Attention-Informed Mixed-Language Training for Zero-Shot Cross-Lingual Task-Oriented Dialogue Systems

Zihan Liu,* Genta Indra Winata,* Zhaojiang Lin, Peng Xu, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
{zliucr, giwinata, zlinao, pxuab}@connect.ust.hk, pascale@ece.ust.hk

Abstract

Recently, data-driven task-oriented dialogue systems have achieved promising performance in English. However, developing dialogue systems that support low-resource languages remains a long-standing challenge due to the absence of high-quality data. In order to circumvent the expensive and time-consuming data collection, we introduce Attention-Informed **M**ixed-**L**anguage **T**raining (**MLT**), a novel zero-shot adaptation method for cross-lingual task-oriented dialogue systems. It leverages very few task-related parallel word pairs to generate code-switching sentences for learning the inter-lingual semantics across languages. Instead of manually selecting the word pairs, we propose to extract source words based on the scores computed by the attention layer of a trained English task-related model and then generate word pairs using existing bilingual dictionaries. Furthermore, intensive experiments with different cross-lingual embeddings demonstrate the effectiveness of our approach. Finally, with very few word pairs, our model achieves significant zero-shot adaptation performance improvements in both cross-lingual *dialogue state tracking* and *natural language understanding* (i.e., intent detection and slot filling) tasks compared to the current state-of-the-art approaches, which utilize a much larger amount of bilingual data.

Introduction

Over the past few years, the demand of task-oriented dialogue systems has increased rapidly across the world, following their promising performance on English systems (Zhong, Xiong, and Socher 2018; Wu et al. 2019). However, most dialogue systems are unable to support numerous low-resource languages due to the scarcity of high-quality data, which will eventually create a massive gap between the performance of low-resource language systems (e.g., Thai) and high-resource systems (e.g., English). A common straightforward strategy to address this problem is to collect more data and train each monolingual dialogue system separately, but it is costly and resource-intensive to collect new data on every single language.

Zero-shot adaptation is an effective approach to circumvent the data collection process when there is no training data available by transferring the learned knowledge from a high-resource source language to low-resource target languages. Currently, a few studies have been performed on the *zero-shot learning* in task-oriented dialogue systems (Chen et al. 2018; Schuster et al. 2019). However, there are two problems that exist in this research: (1) the existing methods require a sufficient parallel corpus, which is not ideal for training models on rare languages where bilingual resources are minimal, and (2) the imperfect alignments of cross-lingual embeddings such as MUSE (Conneau et al. 2018) as well as the enormous cross-lingual models XLM (Lample and Conneau 2019), and Multilingual BERT (Devlin et al. 2019) limit the cross-lingual zero-shot transferability.

To address these problems, we propose the **attention-informed mixed-language training (MLT)**, a new framework that leverages extremely small number of bilingual word pairs to build zero-shot cross-lingual task-oriented dialogue systems. The word pairs are created by choosing words from the English training data using attention scores from a trained English model. Then we pair these English words with target words using existing bilingual dictionaries, and use the target words to replace keywords in the training data and build code-switching sentences.[1] The intuition behind training with code-switching sentences is to help the model to identify selected important keywords as well as their semantically similar keywords in the target language. In addition, we incorporate the MUSE, RCSLS (Joulin et al. 2018), and cross-lingual language models XLM and Multilingual BERT for generating cross-lingual embeddings.

During the training phase, our model learns to capture important keywords in code-switching sentences mixed with source and target language words. We conjuncture that learning with task-related keywords of the target language helps the model to capture other task-related words that have similar semantics, for example, synonyms or words in the same category such as days of the week "Domingo" (Sunday) and "Lunes" (Monday). During the zero-shot testing phase, the inter-lingual understanding learned by the model alleviates the main issue of the imperfect alignment of cross-

*The authors contributed equally to this work.

[1] "code-switching" is interchangeable with "mixed-language".

