across 4 language families, including Indo-European (English, Spanish, German, French, Portuguese and Hindi), Sino-Tibetan (Chinese), Japonic (Japanese) and Altaic (Turkish). It provides annotations for intent recognition (sentence classification) and slot filling (sequence tagging) for each languages. The utterances are professionally translated from English and manually annotated. MultiATIS++ includes 37,084 training examples and 7,859 testing examples.

MTOP (Li et al., 2021) is a recently released multilingual NLU dataset covering 6 languages: English, German, French, Spanish, Hindi, Thai. It's also manually annotated for intent recognition (sentence classification) and slot filling (sequence tagging). MTOP provides a larger corpus consisting of 104,445 examples, of which 10% is validation set and 20% is testing set.

For each language, we randomly split both MultiATIS++ and MTOP into two balanced parts: *annotated* and *unannotated*. The annotated parts are used to train and simulate the off-the-shelf source models while the unannotated parts are used for training the baselines and the proposed 2-step distillation model. We tokenize Chinese, Japanese and Thai utterances using Jieba[2], MeCab[3] and

---

[2]https://github.com/fxsjy/jieba
[3]https://github.com/polm/fugashi

pythainlp[4] respectively.

3.2 Baselines

Translate-Test (Conneau et al., 2018) is a machine translation based method. It performs two-pass inferences to tackle the FreeTransfer-X problem: 1) translate the testing utterances into English (i.e. the source language) from the target language, 2) predict on the translated English utterances with the off-the-shelf English model.

Translate-Train-Pseudo is also based on machine translation. It's a variant of the **Translate-Train** (Conneau et al., 2018), which translates English training examples into target languages and applies English annotations to the translated examples. However, annotations are not provided in the FreeTransfer-X problem. Hence, Translate-Train-Pseudo utilizes the prediction of the off-the-shelf English model to pseudoly annotates the translated examples.

Gold-Supervised is for reference since it's trained with annotations. It replaces the first distillation step of the proposed framework with gold-supervised training, in other words, the mPLM is supervised by gold annotations instead of the off-the-shelf English model. It's supposed to be very strong.

3.3 Experiment Settings

3.3.1 Model Architectures

We experiment with three mainstream NLU model architectures to verify the universality of the proposed framework. They are used as the backbones of the off-the-shelf models θ_{src} and the output models θ_{tgt} in target language.

Transformer encoder (Vaswani et al., 2017) models input sequences fully with Attention mechanism. We follow the language modeling method of BERT (Devlin et al., 2019a). We adopt absolute positional encoding. The contextual representation vector of the first word is used for sentence classification. Sequence tagging is based on the contextual representation of each word.

Bidirectional LSTM (BiLSTM) (Schuster and Paliwal, 1997) models input sequences via leveraging two stacked LSTM layers respectively from backward and forward directions. We take the representation vector of the last word for sentence classification. Word-level representation is used for sequence tagging like Transformer.

---

[4]https://github.com/PyThaiNLP/pythainlp

Models		MTOP			MultiATIS++			Avg
		Transformer	BiLSTM	CNN	Transformer	BiLSTM	CNN	
Reference	Off-the-shelf En source	88.3	86.2	90.5	94.4	90.8	92.7	90.5
	Gold-supervised target	78.4	65.0	79.2	84.6	85.2	86.6	79.8
Baselines	Translate-test	69.6	66.0	73.8	86.4	80.7	86.2	77.1
	Translate-train-pseudo	64.2	57.9	67.4	84.7	81.2	83.2	73.1
Ours	2-step KD	75.1	72.3	75.6	87.7	83.8	85.0	79.9
	+ Balanced distillation	79.3	75.9	77.8	**88.9**	85.2	86.2	82.2
	+ Data augmentation	**79.6**	**79.1**	**78.8**	88.7	**86.4**	**86.9**	**83.3**

Table 1: Classification accuracy averaged over target languages. MTOP: de, es, fr, hi, th. MultiATIS++: de, es, fr, hi, ja, pt, tr, zh. For simplicity, the architecture of a target model is identical to its corresponding English source model.

Convolutional Neural Networks (CNN) (Kim, 2014) encodes input sequences with CNN modules. We adopt three kind of 1-D kernels with kernel size of 3, 4 and 5. Output vectors from all kernels and channels are concatenated as the representation for sentence classification. Dilated CNN (Strubell et al., 2017) is adopted for sequence tagging.

