From Monolingual to Multilingual FAQ Assistant using Multilingual Co-training

Mayur Patidar, Surabhi Kumari, Manasi Patwardhan, Shirish Karande
Puneet Agarwal, Lovekesh Vig, Gautam Shroff
TCS Research, New Delhi, India
{patidar.mayur, surabhi.kumari6, manasi.patwardhan,
shirish.karande, puneet.a, lovekesh.vig,
gautam.shroff}@tcs.com

Abstract

Recent research on cross-lingual transfer show state-of-the-art results on benchmark datasets using pre-trained language representation models (PLRM) like BERT. These results are achieved with the traditional training approaches, such as Zero-shot with no data, Translate-train or Translate-test with machine translated data. In this work, we propose an approach of "Multilingual Co-training" (MCT) where we augment the expert annotated dataset in the source language (English) with the corresponding machine translations in the target languages (e.g. Arabic, Spanish) and fine-tune the PLRM jointly. We observe that the proposed approach provides consistent gains in the performance of BERT for multiple benchmark datasets (e.g. 1.0% gain on MLDocs, and 1.2% gain on XNLI over translate-train with BERT), while requiring a single model for multiple languages. We further consider a FAQ dataset where the available English test dataset is translated by experts into Arabic and Spanish. On such a dataset, we observe an average gain of 4.9% over all other cross-lingual transfer protocols with BERT. We further observe that domain-specific joint pre-training of the PLRM using HR policy documents in English along with the machine translations in the target languages, followed by the joint finetuning, provides a further improvement of 2.8% in average accuracy.

1 Introduction

Achievement of scale, agility, and quality in support functions of large enterprises is a key demand. Conversational systems are increasingly being deployed to this effect. Such systems try to classify users' utterances into one of the FAQ (Khurana et al., 2017), usually referred to as intent, and then show an answer that is mapped to the chosen intent. In specific geographies such as Europe, Latin America, and India such FAQ based conversational systems may be required to work in more than one language. Similar requirements are also presented to us by many international consumer oriented businesses such as airlines, shipping companies, and banks.

The straight-forward approach is to build a different classification model for every language, which is hard to maintain because of manual effort involved in preparing the training data in every language, and training time for every model. We therefore look into cross-lingual transfer learning approaches such as a) *Translate-Train* (Schuster et al., 2019): here we translate the training data from English[1] into all the other languages and train a different model for every language; b) *Translate-Test* (Artetxe and Schwenk, 2018): here we maintain single model (usually for English), and use machine translation at the inference time before using the classification model; c) *Zero-shot* (Artetxe and Schwenk, 2018): here we employ multi-lingual pre-trained language representation model (PLRM) such as LASER (Artetxe and Schwenk, 2018), and train the model in high resource language (English) only and use the target language at the inference time only; d) *Joint training* (Upadhyay et al., 2018a,b): here the same model is trained on all the languages on which it is expected to be used. All these approaches are also shown in Figure 1. Either the accuracy of above mentioned models is low (Zero-Shot, or Translate-Test) or they are too hard to maintain in production system (Translate-Train, or Joint training). We therefore require an approach that performs better than all these approaches and is easier to maintain.

In this paper, we propose a new method for cross-lingual transfer learning, i.e., Multi-Lingual

[1]Most often English is the most common language in all deployments of FAQ systems

Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo), pages 115–123
Hong Kong, China, November 3, 2019.
https://doi.org/10.18653/v1/P17

Co-Training (MCT). Here, we jointly train single model on all the languages (upto 15 languages), using different multi-lingual PLRMs. When the training data is not available for certain language, we use translate-train paradigm and use machine translations as the training data. To the best of our knowledge, such an approach has not been used by prior works in the related area. We demonstrate the efficacy of our approach on a real world dataset taken from "Watt" (Khurana et al., 2017) project. Finally, we also demonstrate the robustness on publicly available datasets such as XNLI (Conneau et al., 2018) and MLDoc (Schwenk and Li, 2018). For MLDoc dataset, MCT provides 1.0% gain for the 8 languages, whereas for the XNLI dataset it provides 1.2% gain for 15 languages.

The rest of the paper is organized as follows: We describe our problem in Section 2 and the proposed approach in Section 3. We present the results of the proposed and other baseline approaches in Section 4. We later describe related work in Section 5 and conclude in Section 6.