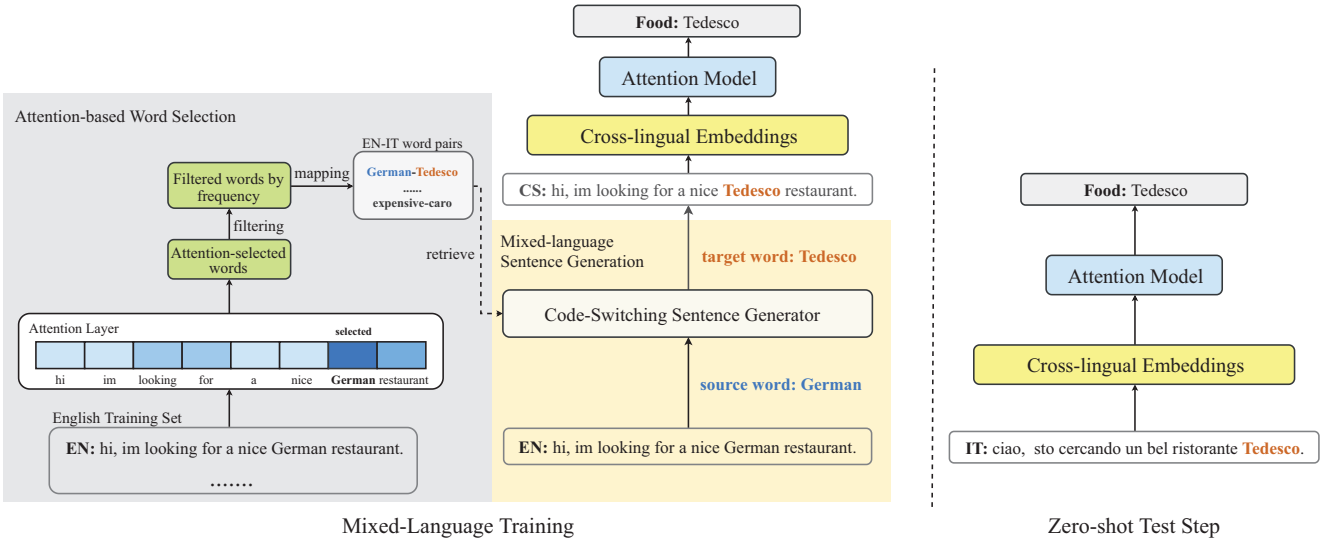


Figure 1: Illustration of the mixed-language training (MLT) approach and zero-shot transfer. **EN** denotes an English text, **IT** denotes an Italian text, and **CS** denotes a code-switching text (i.e., a mixed-language sentence). In the training step, code-switching sentence generator will replace the task-related word with its corresponding translation in the target language to generate code-switching sentences. In the zero-shot transfer step, we leverage cross-lingual word embeddings and directly adapt the trained attention model to the target language.

lingual embeddings. The experimental results on unseen languages show that MLT outperforms existing baselines with significant margins in both dialogue state tracking and natural language understanding tasks on all languages using many fewer resources. This proves that our approach is effective for application to low-resource languages when there is only limited parallel data available.[2]

Contributions in our work are summarized as follows:

- We investigate the extremely low bilingual resources setting for zero-shot cross-lingual task-oriented dialogue systems.
- Our approach achieves state-of-the-art zero-shot cross-lingual performance in both *dialogue state tracking* and *natural language understanding* of task-oriented dialogue systems using many fewer bilingual resources.
- We study the performance of current cross-lingual pre-trained language models (namely Multilingual BERT and XLM) on zero-shot cross-lingual dialogue systems, and conduct quantitative analyses while adapting them to cross-lingual dialogue systems.

Related Work

Task-oriented Dialogue Systems

Dialogue state tracking (DST) and natural language understanding (NLU) are the key components for understanding user inputs and building dialogue systems.

Dialogue State Tracking Mrkšić et al. (2017a) proposed to utilize pre-trained word vectors by composing them into a distributed representation of user utterances and to resolve morphological ambiguity. Zhong, Xiong, and Socher (2018) successfully improved rare slot values tracking through slot-specific local modules.

Natural Language Understanding Liu and Lane (2016) leveraged an attention mechanism to learn where to pay attention in the input sequences for joint intent detection and the slot filling task. Goo et al. (2018) introduced slot-gated models to learn the relationship between intent and slot attention vectors and better captured the semantics of user utterances and queries.

Multilingual Task-oriented Dialogue Systems A number of multilingual task-oriented dialogue systems datasets have been published lately (Mrkšić et al. 2017b; Schuster et al. 2019), enabling evaluation of the approaches for cross-lingual dialogue systems. Mrkšić et al. (2017b) annotated two languages (namely German and Italian) for the dialogue state tracking dataset **WOZ 2.0** (Mrkšić et al. 2017a) and trained a unified framework to cope with multiple languages. Meanwhile, Schuster et al. (2019) introduced a multilingual NLU dataset and highlighted the need for more sophisticated cross-lingual methods.

Cross-lingual Transfer Learning

Cross-lingual transfer learning, which aims to discover the underlying connections between the source and target language, has become a popular topic recently. Conneau et al. (2018) proposed to use zero supervision signals to conduct cross-lingual word embedding mapping and achieved