3.3.2 Training Details

English is regarded as the source language in all the experiments. Off-the-shelf English models are trained on the hold-out annotated English corpus as described in Section 3.1. All the experimented models are controlled in comparable model scale. AdamW (Loshchilov and Hutter, 2019) is adopted as the optimizer with $\epsilon = 1e - 8$. We train the models for 50 epochs and take the checkpoint of the best validation accuracy as the final model. Table 2 reports the hyper-parameters of the model architectures.

Model	Embed size	Hidden size	#Layers	#Params
Transformer	256	256	4	5.3M
BiLSTM	256	512	2	5.3M
CNN	256	768	2	5.0M

Table 2: Hyper-parameters of the experimented models.

Initial learning rate is decided based on a gradient-based searching heuristics proposed by Smith (2015), since in our preliminary experiments Smith (2015) stably finds better learning rates than manual searching. We build vocabularies of 10k words for each language via Byte Pair Encoding (BPE) Sennrich et al. (2016). Experiments are implemented with PyTorch (Paszke et al., 2019) and conducted on a single Nvidia V100 32GB GPU.

3.3.3 Auxiliary Models

M2M-100 (Fan et al., 2021) is adopted as the MT system in our experiments. We apply the 418M model checkpoint from Huggingface[5].

XLM-RoBERTa (Conneau et al., 2020) is adopted as the mPLM in the proposed 2-step distillation framework.

3.4 Results

Average accuracy across languages and models is given in Table 1 and Table 3. Language-wise results are provided in Appendix A.

3.4.1 Sentence Classification

As shown in Table 1, the proposed 2-step KD framework significantly outperforms the MT baselines on most model architectures, except for the CNN of *Translate-test*. Although *Translate-test* is strong in a very few cases, it requires 2-pass inference (MT and classification) that results in a high latency. On the contrary, the proposed framework directly produces classification models in the target languages, which is more efficient. In addition, the *language-balanced distillation* and *language-specific data augmentation* further enhance our model to a large extent, +2.3% and +1.1% respectively. Language-wise results in Table 7 demonstrate the robustness of our method across various languages.

To our surprise, the naive *2-step KD* model even performs on par with the *Gold-supervised* reference on average. We guess it's due to the regularization effects of knowledge distillation that brings a good generalizability to the proposed model. It implies the proposed framework may be a annotation-free alternative to current zero-shot cross-lingual transfer framework.

---

[5]https://huggingface.co/facebook/m2m100_418M

Models		MTOP			MultiATIS++			Avg
		Transformer	BiLSTM	CNN	Transformer	BiLSTM	CNN	
Reference	Off-the-shelf En source	74.8	81.1	72.1	88.4	94.0	89.1	83.3
	Gold-supervised target	64.6	68.6	63.3	71.5	76.5	74.1	69.8
Baselines	Translate-test	37.2	41.4	34.2	24.8	38.8	40.8	36.2
	Translate-train-pseudo	34.4	40.4	28.6	53.9	63.1	61.8	47.0
Ours	2-step KD	**63.7**	**67.6**	**55.7**	**71.7**	**76.9**	**73.5**	**68.2**

Table 3: Sequence tagging F1 score averaged over target languages. MTOP: de, es, fr, hi, th. MultiATIS++: de, es, fr, hi, ja, pt, tr, zh. For simplicity, the architecture of a target model is identical to its corresponding English source model.

Models		Original	Finetuned	Zero-Shot Cross-Lingual Transfer								
		en	en	de	es	fr	hi	**ja**	pt	**tr**	**zh**	Avg
Gold-supervised		-	97.9	97.6	97.4	97.4	92.4	90.6	97.3	83.8	92.8	93.7
Transformer	Naive KD	94.4	97.2	96.9	96.8	96.2	90.8	90.1	95.6	84.3	91.3	92.8
	+ Balanced distillation		97.5	97.6	96.9	96.6	95.4	96.3	96.1	90.8	97.5	95.9
BiLSTM	Naive KD	90.8	93.2	93.5	93.5	93.2	90.4	83.4	93.5	77.2	85.0	88.7
	+ Balanced distillation		92.4	93.3	93.7	92.4	91.8	91.9	93.1	86.6	92.8	92.0
CNN	Naive KD	92.7	94.7	91.8	94.1	93.2	90.3	90.9	94.2	83.5	90.3	91.0
	+ Balanced distillation		92.8	93.2	92.9	92.4	91.8	91.6	93.2	89.2	92.9	92.2

Table 4: Classification accuracy of the finetuned mPLM models, i.e. XLM-RoBERTa. Evaluated on MultiATIS++. *Gold-supervised* is trained with gold annotations. **Bold languages** is not in the Indo-European language family as English.

However, comparing the results of the English source model and those of the target models in Table 7, the cross-lingual transferred models still lag far behind the original English models. There is a great potential of the proposed framework.

