Adversarial Cross-Lingual Transfer Learning for Slot Tagging of Low-Resource Languages

1st Keqing He
Beijing University of Posts and Telecommunications
Beijing, China
keqing@bupt.edu.cn

2nd Yuanmeng Yan
Beijing University of Posts and Telecommunications
Beijing, China
yanyuanmeng@bupt.edu.cn

3rd Weiran Xu
Beijing University of Posts and Telecommunications
Beijing, China
xuweiran@bupt.edu.cn

***Abstract*—Slot tagging is a key component in a task-oriented dialogue system. Conversational agents need to understand human input by training on large amounts of annotated data. However, most human languages are low-resource and lack annotated training data for slot tagging task. Therefore, we aim to leverage cross-lingual transfer learning from high-resource languages to low-resource ones. In this paper, we propose an adversarial cross-lingual transfer model with multi-level language shared and specific knowledge to improve the slot tagging task of low-resource languages. Our method explicitly separates the model into the language-shared part and language-specific part to transfer language-independent knowledge. To refine shared knowledge in the latent space, we add a language discriminator and employ adversarial training to reinforce feature separation. Besides, we adopt a novel multi-level feature transfer in an incremental and progressive way to acquire multi-granularity shared knowledge. To mitigate the discrepancies between the feature distributions of language specific and shared knowledge, we propose the neural adapters to fuse features from different sources. Experiments show that our proposed model consistently outperforms monolingual baseline with a statistically significant margin up to 2.09%, even higher improvement of 12.21% in the zero-shot setting. Further analysis demonstrates that our method could effectively alleviate data scarcity of low-resource languages.**

***Index Terms*—slot tagging, cross-lingual transfer learning, language discriminator, multi-level knowledge representation, neural adapter**

I. Introduction

Goal-oriented dialogue systems rely on a slot tagging component to extract key information from the natural language used in conversation [1]–[3]. Slot tagging aims to obtain the semantic structure for a given utterance. For instance, given an utterance "What flights travel from las vegas to los angeles", slot tagging captures the semantic labels for each tokens where the label "B-fromloc" denotes the corresponding token "las" as the beginning token of original location. The state-of-the-art models [4], [5] guarantee exceptionally high accuracy on the task under the availability of large-scale annotated datasets. However, more than 6,500 low-resource languages around the world lack the labeled data necessary for training these deep models. In this paper, we concentrate on adversarial multi-level cross-lingual transfer learning from high-resource languages to low-resource languages to improve the slot tagging task of low-resource languages.

The challenges of cross-lingual transfer learning are two-folds: (1) Multilingual shared knowledge for transfer learning should comprise different levels of linguistic features from multi-sources. It's essential to define and obtain knowledge in a unified way. (2) Integration at each level of the knowledge hierarchy should be adopted and discrepancies between languages must be considered. Existing approaches [6]–[8] for the cross-lingual SLU task use auxiliary machine translation procedure to either generate supervision in the target language automatically or convert the test data to English. However, these approaches will fail in languages for which machine translation is not reliable, or even unavailable. Other works [8]–[12] employ a language-shared encoding layer, such as character embedding, aligned word embedding or context encoder. All the methods only focus on weight sharing but ignore the discrepancy between languages. Besides, most of the models just consider single-layer parameter transfer. Nevertheless, multi-level linguistic knowledge and integration at each layer of the feature hierarchy are essential to cross-lingual transfer learning.

Inspired by the previous works, our motivations are two-folds: (1) Linguistic knowledge between languages consists of language shared and specific parts and should be separated explicitly. Integration of two parts needs to tackle discrepancies between languages. (2) Multi-level knowledge transfer and feature hierarchy should be considered. In this paper, we propose a multi-level cross-lingual transfer model with language shared and specific knowledge to boost the performance of monolingual slot tagging for low-resource languages. Specifically, we explicitly separate the model into the language-shared part and language-specific part to issue the discrepancies between languages. To refine shared knowledge, we add a language discriminator acting as a classifier to determine which language the knowledge encoded in the language-shared feature extractor belongs to. Besides, we adopt multi-level knowledge transfer including char-level embeddings, word-level representation, sentence-level semantics, and tag-level correlation in an incremental and progressive way

* This work was partially supported by National Key R&D Program of China No. 2019YFF0303300 and Subject II No. 2019YFF0303302, DOCOMO Beijing Communications Laboratories Co., Ltd, MoE-CMCC "Artifical Intelligence" Project No. MCM20190701.