2 Problem Description

A labeled dataset (D) for the deployed FAQ assistant in the source language (i.e., English) was created by HR domain experts using policy documents. It consists of a set of intents i.e. $D = \{I_1, I_2, ..., I_n\}$ where, each I_j comprises of a set of semantically similar queries $Q_j = \{q_{j1}, q_{j2}, ..., q_{jm}\}$ and a common corresponding answer ans_j i.e. $I_j =< Q_j, ans_j >$. Our objective here is to find a relevant intent I corresponding to a user's query q and then retrieve and show the answer associated with that intent. This can be modeled as a multiclass sentence classification where $I = \underset{I_j \epsilon D}{\operatorname{argmax}} P(I_j/q)$.

In the context of a multilingual FAQ assistant, we assume that there exists complete overlap between the intents of source and target languages (T_i, $i = 1, 2, ..., k$) with no availability of human labeled data in any target language. The objective in the case of multilingual FAQ assistant is similar to the monolingual case except that user is free to ask a query in any language. In a multilingual FAQ system, along with intent identification, a language detection module is also required to respond to a user's query in an appropriate language.

3 Proposed Approach

In the context of a multilingual FAQ assistant, we assume that there exists complete overlap between the intents of source and target languages (T_i, $i = 1, 2, ..., k$) with no availability of human-labeled data in any target language. To create a labeled dataset D_{T_i} for a target language T_i, each set of semantically similar queries Q_j in the source language are translated to the target language to obtain Q_{jT_i}, using machine translation (MT)[2] and ans_{jT_i} is created by the respective HR domain experts.

To obtain a single multilingual labeled dataset $D^{'}$ comprising of data from the source as well as all the target languages, we combine D with all the datasets D_{T_i} created for all T_i. Each intent $I_j =< Q^{'}_j, ans^{'}_j >$ in the final labeled dataset ($D^{'}$) is comprised of queries $Q^{'}_j = \{Q_j \cup Q_{jT_i} \cup ... \cup Q_{jT_k}\}$ and answers $ans^{'}_j = \{ans_j \cup ans_{jT_i} \cup ... \cup ans_{jT_k}\}$ from the source and target languages.

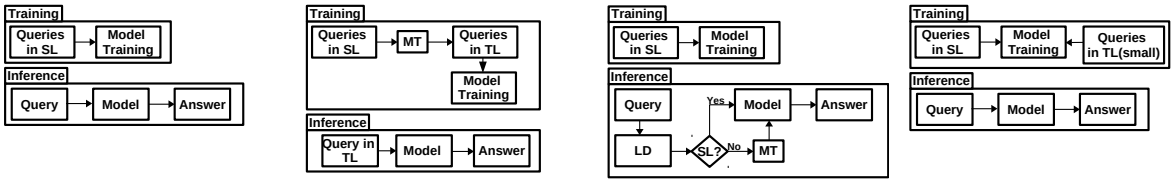
We propose an approach referred to as Multilingual Co-training (MCT), where we use multilingual labeled dataset $D^{'}$ to train a multiclass classifier for intent identification. In this work, we propose three variants of MCT, which differ in terms of how we train a classifier given multilingual labeled dataset $D^{'}$, which we discuss in next subsections.

In all variants of MCT, we need a translation system only to create the dataset $D^{'}$. Unlike translate-test, we do not require to translate each user query to source language during the inference. Also, we need to maintain only a single multilingual FAQ assistant for all languages. However, in case of translate-train, in general, we need to create multiple FAQ assistants, one for each language. We use $D^{'}$ to train a multilingual FAQ system, which may not be the best but perform better than solely relying on representations from PLRM (zero-shot) for cross-lingual transfer.

3.1 MCT using Multilingual Sentence Representation (MCT-MSR)

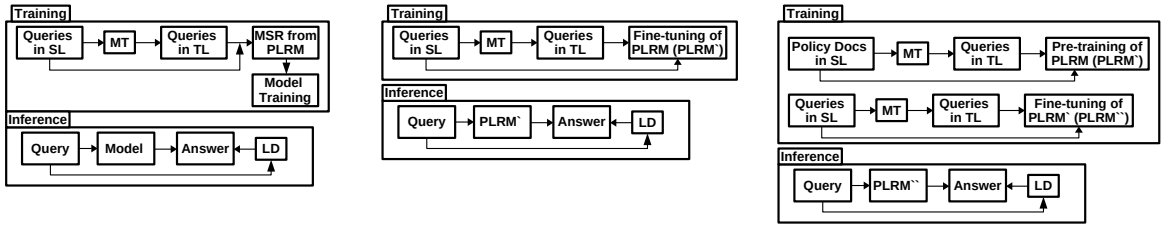
MCT-MSR is the simplest variant of MCT, where we obtain vector representation for all the queries present in dataset $D^{'}$ from the PLRM. Corresponding to each user query q_t, we obtain vector representation $\boldsymbol{q_t} \in R^d$ where d is the dimension

[2] we use google translation api for machine translation.