[2]The code is available at: https://github.com/zliucr/mixed-language-training

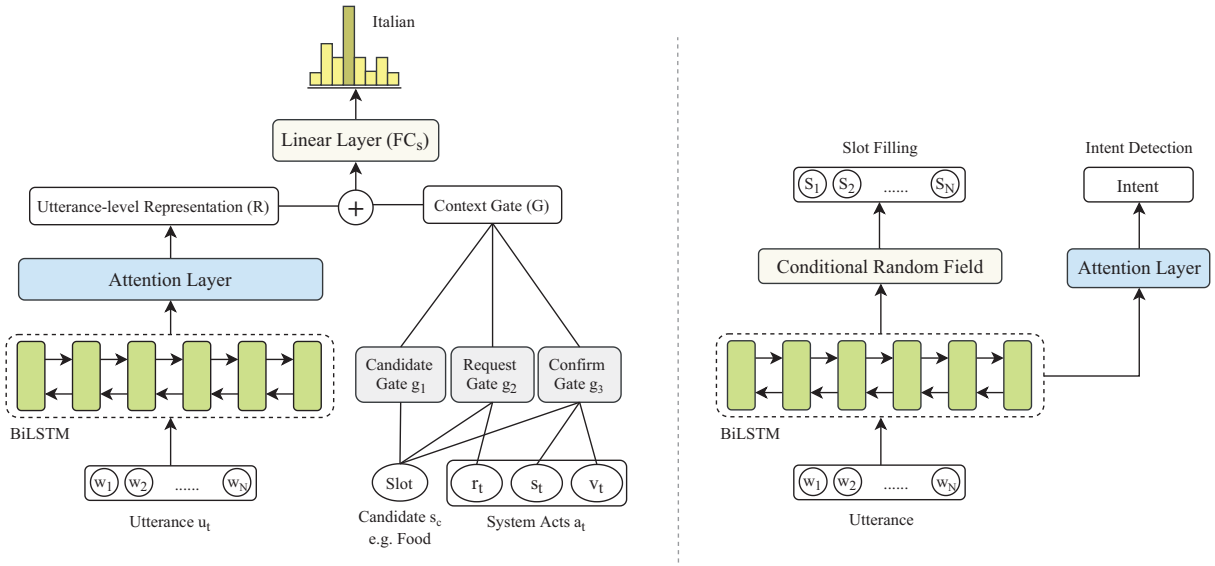


Figure 2: Dialogue State Tracking Model **(left)** and Natural Language Understanding Model **(right)**. For each model, we apply an attention layer to learn important task-related words.

promising results. Devlin et al.; Lample and Conneau (2019; 2019) leveraged large monolingual and bilingual corpus to align cross-lingual sentence-level representations and achieved the state-of-the-art performance in many cross-lingual tasks. Recently, studies have applied cross-lingual transfer algorithms to natural language processing tasks, such as *named entity recognition* (NER) (Ni, Dinu, and Florian 2017), *entity linking* (Pan et al. 2017), *POS tagging* (Kim et al. 2017; Zhang et al. 2016), and *dialogue systems* (Chen et al. 2018; Upadhyay et al. 2018; Liu et al. 2019). Nevertheless, to the best of our knowledge, only a few studies have focused on task-oriented dialogue systems, and none of them investigated the extremely low bilingual resources scenario.

Mixed-Language Training

As shown in Figure 1, in the mixed-language training step, our model is trained using code-switching sentences generated from source language sentences by replacing the selected source words with their translations. In the zero-shot test step, our model directly transfers into the unseen target language.

Attention-based Selection

Intuitively, the attention layer in a trained model can focus on the keywords that are related to the task. As shown in Figure 1, we propose to utilize the scores computed from the attention layer of a trained model on source language (English) data to select keywords for completing the task. Concretely, we first collect source words by taking the top-1 attention score for each source utterance since the source words with the highest attention score are the most important for the given task. However, some noisy words (unimportant words) might still exist in the collection. Hence, we first count the times that the words are selected and filter the words that are seldom selected, and then we choose the top-n most frequent words in the training set as our final word candidates and pair them using an existing bilingual dictionary. We denote the selected n word pairs as a key-value dictionary $D = ((x_1; y_1), \ldots, (x_n; y_n))$, where x and y represent the source and target language, respectively.

Training and Adaptation

Given a source language sentence $\mathbf{w} = [w_1, w_2, \ldots, w_N]$, we replace the words in $\mathbf{w}$ with their corresponding target words if they are present in D to generate a code-switching sentence $\mathbf{w}_{cs}$. As illustrated in Figure 1, we use cross-lingual word embeddings for source and target language words.

$$\mathbf{w}_{cs} = \text{CS}_{gen}(\mathbf{w}), \quad (1)$$

$$out = \text{AttnModel}(E(\mathbf{w}_{cs})), \quad (2)$$

where CS_{gen} represents the code-switching sentence generator in Figure 1, *AttnModel* represents the attention model, and E denotes cross-lingual word embeddings. We specifically use cross-lingual word embeddings from MUSE (Conneau et al. 2018) and RCSLS (Joulin et al. 2018), aligned representations of source and target languages to transfer the learned knowledge from the source language to the target language. By applying mixed-language training, our model can cope with the problem of imperfect alignment of cross-lingual word embeddings. In the zero-shot test step, the attention layer is still able to focus on the same or semantically similar target language keywords, as it does in the mixed-language training step, which improves the robustness of cross-lingual transferability.