3.4.2 Sequence Tagging

On the sequence tagging task, the proposed model beats the baselines by a wide margin. The MT-based baselines perform very poor on this task due to the error from word-level annotation alignment. Also because of the alignment error, we do not apply *language balanced distillation* and *language-specific data augmentation* on this task.

As to the comparison with the *Gold-supervised* reference, our model performs slightly worse than it. It may due to the insufficient knowledge distillation from the teacher to the student, which comes from the discrepancy between teacher's and student's tokenizations. Although, as described in Section 2.2.2, we perfectly align their prediction at word-level, only the first subword of each word is used for distillation. More informative subword-level aligning and distillation methods can be explored. We leave this problem for the future research. Besides, similar to sentence classification, gap between the English source model and the transferred target models is huge, as shown in Table 3.

In sum, both experimental results on sentence classification and sequence tagging demonstrate that the proposed model is significantly stronger than MT-based cross-lingual transfer methods. Furthermore, the proposed model only slightly lags behind or even performs on par with the strong *Gold-supervised* reference, which is not able to address the FreeTransfer-X problem.

4 Further Analysis

In order to explore the potential of the proposed framework, we analyze it in more details. For simplicity, experiments in this Section are conducted only on the MultiATIS++ sentence classification task.

4.1 Effects of the Distillation

Table 4 reports the accuracy of the XLM-RoBERTa finetuned from gold annotations, Transformer teacher, BiLSTM teacher and CNN teacher.

First, compare the *Original* with the *Naive KD Finetuned* of each model respectively. It's very interesting that the accuracy of the student mPLM is consistently higher than its teacher. The XLM-RoBERTa students gain 2.8%, 2.4% and 2.0% improvement from the *Original* teachers as Transformer, BiLSTM and CNN respectively. The phenomenon implies the general effectiveness of language modeling of mPLMs. We conjecture the improvement comes from two aspects: 1) mPLMs'

Models		Original en	Finetuned en	Zero-Shot Cross-Lingual Transfer								
				de	es	fr	hi	ja	pt	tr	zh	Avg
Transformer	XLM-RoBERTa	94.4	97.2	96.9	96.8	96.2	90.8	90.1	95.6	84.3	91.3	92.8
	mBERT		96.9	88.4	92.4	93.8	81.1	85.7	94.0	73.7	83.2	86.5
BiLSTM	XLM-RoBERTa	90.8	93.2	93.5	93.5	93.2	90.4	83.4	93.5	77.2	85.0	88.7
	mBERT		92.3	80.4	87.5	82.5	79.2	79.4	82.3	76.5	75.0	80.3
CNN	XLM-RoBERTa	92.7	94.7	91.8	94.1	93.2	90.3	90.9	94.2	83.5	90.3	91.0
	mBERT		93.3	82.6	86.9	87.9	78.1	78.7	88.2	72.9	80.9	82.0

Table 5: Classification accuracy of XLM-RoBERTa and mBERT. *Step-1 KD*: off-the-shelf English model -> mPLM. The mPLMs are finetuned and evaluated on MultiATIS++.

Models		Original en	Transferred en	Transferred Target Languages									
				de	es	fr	hi	ja	pt	tr	zh	Avg	Δ
Transformer	XLM-RoBERTa	94.4	94.0	94.4	93.3	90.8	85.7	81.3	92.4	76.5	87.1	87.7	-5.1
	mBERT		95.6	86.2	91.8	92.8	79.8	81.7	90.9	72.9	80.3	84.6	-1.9
BiLSTM	XLM-RoBERTa	90.8	89.8	89.8	89.7	89.5	84.1	78.1	84.9	71.5	83.1	83.8	-4.9
	mBERT		89.6	80.9	82.9	81.0	78.5	76.0	83.7	71.7	78.6	79.2	-1.1
CNN	XLM-RoBERTa	92.7	90.7	87.9	88.1	87.8	85.0	83.2	86.3	75.2	86.7	85.0	-6.0
	mBERT		89.1	79.4	82.9	81.5	76.0	78.2	82.4	72.0	79.5	79.0	-3.0

Table 6: Classification accuracy of the target models, distilled from XLM-RoBERTa and mBERT respectively. *Step-2 KD*: mPLM -> target model. Δ: changes w.r.t Table 5. The target models are transferred and evaluated on MultiATIS++.

generalizability learn from the large-scale pre-training, 2) the large model scale of mPLMs, which enhances its NLU capability. Besides, the improvement with respect to the *Original* varies across model architectures. Especially when compare Transformer (+2.8%) to CNN (+2.0%), although the Transformer's student XLM-RoBERTa performs much closer to the *Gold-supervised*, it still improves greater than the CNN's student. Since the XLM-RoBERTa is Transformer-based network, it implies that the knowledge distillation performs better if the architectures of the teacher and the student are more similar.