(a) Zero-Shot (Z-shot) (b) Translate-train (T-train) (c) Translate-test (T-test) (d) Joint Training

Figure 1: Baseline approaches for cross-lingual transfer



(a) MCT-MSR (b) MCT-FT (c) MCT-PT-FT

Figure 2: Proposed approaches for MCT

of query representation. We use these query representations to train a multiclass classifier by minimizing the categorical cross entropy loss as shown in Equation 3, where $I_i \in I$, N is the total number of queries in D', n is the number of intents in D' and y is 1 only for the target intent and zero otherwise. We build the classifier using a two layered feed forward network as described in Equations 1 and 2 where W_1, W_2 represent the weights and b_1, b_2 represent the biases of the two layers. We also use dropout (Srivastava et al., 2014) for regularization and *tanh* as the nonlinear activation function.

Finally, as shown in Figure 2a, we use trained classifier with language detection module i.e., Multilingual FAQ assistant to answer user's query in source or target language.

$$\boldsymbol{o_t} = dropout(tanh(W_1 * \boldsymbol{q_t} + b_1)) \quad (1)$$

$$p(I \mid \boldsymbol{q_t}) = \text{softmax}(W_2 * \boldsymbol{o_t} + b_2) \quad (2)$$

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{n} y \cdot log(p(I_i \mid \boldsymbol{q_t})) \quad (3)$$

3.2 MCT using Fine Tuning (MCT-FT)

In the recent work (Devlin et al., 2019; Lample and Conneau, 2019; Wu and Dredze, 2019), it is shown that fine-tuning of all or a few layers of PLRM on end task performs better than task-specific models. Unlike in MCT-MSR, in MCT-FT we use D' to fine-tune all the parameters of PLRM along with the weights W_3 and biases b_3 of a task-specific linear layer as shown in Equation 4 which is similar to (Devlin et al., 2019). In Equation 4, $\boldsymbol{q_t}$ refers to vector representation of user's query i.e., $\boldsymbol{q_t}$ obtained from PLRM.

$$p(I \mid \boldsymbol{q_t}) = \text{softmax}(W_3 * \boldsymbol{q_t} + b_3) \quad (4)$$

Finally, as shown in Figure 2b, we use PLRM, obtained after fine-tuning with language detection module i.e., Multilingual FAQ assistant to answer user's query in source or target language.

3.3 MCT via pre-training followed by fine-tuning (MCT-PT-FT)

In (Devlin et al., 2019), it is shown that additional pre-training of PLRMs on domain-specific text corpus improve the performance on the end task. In this work, addition to D', we also use policy documents to create our multilingual FAQ assistant. A policy document is a semi-structured document which contains information (e.g., purpose, applicability, approval workflow, etc.) about leave type in the form of tables, plain text, etc. In this work, we only use plain text from policy documents. Due to unavailability of policy documents in the target languages, we use MT to translate them to target languages.

In MCT-PT-FT, we perform domain-specific pre-training of existing PLRM using policy documents on tasks specific to PLRM. For example, we pre-train BERT (Devlin et al., 2019) using MLM and NSP tasks. Similar to MCT-FT, we fine-tune the PLRM obtained after pre-training on policy documents.

Finally, as shown in Figure 2c, we use a PLRM, obtained after domain-specific pre-training and fine-tuning with language detection (LD) module forming the Multilingual FAQ assistant to answer user's query in source or target language.

Table 1: We compare the existing baselines with **MCT-MSR** using **LASER**, **BERT** and **XLM** as PLRMs **without fine-tuning** on "Leave Dataset". Bold with * denotes the best and underline denotes the second best average classification accuracy on test set.