Cross-lingual Dialogue Systems

In this section, we focus on applying our mixed-language training approach to cross-lingual task-oriented dialogue

systems. We design model architectures for dialogue state tracking and natural language understanding (i.e., intent detection and slot filling) as follows.

Dialogue State Tracking

Our dialogue state tracking (DST) model, illustrated in Figure 2, is modified from Chen et al. (2018). We model DST into a classification problem based on three inputs: (i) the user utterance u_t, (ii) the slot candidate s_c, and (iii) the system dialogue acts $a_t = (r_t, s_t, v_t)$[3], where we use subscript t to denote each dialogue turn. In short, our model can be decomposed into the following three components:

Utterance Encoder We use a bi-directional LSTM (BiLSTM) to encode the user utterance $u_t = [w_1, w_2 \dots w_N]$ and an attention mechanism (Felbo et al. 2017) on top of the BiLSTM to generate an utterance representation R, where w_i is the word vector of the i-th token and N is the length of the utterance. We formalize the utterance encoder as:

$$[h_1, h_2 \dots h_N] = \text{BiLSTM}([w_1, w_2 \dots w_N]), \quad (3)$$

$$e_i = h_i w_a, \quad \alpha_i = \frac{exp(e_i)}{\sum_{j=1}^{N} exp(e_j)}, \quad R = \sum_{i=1}^{N} \alpha_i h_i, \quad (4)$$

where w_a is a trainable weight vector in the attention layer, and α_i is the attention score of each token i.

Context Gate Given a candidate slot s_c and system acts (r_t, s_t, v_t) as inputs, we compute the context gate G by summing three individual gates: (i) the candidate gate (g_1), (ii) the request gate (g_2), and (iii) the confirm gate (g_3). The context gate is defined as follows:

$$g_1 = E(s_c), \quad g_2 = \sigma(E(s_c) \odot W_1 E(r_t)), \quad (5)$$

$$g_3 = \sigma(E(s_c) \odot W_2(E(s_t) + E(v_t))), \quad (6)$$

$$G = g_1 + g_2 + g_3, \quad (7)$$

where E denotes the word embedding look-up table, $\odot$ denotes a Hadamard product, W_1 and W_2 represent trainable parameter matrices, and σ represents a sigmoid function.

Slot Value Prediction Finally, we concatenate the utterance representation (R) and the context gate (G), which are then passed into a linear layer FC_s and a softmax layer for prediction.

Natural Language Understanding

Our NLU model is illustrated in Figure 2 as a multi-task problem. We describe our model as follows:

Slot Filling We use a BiLSTM-CRF combining a BiLSTM with a conditional random field (CRF) sequence labeling model (Lample et al. 2016) for slot prediction. We pass the hidden states of the BiLSTM through a softmax layer and then pass the resulting label probability vectors through the CRF layer for computing final predictions.

[3] r_t represents the system request, and s_t and v_t represent the system confirmation. For example, when the system requests more information by asking "Do you have an area preference?", then r_t = **"area"**, or when the system confirms by saying "The Vietnamese food is in the cheap price range," then s_t = **"price range"** and v_t = **"cheap"**.

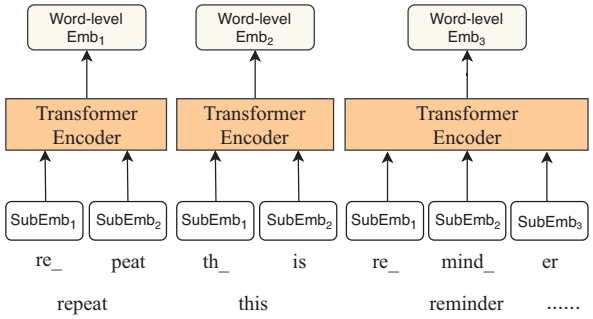


Figure 3: Illustration of how we leverage a *transformer encoder* to incorporate subword embeddings into word-level representations. The parameters in the *transformer encoder* are shared for all subword embeddings.

Intent Prediction We place an attention layer over the hidden states of the BiLSTM and predict the intent for the user utterance through a softmax projection layer. The attention layer is similar to the one in the dialogue state tracking shown in equation (4).

Cross-lingual Language Model

We investigate the effectiveness of current powerful cross-lingual pre-trained language models XLM and Multilingual BERT, and deploy MLT into them for the zero-shot cross-lingual DST and NLU tasks. Lample and Conneau (2019) proposed cross-lingual language model pre-training (XLM) and two objective functions *masked language modeling* (MLM) and *translation language modeling* (TLM). The MLM leveraged a monolingual corpus, the TLM utilized a bilingual corpus, and MLM+TLM incorporated both MLM and TLM. Pre-trained XLM models on 15 languages are publicly available.[4] Multilingual BERT is trained on the monolingual corpora of 104 languages, and the model is also publicly available.[5]

In order to handle multiple languages and reduce the vocabulary size, both methods leverage subword units to tokenize each sentence. However, the outputs of the DST and NLU tasks depend on the word-level information. Hence, we propose to learn the mapping between the subword-level and word-level by adding a transformer encoder (Dehghani et al. 2019) on top of subword units and learn to encode them into word-level embeddings, which we describe in Figure 3. After that, we leverage the same model structures as illustrated in Figure 2 for the DST and NLU tasks.