* Weiran Xu is the corresponding author.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

to acquire multi-granularity shared knowledge. To mitigate the disparities between the feature distributions of language specific and shared knowledge, we propose the neural adapters to fuse them automatically rather than direct concatenation.

Generally, our main contributions are:

1) We propose a cross-lingual slot tagging framework explicitly leveraging both language shared and specific knowledge.
2) Our model adopts multi-level cross-lingual knowledge transfer including char-level embeddings, word-level representation, sentence-level semantics, and tag-level correlation. Linguistic knowledge can be integrated at each layer of the feature hierarchy via the neural adapters.
3) We add a language discriminator to the shared feature extractor and employ adversarial training for the whole model to reinforce the performance of feature separation and slot tagging simultaneously.

II. Monolingual Baseline Architecture

In this paper, we mainly focus on the fundamental slot tagging task. The goal of slot tagging is to assign a categorical label to each token in a given sentence. Though traditional methods such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) [13], [14] achieved high performance on slot tagging tasks, they typically relied on hand-crafted features, therefore it is difficult to adapt them to new tasks or languages. Recently plentiful proposals based on the neural network [15], [16] have been proved to make a significant difference.

Therefore, we design our monolingual baseline architecture adopted from the LSTM-CNNs [16] as shown in Fig 1. First, each word x_i is represented as the concatenation r_i of word embedding and character embedding which is extracted from a CharCNN network. Then, a bidirectional LSTM processes the sequence $R = \{r_1, ..., r_L\}$ where L is the length of the input sentence. The output of biLSTM is a sequence $H = \{h_1, ..., h_L\}$ consisting of L fixed-size vectors. Next, a linear layer transforms each h_i to a score vector y_i, in which each component represents the predicted score of a target tag. To model correlations between tags, a CRF layer is added at the top to generate the best tagging path for the whole sequence.

III. Adversarial Cross-Lingual Transfer Learning with Language Shared and Specific Knowledge

In this part, we will adequately delineate our multi-level cross-lingual model with language shared and specific knowledge via transfer learning. We start from a brief description of the overall architecture and then dive into the details of each part of the proposed model.

Suppose that we have a dataset $D_s = \{(X_i, Y_i)\}_{i=1}^{N_s}$ of English, where s represents the source language, $X_i = (x_1, ..., x_L)$ is the input sequence and $Y_i = (y_1, ..., y_L)$ is the corresponding tag sequence. N_s is the size of the dataset D_s. Analogical to the notations of English dataset D_s, the Spanish dataset is formalized as $D_t = \{(X_j, Y_j)\}_{j=1}^{N_t}$. In this paper, we aim at improving the performance of low-resource languages via multi-level cross-lingual transfer learning. Therefore, we use the dataset from [8] where $N_s \approx 8.5N_t$. See more detailed statistical summary in the dataset section. [1]

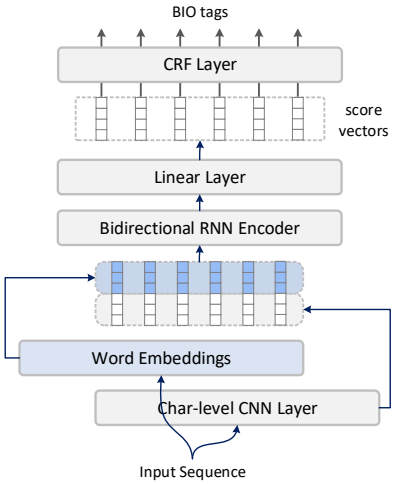


Fig. 1. Monolingual baseline architecture for slot tagging.