Language Code	LASER (Artetxe and Schwenk, 2018)				XLM (MLM+TLM) (Lample and Conneau, 2019)				BERT (Devlin et al., 2019)			
	Z-shot	T-train	T-test	MCT-MSR (Ours)	Z-shot	T-train	T-test	MCT-MSR (Ours)	Z-shot	T-train	T-test	MCT-MSR (Ours)
en	85.5	85.5	85.5	82.4	47.1	47.1	47.1	44.2	59.5	59.5	59.5	53.9
ar	45.6	52.4	48.7	58.1	4.7	13.1	23.1	17.4	3.9	18.5	22.8	19.1
es	63.6	73.3	63.9	72.3	17.5	29.9	28.0	29.8	6.6	35.1	28.7	31.5
Average	64.9	70.4	66.0	**70.9***	23.1	30.0	**32.7***	30.5	23.3	**37.7***	37.0	34.8

Table 2: We compare the existing baselines with **MCT-FT** using **Watt**, **BERT** and **XLM** as PLRMs **with fine-tuning** on "Leave Dataset". Bold with * denotes the best and underline denotes the second best average classification accuracy on test set.

Language Code	Watt (BiLSTM + SQRT-KLD) (Khurana et al., 2017)				XLM (MLM+TLM) (Lample and Conneau, 2019)				BERT (Devlin et al., 2019)			
	Z-shot	T-train	T-test	MCT-FT (Ours)	Z-shot	T-train	T-test	MCT-FT (Ours)	Z-shot	T-train	T-test	MCT-FT (Ours)
en	83.5	83.5	83.5	79.4	82.3	82.3	82.3	85.6	90.0	90.0	90.0	89.8
ar	-	18.1	24.1	31.5	15.0	39.6	41.5	52.4	9.4	44.9	44.4	56.5
es	-	56.3	34.0	63.9	28.0	62.7	50.1	72.2	26.5	78.5	58.5	81.8
Average	-	52.6	47.2	**58.3***	41.8	61.5	58.0	**69.9***	42.0	71.1	64.3	**76.0***

Table 3: Dataset description. SPL refers to Samples Per Language

Property ↓ / Dataset →	Leave	MLDoc	XNLI
Train-SPL	2801	1000	392,702
Validation-SPL	934	1000	2490
Test-SPL	832	4000	5010
No. of classes	199	4	3
No. of languages	3	8	15

4 Experiments and Results

In this section, we describe the various datasets and also give details of the different hyper-parameters of the models used in our experiments. We later present all the results and note some key observations from them.

4.1 Dataset Description

We evaluate proposed approaches on three datasets as shown in Table 3.

Leave dataset (Khurana et al., 2017) is created by HR domain experts in English and for our purpose, we translate training and validation set in target languages (Arabic and Spanish) using MT, while test set is translated by respective target language experts.

MLDoc[3] (Schwenk and Li, 2018) is a four class, multilingual document classification dataset containing news stories in eight languages, where stories in target languages are written by respective target language experts. Similar to (Wu and Dredze, 2019), we take first two sentences from each document in our experiments and use NLTK[4] for sentence tokenization.

The Cross-lingual Natural Language Inference (XNLI) (Conneau et al., 2018) dataset is an extension of Multi-Genre Natural Language Inference (MultiNLI)[5] corpus, where objective is to classify a pair of sentences (premise and hypothesis) in one of the three classes. Validation and test set are translated by domain experts and the training set by a machine translation system in 14 target languages.

4.2 Training Details

For training Watt, the final hyper-parameters are selected from the sets as mentioned in (Khurana et al., 2017). The datasets mentioned in the Table 3 are not pre-processed in any form during our experiments. All the final hyper-parameters are selected based on the performance on a validation set. We use Adam (Kingma and Ba, 2014) for optimization and dropout for regularization (Srivastava et al., 2014). The batch size is selected from the set $\{16, 32\}$.

We use multilingual variants of PLRMs, viz. BERT[6], XLM(MLM+TLM)[7] (Lample and Conneau, 2019) and LASER[8] (Artetxe and Schwenk, 2018) in our experiments.

MCT-MSR The number of hidden units and layers are selected from the sets {512, 1024,

[3] https://trec.nist.gov/data/reuters/reuters.html

[4] https://www.nltk.org/

[5] https://www.nyu.edu/projects/bowman/multinli/

[6] https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

[7] https://dl.fbaipublicfiles.com/XLM/mlm_tlm_xnli15_1024.pth

[8] https://github.com/facebookresearch/LASER

Table 4: We compare the results of **Fine-tuning** vs **Pre-training followed by Fine-tuning** of various models on "Leave Dataset". For MCT we use BERT as a PLRM. Bold with * denotes the best and underline denotes the second best average classification accuracy on test set.