Experiments

Datasets

Dialogue State Tracking Wizard of Oz (WOZ), a restaurant domain dataset, is used for training and evaluating dialogue state tracking models on English. It was enlarged into

[4] https://github.com/facebookresearch/XLM

[5] https://github.com/google-research/bert/blob/master/multilingual.md

Model	German								
	slot acc.			joint goal acc.			request acc.		
	BASE	**MLT**$_O$	**MLT**$_A$	**BASE**	**MLT**$_O$	**MLT**$_A$	**BASE**	**MLT**$_O$	**MLT**$_A$
MUSE	60.69	68.58	**71.38**	21.57	30.61	**36.51**	74.22	80.11	**82.99**
XLM (MLM)*	52.21	66.26	**68.25**	14.09	29.45	**31.29**	75.15	78.48	**80.22**
+ Transformer	53.81	65.81	**68.55**	13.97	30.87	**32.98**	76.83	78.95	**81.34**
XLM (MLM+TLM)*	58.04	65.39	**66.25**	16.34	29.22	**29.83**	75.73	78.86	**79.12**
+ Transformer	56.52	66.81	**68.88**	16.59	31.76	**33.12**	78.56	81.59	**82.96**
Multi. BERT*	57.61	67.49	**69.48**	14.95	30.69	**32.23**	75.31	83.66	**86.27**
+ Transformer	57.43	68.33	**70.77**	15.67	31.28	**34.36**	78.59	84.37	**86.97**
Ontology Matching†		*24*			-			*21*	
Translate Train†		*41*			-			*42*	
Bilingual Dictionary‡		*51.74*			*28.07*			*72.54*	
Bilingual Corpus‡		*55*			*30.84*			*68.32*	
Supervised Training		*85.78*			*78.89*			*84.02*	
Model	Italian								
	slot acc.			joint goal acc.			request acc.		
	BASE	**MLT**$_O$	**MLT**$_A$	**BASE**	**MLT**$_O$	**MLT**$_A$	**BASE**	**MLT**$_O$	**MLT**$_A$
MUSE	60.59	73.55	**76.88**	20.66	36.88	**39.35**	79.09	82.24	**84.23**
Multi. BERT*	53.34	65.49	**69.48**	12.88	26.45	**31.41**	76.12	84.58	**85.18**
+ Transformer	54.56	66.87	**71.45**	12.63	28.59	**33.35**	77.34	82.93	**84.96**
Ontology Matching†		*23*			-			*21*	
Translate Train†		*48*			-			*51*	
Bilingual Dictionary‡		*73*			*39.01*			*77.09*	
Bilingual Corpus‡		*72*			*41.23*			*81.23*	
Supervised Training		*88.92*			*80.22*			*91.05*	

Table 1: Zero-shot results for the target languages on **Multilingual WOZ 2.0**. **MLT**$_A$ denotes our approach (attention-informed MLT), which utilizes the same number of word pairs as **MLT**$_O$ (90 word pairs). ‡ denotes the results of XL-NBT. Note that, we realize that the goal accuracy in Chen et al. (2018) is calculated as slot accuracy in our paper, so we rerun the models using the provided code (https://github.com/wenhuchen/Cross-Lingual-NBT) to calculate joint goal accuracy. † denotes the results from Chen et al. (2018). Instead of using the *transformer encoder*, we sum the subword embeddings based on the word boundaries to get word-level representations. Due to the absence of the Italian language in the XLM models, we cannot report the results.

WOZ 2.0 by adding more dialogues, and recently, Mrkšić et al. (2017b) expanded WOZ 2.0 into Multilingual WOZ 2.0 by including two more languages (German and Italian). Multilingual WOZ 2.0 contains 1200 dialogues for each language, where 600 dialogues are used for training, 200 for validation, and 400 for testing. The corpus contains three goal-tracking slot types: food, price range and area, and a request slot type. The model has to track the value for each goal-tracking slot and request slot.

Natural Language Understanding Recently, a multilingual task-oriented natural language understanding dialogue dataset was proposed by Schuster et al. (2019), which contains English, Spanish, and Thai across three domains (alarm, reminder, and weather). The corpus includes 12 intent types and 11 slot types, and the model has to detect the intent of the user utterance and conduct slot filling for each word of the utterance.