Second, under the cross-lingual transfer condition, although the *Gold-supervised* outperforms the *Naive KD* on most target languages, it performs weaker on Turkish (tr). It demonstrates the better generalizability and few-shot performance of the *Naive KD*, since Turkish is a low-resource language in MultiATIS++. The number of training examples of Turkish (578) is less than other languages (4488).

Third, the effectiveness of the proposed *language balanced distillation* is very clear. In the comparison between the *Naive KD* and + *Balanced distillation*, the accuracy is highly boosted almost on all the target languages. This improvement is particularly significant on the languages that is not in the same family of English: Hindi (hi), Japanese (ja), Turkish (tr) and Chinese (zh). A future research topic is to improve *language balanced dis-*

tillation on the languages similar to the source language, e.g. European languages to English. Data selection algorithms may have potentials.

In sum, the proposed framework and distillation method is effective and of strong generalizability. Future researches on heterogeneous distillation and data selection may benefit the proposed framework.

4.2 Effects of mPLM Models

Table 5 and Table 6 respectively reports accuracy of the step-1 KD and step-2 KD in the proposed framework. According to Table 6, the choice of mPLM is critical to the target models' performance. Performance with XLM-RoBERTa as the mPLM is stronger than with mBERT. However, there are interesting observations we should notice.

First, observe the performance changes (Δ) of the *Step-2 KD*: from the mPLM teacher to the target model student. We notice that the performance drop of mBERT is slighter than the XLM-RoBERTa's, based on the results of the average score in Table 6 minus those in Table 5. It implies that as the capability of mPLM increases, the KD dissipation tends to increase as well. Similar to the analysis in Section 4.1, the KD dissipation may come from: 1) the pre-trained language model that the target models lack of, 2) discrepancy between the model size of the mPLM and the target models. Hence, performance based on XLM-RoBERTa drops more due to its gap to the target models is greater than mBERT's in both the two aspects of

discrepancy. To reduce the KD dissipation, researches should focus on how to reduce the model discrepancy between mPLM and the target model, e.g. improve the language modeling capability of the target model. Besides, the performance difference among model architectures is consistent, either based on XLM-RoBERTa or mBERT. It further evidences that the proposed framework is general and works well for different model architectures.

4.3 Cross-Architecture Transfer

To analyze the proposed framework in a more general setting, we free the architecture ties of the off-the-shelf English models and the target models to be heterogeneous, that the source and target models can be different.
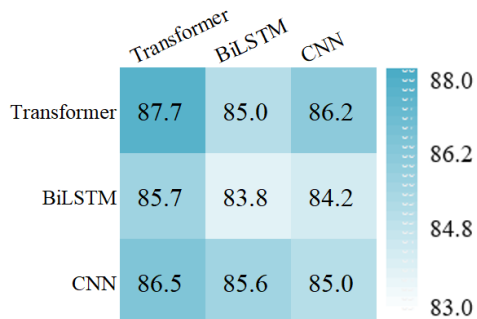

Figure 4: Classification accuracy of the target models via cross-architecture transfer, averaged over all target languages. Transfer from rows to columns. *Row*: architectures of source English models, *Column*: architectures of the target models. Experimented on Multi-ATIS++.

As depicted in Figure 4, the transfer performs the best when taking Transformer as both the source and target models. The worst comes to the transfer between BiLSTM models. On one side, the advantage of the Transformer architecture may be a reason. On the other side, it reconfirms the observation that the more similar teacher and student models are, the better transfer performance comes.

Besides, taking the BiLSTM as the source or target model consistently result in lower accuracy, no matter what the corresponding target or source models are. Hence, we guess the architecture similarity between BiLSTM and the Transformer-based mPLM is lower than that between CNN and the mPLM. We leave this for future work.