Language Code	Fine-tuning				Pre-training followed by Fine-tuning			
	Z-shot	T-train	T-test	MCT-FT (Ours)	Z-shot	T-train	T-test	MCT-PT-FT (Ours)
en	90.0	90.0	90.0	89.8	90.3	90.3	90.3	90.5
ar	9.4	44.9	44.4	56.5	10.9	44.9	44.8	59.4
es	26.5	78.5	58.5	81.8	30.8	79.2	60.0	86.4
Average	42.0	<u>71.1</u>	64.3	**76.0***	44.0	<u>71.5</u>	65.0	**78.8***

Table 5: We compare the results of different approaches to bilingual co-training (BCT) on "Leave Dataset". We use LASER (Artetxe and Schwenk, 2018), BERT (Devlin et al., 2019) and XLM (MLM+TLM) (Lample and Conneau, 2019) as PLRMs. Bold with * denotes the best and underline denotes the second best average classification accuracy on test set for each approach. Bold with ** denotes the overall best. The first three rows showcase results of the billingual models created on en-ar pair and later 3 rows on en-es pair.

Language Code	BCT-MSR			BCT-FT			BCT-PT-FT
	LASER	XLM (MLM+TLM)	BERT	Watt	XLM (MLM+TLM)	BERT	BERT
en	81.9	39.5	53.3	75.0	82.4	89.5	90.0
ar	55.1	16.3	18.1	20.7	48.3	59.8	63.5
Average	**68.5***	27.9	<u>35.7</u>	47.8	<u>65.3</u>	**74.6***	**76.7****
en	84.7	48.3	56.6	78.1	85.1	90.5	90.6
es	73.9	34.7	36.0	61.6	72.5	82.7	84.0
Average	**79.3***	41.5	<u>46.3</u>	69.8	<u>78.8</u>	**86.6***	**87.3****

2048} and {1, 2} respectively with *tanh* as the non-linearity. The learning rate and dropout are selected from the sets $\{1e-2, 1e-3\}$ and {0.1, 0.2, 0.3} respectively.

MCT-FT For fine-tuning of PLRMs based on the end task, we use dropout of 0.1 and the learning rate is selected from the set {2e-5, 3e-5, 5e-5}. For MLDoc dataset, we also use L2 weight decay of 0.01 in addition to dropout for regularization. For XNLI (Conneau et al., 2018) and MLDoc (Schwenk and Li, 2018) datasets, the number of epochs for fine-tuning are selected from the set {3,4}. However for Leave dataset due to high number of classes and small data size, we have used early stopping.

MCT-PT-FT To utilize domain-specific corpus i.e., policy documents, we have run additional steps of pre-training starting from the existing Multilingual BERT model. We have used a masking probability of 0.15, learning rate of $2e{-}5$, 50% noise for data creation for NSP, batch size of 32 and a maximum of 20 masked LM predictions per sequence. The number of epochs for pre-training of BERT are selected from the set {5, 10, 15}.

4.3 Results And Discussion

In this work, we compare proposed approaches with existing baselines. For fair comparison, we compare proposed approaches with existing baselines under different scenarios, i.e. use of PLRMs with/without finetuning and/or pre-training. In all our experiments we assume that the accuracy of the language detection (LD) module is 100%. This is not an unreasonable assumption, as IP address, employee number, scripts and vocabulary can all be used together for language detection.

MCT-MSR vs Baselines In first scenario (without fine-tuning of PLRMs), we obtain multilingual sentence representations (MSRs) for each sentence in a given dataset and train a classifier as described in subsection 3.1. According to Table 1, on Leave dataset, for LASER, MCT-MSR perform slightly better than other baseline approaches. However, in case of BERT and XLM, baseline approaches perform better than MCT-MSR. Overall, LASER-based approaches perform better than BERT and XLM since, pre-training objective of LASER, "machine translation using single encoder for 93 languages", seems to explicitly force alignment of sentence representations in

Table 6: We compare **MCT-FT** with the existing baselines on "MLDoc" (Schwenk and Li, 2018) Dataset. We use BERT (Devlin et al., 2019; Wu and Dredze, 2019) as a PLRM for **MCT-FT**. Bold with * denotes the best classification accuracy on test set for each language and also for the average across all languages.