Experimental Setup

We explore two training settings: (1) without Mixed-language Training (BASE), and (2) Mixed-language Training (MLT). The former trains models only using English data, and then we directly transfer to the target language by leveraging the same cross-lingual word embeddings as our model. The latter utilizes code-switching sentences as the train data. We evaluate our model with cross-lingual embeddings: MUSE (Conneau et al. 2018), RCSLS (Joulin et al. 2018), XLM (Lample and Conneau 2019), and Multilingual BERT (Multi. BERT) (Devlin et al. 2019).

We describe our baselines for the *dialogue state tracking* task in the following:

Ontology-based Word Selection (MLT$_O$) We use dialogue ontology word pairs for mixed-language training since ontology words are all task-related and essential for the DST task.

XL-NBT Chen et al. (2018) proposed a teacher-student framework for cross-lingual neural belief tracking (i.e., dialogue state tracking) by leveraging a bilingual corpus or bilingual dictionary. The model learns to generate close representations for semantically similar sentences across languages.

Ontology Matching Chen et al. (2018) directly used exact string matching for the user utterance according to the ontology words to discover the slot value for each slot type.

Translate Train Chen et al. (2018) used an external bilingual corpus to train a machine translation system, which translates English dialogue training data into target lan-

Model	Spanish						Thai					
	Intent acc.			Slot F1			Intent acc.			Slot F1		
	BASE	**MLT$_H$**	**MLT$_A$**	**BASE**	**MLT$_H$**	**MLT$_A$**	**BASE**	**MLT$_H$**	**MLT$_A$**	**BASE**	**MLT$_H$**	**MLT$_A$**
RCSLS	37.67	77.59	**87.05**	22.23	**59.12**	57.75	35.12	68.63	**81.44**	8.72	29.44	**30.42**
XLM (MLM)	60.8	75.11	**83.95**	38.55	63.29	**66.11**	37.59	46.34	**65.31**	8.12	19.03	**20.43**
+ Transformer	62.33	82.83	**85.63**	41.67	66.53	**67.95**	40.31	57.27	**68.55**	11.45	26.02	**27.45**
XLM (TLM+MLM)	62.48	81.34	**84.91**	42.27	65.71	**66.48**	31.62	50.34	**65.25**	7.91	19.22	**19.88**
+ Transformer	65.32	83.79	**87.48**	44.39	66.03	**68.55**	37.53	68.62	**72.59**	12.84	26.56	**27.98**
Multi. BERT	73.73	77.51	**86.54**	51.73	**74.51**	74.43	28.15	52.25	**70.57**	10.62	24.41	**28.47**
+ Transformer	74.15	82.9	**87.88**	54.28	**74.88**	73.89	26.54	53.84	**73.46**	11.34	26.05	**27.12**
Zero-shot SLU†		*46.64*			*15.41*			*35.64*			*12.11*	
Multi. CoVe		*53.34*			*22.50*			*66.35*			*32.52*	
Multi. CoVe w/ auto		*53.89*			*19.25*			*70.70*			*35.62*	
Translate Train		*85.39*			*72.87*			*95.85*			*55.43*	

Table 2: Results on multilingual NLU dataset (Schuster et al. 2019), and the number of word pairs on both MLT$_H$ and MLT$_A$ is 20. † We implemented the model (Upadhyay et al. 2018) and tested it on the same dataset.

guages (German and Italian) as "annotated" data to supervise the training of DST systems in target languages.

Supervised Training We assume the existence of annotated data for the target languages dialogues state tracking. It indicates the upper bound of the DST model.

We describe our baselines for the *natural language understanding* task in the following:

Human-based Word Selection (MLT$_H$) Due to the absence of ontology in the NLU task, we crowd-source the top-20 task-related source words in the English training set.

Zero-shot SLU Upadhyay et al. (2018) used cross-lingual word embeddings (Bojanowski et al. 2017) to conduct zero-shot transfer learning in the NLU task.

Multi. CoVe Schuster et al. (2019) used Multilingual CoVe (Yu, Li, and Oguz 2018) to encode phrases with similar meanings into similar vector spaces across languages.

Multi. CoVe w/ auto. Based on Multilingual CoVe, Schuster et al. (2019) added an autoencoder objective to produce more general representations for semantically similar sentences across languages.

Translate Train Schuster et al. (2019) trained a machine translation system using a bilingual corpus, and then translated English NLU data into the target languages (Spanish and Thai) for supervised training.

Evaluation Metrics

Dialogue State Tracking We use *joint goal accuracy* and *slot accuracy* to evaluate the model performance on goal-tracking slots. The joint goal accuracy compares the predicted dialogue states to the ground truth at each dialogue turn, and the prediction is correct if and only if the predicted values for all slots exactly match the ground truth values. While the slot accuracy individually compares each slot-value pair to its ground truth. We use *request accuracy* to evaluate the model performance on "request" slot. Similar to joint goal accuracy, the prediction is correct if and only if all the user requests for information are correctly identified.