In addition, we study the accuracy drop from the source English models to the target models, as shown in Figure 5. From the perspective of the source model, the drop is the least when BiLSTM
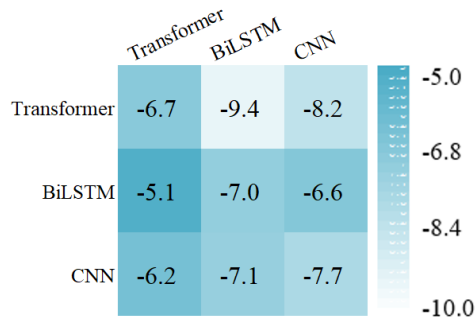

Figure 5: Accuracy drop from the source English models to the target models, averaged over all target languages. Transfer from rows to columns. *Row*: architectures of source English models, *Column*: architectures of the target models. Experimented on MultiATIS++.

is the source. From the perspective of the target model, the drop is the least when Transformer is the target. It reveals an asymmetry between the two KD steps with respect to the mPLM. To reduce the KD dissipation to the largest extent, it seems mPLM should be distilled from a weaker teacher architecture (e.g. BiLSTM) before teaching a stronger student architecture (e.g. Transformer).

In brief, the proposed framework works for heterogeneous cross-lingual transfer. The future work may focus on how to define the similarity between model architectures and how to evaluate the source-target model pairs.

5 Conclusions

In this paper, we define a novel cross-lingual transfer (CLT) problem - FreeTransfer-X, especially for CLT in private scenarios such as medical and business. The FreeTransfer-X is defined to transfer knowledge from off-the-shelf models in rich-resource languages to poor-resource languages, without labeled corpora. To address the FreeTransfer-X, we propose a 2-step knowledge distillation (2-step KD) framework based on multilingual pre-trained language models. In addition, two data augmentation methods for cross-lingual KD are proposed to boost the performance of the 2-step KD framework. Experimental results clearly demonstrate the effectiveness of the proposed framework. It's worth noting that the proposed KD framework can be applied between heterogeneous models, which benefits the deployment in different environment. Further analyses point out various research directions for future work.

References

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Multilingual bert.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Yinpeng Guo, Yi Liao, Xin Jiang, Qing Zhang, Yibo Zhang, and Qun Liu. 2019. Zero-shot paraphrase generation with multilingual language models. *CoRR*, abs/1911.03597.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

James M. Joyce. 2011. *Kullback-Leibler Divergence*, pages 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 33–43, Hong Kong, China. Association for Computational Linguistics.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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.

Leslie N. Smith. 2015. No more pesky learning rate guessing games. *CoRR*, abs/1506.01186.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, Copenhagen, Denmark. Association for Computational Linguistics.

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)*, pages 6034–6038.

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.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

A   Language-Wise Results

Here we list the detailed language-wise experimental results of Table 1 and Table 3 for reference.

Models			Source en	Targets									
				en	de	es	fr	hi	ja	pt	tr	zh	Avg
Reference		Gold-supervised	-	88.5	89.7	89.6	91.9	81.9	79.5	86.3	73.4	89.1	85.2
Transformer	Baselines	Translate-test	94.4	92.5	90.0	88.1	90.6	83.7	86.8	88.1	75.1	88.7	86.4
		Translate-train-pseudo		92.7	89.4	90.0	90.5	83.1	74.8	90.1	74.1	85.9	84.7
	Ours	2-step KD		94.0	94.4	93.3	90.8	85.7	81.3	92.4	76.5	87.1	87.7
		+ Balanced distillation		94.3	93.5	92.9	95.0	84.0	83.0	93.3	78.5	90.7	88.9
		+ Data augmentation		94.7	93.6	93.3	94.7	84.2	83.7	93.2	77.6	89.7	88.7
BiLSTM	Baselines	Translate-test	94.4	87.1	84.2	83.4	85.8	77.9	81.5	84.4	64.1	84.2	80.7
		Translate-train-pseudo		87.8	85.0	85.4	86.9	80.4	73.5	86.1	72.0	80.4	81.2
	Ours	2-step KD		89.8	89.8	89.7	89.5	84.1	78.1	84.9	71.5	83.1	83.8
		+ Balanced distillation		90.7	89.4	88.7	88.6	82.5	81.7	86.1	76.2	88.6	85.2
		+ Data augmentation		89.1	90.6	90.9	88.7	83.4	82.5	86.7	77.9	90.3	86.4
CNN	Baselines	Translate-test	94.4	90.7	86.6	86.3	88.7	86.9	85.3	88.2	81.0	86.3	86.2
		Translate-train-pseudo		86.3	84.4	85.6	86.5	83.8	79.8	82.2	77.5	85.9	83.2
	Ours	2-step KD		90.7	87.9	88.1	87.8	85.0	83.2	86.3	75.2	86.7	85.0
		+ Balanced distillation		89.1	88.5	87.8	88.8	85.7	83.0	86.2	79.2	90.6	86.2
		+ Data augmentation		89.6	89.8	89.2	89.0	86.0	83.8	87.1	79.4	90.9	86.9

Table 7: Sentence classification accuracy on MultiATIS++.