Language Code	Z-shot			T-train		MCT-FT(Ours)
	MLDoc	LASER	BERT	MLDoc	BERT	BERT
en	92.2	89.9	94.2	92.2	94.2	**94.3***
de	81.2	84.8	80.2	93.7	93.3	**96.6***
zh	74.7	71.9	76.9	87.3	89.3	**91.7***
es	72.5	77.3	72.6	94.5	95.7	**96.0***
fr	72.4	78.0	72.6	92.1	93.4	**94.2***
it	69.4	69.4	68.9	85.6	**88.0***	87.7
ja	67.6	60.3	56.5	85.4	88.4	**89.6***
ru	60.8	67.8	73.7	85.7	87.5	**87.7***
Average	73.9	74.9	74.5	89.5	91.2	**92.2***

Table 7: We compare **MCT-FT (Ours)** with the existing baselines on "XNLI Dataset" (Conneau et al., 2018) Dataset. We use BERT and XLM (MLM+TLM) as PLRMs for **MCT-FT (Ours)**. Bold with * denotes the best classification accuracy on test set for each language and also for the average across all languages.

Language Code	XLM (MLM+TLM) (Lample and Conneau, 2019)				BERT (Devlin et al., 2019; Wu and Dredze, 2019)		
	Z-shot	T-train	T-test	MCT-FT (Ours)	Z-shot	T-train	MCT-FT (Ours)
en	**85.0***	**85.0***	**85.0***	83.5	**82.1***	**82.1***	80.6
fr	78.7	**80.2***	79.0	79.3	73.8	76.9	**77.4***
es	78.9	**80.8***	79.5	80.2	74.3	**78.5***	78.2
de	77.8	**80.3***	78.1	78.7	71.1	74.8	**76.3***
el	76.6	**78.1***	77.8	78.0	66.4	72.1	**74.3***
bg	77.4	**79.3***	77.6	77.8	68.9	**75.4***	75.1
ru	75.3	**78.1***	75.5	75.6	69.0	**74.3***	73.6
tr	72.5	**74.7***	73.7	72.8	61.1	70.6	**71.2***
ar	73.1	**76.5***	73.7	75.0	64.9	**70.8***	70.5
vi	76.1	76.6	70.8	**77.1***	69.5	67.8	**75.3***
th	73.2	75.5	70.4	**76.4***	55.8	63.2	**65.7***
zh	76.5	**78.6***	73.6	78.5	69.3	**76.2***	75.9
hi	69.6	**72.3***	69.0	71.9	60.0	65.3	**67.2***
sw	68.4	**70.9***	64.7	70.4	50.4	65.3	**66.3***
ur	67.3	63.2	**65.1***	63.8	58.0	60.6	**64.53***
Average	75.1	**76.7***	74.2	75.9	66.3	71.6	**72.8***

multiple languages.

MCT-FT vs Baselines In second scenario (fine-tuning of PLRMs), we fine-tune PLRMs as described in subsection 3.2. However, Watt is not based on PLRMs and for comparison we train it from scratch as described in (Khurana et al., 2017). As LASER (Artetxe and Schwenk, 2018) is typically used to obtain MSRs, we have not considered it for comparison here. According to Table 2, for Leave dataset, proposed approach MCT-FT performs significantly better than baseline approaches in all cases and for BERT we gain 4.9% in terms of average classification accuracy compared to translate-train. For MLDoc dataset we achieve better accuracy in seven out of eight languages with 1.0% average improvement over existing baselines as shown in Table 6. According to Table 7, for XLM, baseline translate-train performs better than the proposed approach by 0.8%. However, in case of BERT we achieve better accuracy in nine out of fifteen languages with an improvement of 1.2% in terms of average classification accuracy compared to translate-train.

MCT-PT-FT vs Baselines In third scenario (pre-training followed by fine-tuning of PLRMs), we pre-train PLRM using domain-specific unlabeled text corpus (policy documents) and fine-tune it on labeled dataset as discussed in subsection 3.3. Since BERT outperformed XLM during fine-tuning we use BERT as a PLRM for all

baselines and as well as the proposed approach MCT-PT-FT. According to Table 4, MCT-PT-FT outperforms translate-train by a margin of 7.3% and gains an improvement of 2.8% over MCT-FT. MCT-PT-FT was tested on Leave dataset only as for other datasets their domain-specific unlabeled text corpora were unavailable.