Natural Language Understanding We use *accuracy* and *BIO-based f1-score* to evaluate the performance of intent prediction and slot filling, respectively.

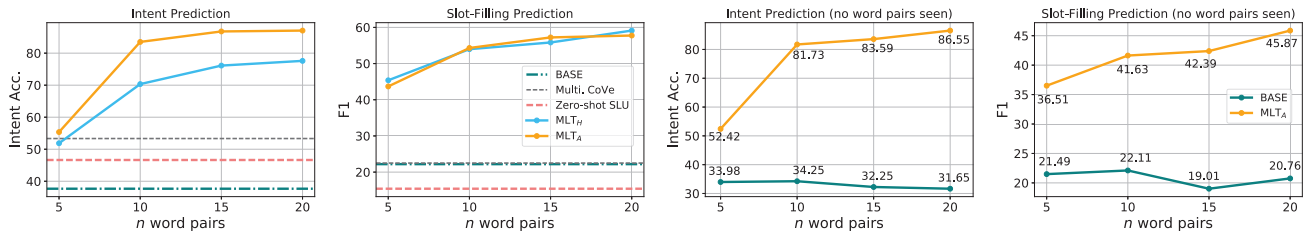
Results & Discussion

Quantitative Analysis

The DST and NLU results are shown in Table 1 and 2. In most cases, our models using MLT significantly outperform the existing state-of-the-art zero-shot baselines, and we achieve a comparable result to the *Multi. CoVe w/ auto* on Thai. Notably, our models achieve impressive performance since we only use a few word pairs and many fewer bilingual resources than sophisticated models such as *Multi. Cove* or *Bilingual Corpus*.

We observe that *ontology matching* is an intuitive method to attempt zero-shot in low-resource languages. However, this method is ineffective because it does not seem able to detect synonyms or paraphrases. Applying ontology pairs into the MLT models copes with this problem and outperforms the BASE models with vast improvements. Interestingly, MLT$_A$ consistently outperforms MLT$_O$ because the attention-based selection mechanism is not only capturing important ontology keywords but also keywords which are not listed in the ontology (i.e., synonyms or paraphrases to the ontology words). For example, word "moderate" is interchangeable with "fair" when users describe the food price during the conversation, which is not listed in the ontology. Since we do not have an ontology in the NLU task, we compare our results with human crowd-sourcing-based word selection (MLT$_H$). Results show that MLT$_A$ significantly outperforms human word pairs selection MLT$_H$ in the intent detection, which further proves the high quality of words selected by the attention layer.

Due to the imperfect alignment of cross-lingual word embeddings, our BASE models with MUSE or RCSLS still suffer from low performance in the zero-shot adaptation. Although we replace these cross-lingual word embeddings with large pre-trained language models such as XLM and Multi. BERT, the performance is not consistently better. This is because the quality of alignment degrades when we combine subword-based embeddings into word-level represen-



(a) (b) (c) (d)

Figure 4: The dynamics of the NLU task: intent and slot-filling results with different numbers of word pairs on Spanish test data using RCSLS. The words are decided according to the frequency in the source language (English) training set. We evaluate on all test data for **(a)** and **(b)**. For **(c)** and **(d)**, we only evaluate on filtered test data that do not contain any word pairs.

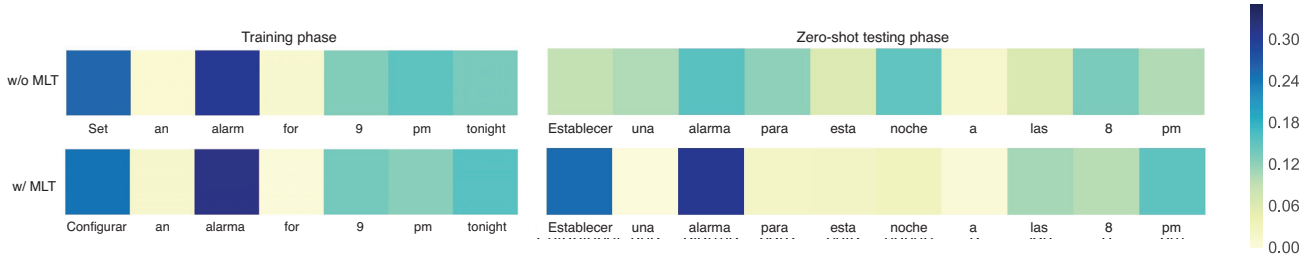


Figure 5: Attentions on words in both training and testing phases. A darker color shows a higher attention score and importance.

tations. The performance of the XLM-based models and Multi. BERT-based models are improved remarkably by applying MLT. Surprisingly, MLT-based models with RCSLS surpass XLM and Multi. BERT by a substantial margin on the Thai language. We find that the length of Thai subword sequences is approximately twice as long as other languages. Hence, the quality of subword-to-word alignments degrades severely.

Performance vs. Number of Word Pairs

Figure 4a and 4b compare the performance of intent and slot-filing predictions on Spanish data with respect to the number of word pairs, and investigates the gap between *human crowd-sourcing-based word selection* (MLT_H) and *attention-based word selection* (MLT_A). Interestingly, with only five word pairs, MLT_A achieves notable gains of 17.69% and 21.45% in intent prediction and slot filling performance, respectively, compared to the BASE model. Compared with human word pairs selection MLT_H, in the intent prediction, MLT_A beats the performance of human-based word selection, and in slot-filling prediction, the result is on par with the MLT_H.

Model Transferability

In Figure 4c and 4d, we show the transferability of MLT_A on the target language data that does not have any target keywords selected from the word pair list. Our model with MLT_A is still able to achieve impressive gains on both intent and slot-filling performance on these data. The results emphasize that the MLT-based model not only memorizes target word replacements, but captures the generic semantics of words and learns to generalize to other words that have a similar vector space, for example, the synonyms "configurer" and "establecer" (both mean "set" in English) or word from the same domain, like "Domingo" (Sunday) and "Lunes" (Monday).

To further support our claims, we extract the attention scores from the attention layer and elaborate on the findings. Figure 5 displays that, in the training phase, our model puts attentions on parallel task-related words in both the source and target languages, such as "Set" and "alarm" in English, and "Configurar" and "alarma" in Spanish. In the zero-shot test phase, our attention layer in the MLT-based models puts an attention on identical or synonym words because they have the same or similar vector representations, respectively, but without MLT, our attention layer fails to do so. Interestingly, we can see clearly in Figure 5 that word 'Establecer" is as equally important as "Configurer", although "Establecer" is not found in the code-switching sentence.

Conclusion

We propose attention-informed mixed-language training (MLT), a novel zero-shot adaptation method for cross-lingual task-oriented dialogue systems using code-switching sentences. Our approach utilizes very few task-related parallel word pairs based on the attention layer and has a better generalization to words that have similar semantics in the target language. The visualization of the attention layer confirms this. Experimental results show that MLT-based models outperform existing zero-shot adaptation approaches in dialogue state tracking and natural language understanding with many fewer resources.

References

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

Chen, W.; Chen, J.; Su, Y.; Wang, X.; Yu, D.; Yan, X.; and Wang, W. Y. 2018. Xl-nbt: A cross-lingual neural belief tracking framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 414–424.

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

Dehghani, M.; Gouws, S.; Vinyals, O.; Uszkoreit, J.; and Kaiser, L. 2019. Universal transformers. In *International Conference on Learning Representations*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.

Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; and Lehmann, S. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1615–1625.

Goo, C.-W.; Gao, G.; Hsu, Y.-K.; Huo, C.-L.; Chen, T.-C.; Hsu, K.-W.; and Chen, Y.-N. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 753–757.

Joulin, A.; Bojanowski, P.; Mikolov, T.; Jégou, H.; and Grave, E. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2979–2984.

Kim, J.-K.; Kim, Y.-B.; Sarikaya, R.; and Fosler-Lussier, E. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2832–2838.

Lample, G., and Conneau, A. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Liu, B., and Lane, I. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *Interspeech 2016* 685–689.

Liu, Z.; Shin, J.; Xu, Y.; Winata, G. I.; Xu, P.; Madotto, A.; and Fung, P. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. 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)*, 1297–1303.

Mrkšić, N.; Séaghdha, D. Ó.; Wen, T.-H.; Thomson, B.; and Young, S. 2017a. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1777–1788.

Mrkšić, N.; Vulić, I.; Séaghdha, D. Ó.; Leviant, I.; Reichart, R.; Gašić, M.; Korhonen, A.; and Young, S. 2017b. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics* 5(1):309–324.

Ni, J.; Dinu, G.; and Florian, R. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1470–1480.

Pan, X.; Zhang, B.; May, J.; Nothman, J.; Knight, K.; and Ji, H. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1946–1958.

Schuster, S.; Gupta, S.; Shah, R.; and Lewis, M. 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, Volume 1 (Long and Short Papers)*, 3795–3805.

Upadhyay, S.; Faruqui, M.; Tür, G.; Dilek, H.-T.; and Heck, L. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6034–6038. IEEE.

Wu, C.-S.; Madotto, A.; Hosseini-Asl, E.; Xiong, C.; Socher, R.; and Fung, P. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. *arXiv preprint arXiv:1905.08743*.

Yu, K.; Li, H.; and Oguz, B. 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, 175–179.

Zhang, Y.; Gaddy, D.; Barzilay, R.; and Jaakkola, T. 2016. Ten pairs to tag–multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1307–1317.

Zhong, V.; Xiong, C.; and Socher, R. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1458–1467.