Bilingual Co-training (BCT) For completeness, we also report results on bilingual co-training which is type of MCT, where unlike bilingual joint-training we use machine translated data for target language. According to Table 5, BERT based MCT-PT-FT performs better for both language pairs i.e., en-es and en-ar as compared to MCT-MSR and MCT-FT.

Does noisy translation affect MCT ?

It is interesting to note, from Table 4, the gains obtained by MCT-PT-FT over Translate-train on Spanish (*es*) (86.4% over 79.2%) and Arabic (*ar*) (59.4% over 44.9%). Apart from these gains in performance, the poorer performance on *ar* compared to *es* can be attributed to the noise induced by MT when translating the domain-specific words from English to target languages. To verify this, we translate the test set of English into *es* and *ar* (one set for each) using MT. We then evaluate the performance of MT in terms of BLEU (Papineni et al., 2002) score by considering the manually labeled test sets of *es* and *ar* as the reference translations. These are found to be 40.0 for *ar* and 63.0 for *es*, further validating our observation regarding the noisy MT system. In future, one can consider approaches which compensate for the translator noise. For example, during MCT one could use different weights for each language in the cost function.

5 Related Work

In this section, we provide an outline of existing FAQ assistants, followed by an overview of the recent work on multilingual language modelling and cross-lingual transfer methods.

5.1 FAQ Assistants

Recent years have seen significant advances in conversational systems, with various models considering context, affect, goal, external knowledge etc. However, all these systems can be categorized into two types i.e. those which seek to generate responses or those which use a retrieval based approach. (Zhou et al., 2018; Pei and Li, 2018) are examples where the ability to generate responses is learnt from patterns in dialogues found in the training set. On the contrary, there exist several industrial scenarios where the domain is sufficiently restricted, or there exist legal ramifications associated with the responses, and hence pre-defined answers are preferred. Therefore, research on retrieval based conversational models continues to be active, for example see (Das et al., 2016; Lai et al., 2015). Our work builds upon the retrieval-based model for a domain-specific leave dataset used in (Khurana et al., 2017), where a Bi-LSTM based architecture was employed.

Multilingual and cross-lingual conversational models for virtual assistants are an emerging field of research. Some research work has been done to capture different languages in one conversational system. In (Gupta et al., 2018), machine translation and information retrieval approaches were used for multilingual question answering in English and Hindi languages. In (Schuster et al., 2019), the authors use different cross-lingual embeddings eg. XLU (Schuster et al., 2019), ELMo (Peters et al., 2018), CoVe (McCann et al., 2017), etc. for cross-lingual learning in English, Spanish and Thai. In this paper, we propose an approach to extend an FAQ system to other languages such as Arabic and Spanish.

5.2 Multilingual Sentence Representation

There are approaches which have specifically been developed for capturing cross-lingual sentence representations. An encoder was used to align a parallel set of sentences to learn joint space embeddings in (Hermann and Blunsom, 2014; Conneau et al., 2018), an encoder pre-trained on the translation task with multiple source languages was utilized in (Artetxe and Schwenk, 2018; Eriguchi et al., 2018; Yu et al., 2018; Schwenk et al., 2017), Transformer based approaches such as BERT further extended to the multilingual setting (Wu and Dredze, 2019) and XLM (Lample and Conneau, 2019) having a cross-lingual objective for language modeling can be used to obtain multilingual sentence representations for cross-lingual transfer. The cross-lingual sentence representation obtained from these models can be further utilized for multilingual downstream tasks, e.g. (Schwenk and Li, 2018; Conneau et al., 2018). In our work, as we are trying to extend our FAQ assistant to the multilingual setting, we use the

BERT, XLM models (Devlin et al., 2019; Lample and Conneau, 2019; Artetxe and Schwenk, 2018) as base models and further fine-tune them with domain-specific and task-specific data.

5.3 Cross-Lingual Transfer

For low resource languages, due to insufficient (or no) data availability, it is difficult to get good task-specific accuracies. In case of complete unavailability of low resource language data, various approaches are defined in the literature: (i) zero-shot approaches, which train task-specific models on high resource languages and then use these models directly for low resource languages (Artetxe and Schwenk, 2018). (ii) Using predefined word or sentence embeddings (Schwenk and Li, 2018). (iii) Making use of translated high resource language data for training a low resource language model (Schuster et al., 2019). For cases where a small amount of low resource language data is available, there are approaches which make use of joint training using high resource language data augmented with a small amount of target (low resource) language data, which leads to better task-specific accuracy for target languages than zero shot (Upadhyay et al., 2018a,b). These approaches are applicable for the bi-lingual as well as multilingual settings. There are studies which help determine the applicability of using a particular high resource source language for a (set-of) low resource target language(s) (Lin et al., 2019). Our work is inspired by the joint-training approach of cross-lingual transfer, however, we assume unavailability of target language data and use machine translations for the same.

6 Conclusions and Future Work

There are a few baseline observations that need to be highlighted before commenting upon the key conclusions about the proposed "Multi-lingual Co-training". With regards to our Multilingual FAQ bot, when compared with Watt (Khurana et al., 2017), it seems that the use of PLRMs can improve the performance even for English. Therefore, it was reasonable to base the study in this work on the three recently proposed PLRMs, viz. LASER, BERT and XLM. With regards to cross-lingual transfer, if one were to use the PLRMs purely as feature extractors, then LASER provide the best baselines. Meanwhile, if one were to allow fine-tuning, then BERT provides the best baselines. In both cases the best baseline is provided by Translate-Train. The proposed variants MCT-MSR of LASER and MCT-FT of BERT are able to beat the corresponding baselines. In fact, one can observe that while Watt and XLM do not provide the best baselines for fine-tuning, even for these models, multilingual co-training does help. Finally, we explored the use of pre-training in the multilingual setting. While human translations have been used by LASER, the use of machine translations as a self-supervised language modeling task has not been explored in the past. Translation noise can potentially lead to a lot of error propagation. However, we observed that use of translations for pre-training provides the best baseline with BERT, and a joint multilingual pre-training is able to beat this baseline.

As a part of the future work, we would like to explore distinct strategies for a further boost in the performance. We comment upon a few possibilities. (1) The essence of MCT lies in the use of target language translations. Translator noise can have a big impact on the performance, and training bias in favor of a particular language. We believe that there are several approaches that can be attempted to overcome such a challenge. One could identify a set of languages that can mutually benefit from and share a quality MT. Thus, instead of training a single model for all languages, one could train a model for each set. One can also bias the cost function by using language-specific weights; these weights could potentially be used to model translator noise. One could also use a training schedule (along with adapting the learning rate) instead of weights to bias the training in favor of a language or language set. (2) Finally, we admit, for the purposes of illustration, we have made a rather strict assumption of zero human-translated data. It remains of interest to explore the impact of a small volume of human translated data on the performance of MCT, further whether an MCT can be used to sample queries which if translated by a human can help to maximally boost performance, in an active learning framework.

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*.

Alexis Conneau, Ruty Rinott, et al. 2018. XNLI: Evaluating cross-lingual sentence representations. In

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

Arpita Das, Harish Yenala, et al. 2016. Together we stand: Siamese networks for similar question retrieval. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin et al. 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*.

Akiko Eriguchi, Melvin Johnson, et al. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *CoRR*.

Deepak Gupta, Surabhi Kumari, et al. 2018. MMQA: A Multi-domain Multi-lingual Question-Answering Framework for English and Hindi. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Prerna Khurana, Puneet Agarwal, et al. 2017. Hybrid bilstm-siamese network for faq assistance. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM 2017)*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*.

Siwei Lai, Liheng Xu, et al. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*.

Yu-Hsiang Lin, Chian-Yu Chen, et al. 2019. Choosing transfer languages for cross-lingual learning. In *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Bryan McCann, James Bradbury, et al. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Jiaxin Pei and Chenliang Li. 2018. S2SPMN: A simple and effective framework for response generation with relevant information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Matthew Peters, Mark Neumann, et al. 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)*.

Sebastian Schuster, Sonal Gupta, et al. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Holger Schwenk, Ke Tran, et al. 2017. Learning joint multilingual sentence representations with neural machine translation. *CoRR*.

Nitish Srivastava, Geoffrey E Hinton, et al. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*.

Shyam Upadhyay, Manaal Faruqui, et al. 2018a. (almost) zero-shot cross-lingual spoken language understanding. In *Proceedings of the IEEE ICASSP*.

Shyam Upadhyay, Nitish Gupta, et al. 2018b. Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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

Katherine Yu, Haoran Li, et al. 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *The 56th Annual Meeting of the Association for Computational Linguistics*.

Xiangyang Zhou, Lu Li, et al. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.