A. Model Overview

Fig 2 illustrates the full architecture of our multi-level cross-lingual model. Generally, the proposed model contains three submodules represented by the three columns in the figure respectively. Thereof, the left and right columns denote language-specific slot tagging models for English and Spanish respectively and the middle column performs as a language-shared feature extractor. The main idea of our method is to transfer language-shared knowledge from the multilingual setting to improve the monolingual slot tagging, especially for low-resource languages. To acquire multi-granularity shared knowledge with significant generalization capability and avoid catastrophic forgetting, we adopt multi-level knowledge transfer including char-level embeddings, word-level representation, sentence-level semantics, and tag-level correlation in an incremental and progressive way via neural adapters. By combining previously learned features in this manner, our model achieves richer compositionality, in which prior knowledge is no longer transient and can be integrated at each layer of the feature hierarchy. Besides, we add a language discriminator acting as a classifier to determine whether the knowledge encoded in the language-shared feature extractor is from English or Spanish. When a well-trained discriminator can't classify the language of the input sequence properly, we can think the shared knowledge is language invariant [17].

In the subsequent sections, we will elaborate each component of our multi-level cross-lingual model, and how they

[1]We describe our model based on the assumption that English is of high-resource languages and Spanish is of low-resource languages. Our model can still be applied to other languages.

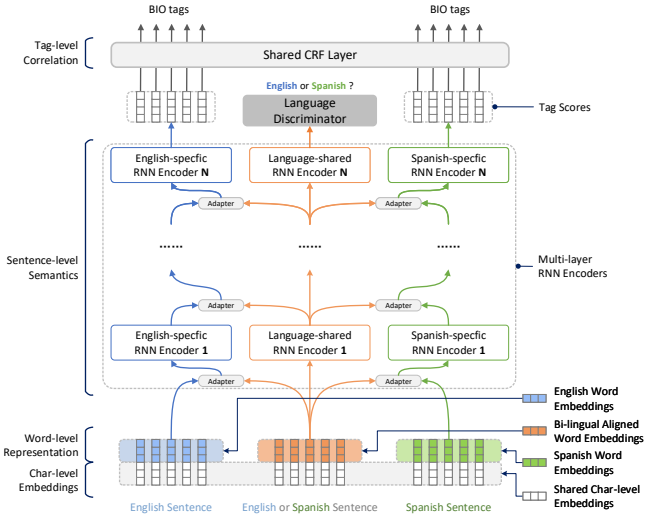


Fig. 2. The overall architecture of our multi-level cross-lingual model.

are combined to boost the performance of the monolingual baseline model for slot tagging.

B. Shared Character Embeddings

On the first level of the proposed model, we construct the basis of the architecture by sharing character-level embeddings among all languages. This level of parameter sharing aims to provide universal character representation and morphological feature extraction capability for English and Spanish. We suppose sharing character-level embeddings results in a common character embedding space between the two languages, and intuitively should allow for more efficient transfer learning at the character level. [2]

Character-level features can represent morphological and semantic information; e.g., the English morpheme *"dis-"* usually indicates negation and reversal as in *"disagree"* and *"disapproval"*. For low-resource languages lacking data to suffice the training of high-quality word embeddings, character embeddings learned from other languages may provide crucial information for labeling, especially for rare and out-of-vocabulary words. For example, the English word *"unhappiness"* acts as a noun, meaning *"not happy"*. Even if it's absent in the vocabulary, the model may infer its meaning from the suffix *"un"* (present negation), root *"happy"*, and the prefix *"ness"* (represent nominalization).

Our implementation follows [16], employing a character-level CNN to extract character-level features for each word. Specifically, the model uses a character embedding matrix to lookup a sequence of vectors, corresponding to each character in a word. These vectors are subsequently feed into a convolution layer with multiple kernels, followed by a max-pooling layer.

[2]Our work mainly focus on languages with the same alphabet, not considering languages like Japanese which is written using an unsegmented mixture of two syllabaries as well as thousands of Chinese characters. [10] proposed a Mixed Orthographic Model (MOM) to tackle the issue. We leave it to our future work.

C. Monolingual and Aligned Cross-Lingual Word Embeddings

In addition to character-level embeddings, word embeddings are still essential to represent lexical and semantic information. We develop an integration strategy of combining two kinds of word embeddings for language shared and specific information. For monolingual slot tagging submodules of English and Spanish represented by the left and right columns in Fig 2, we apply pre-trained English and Spanish word embeddings respectively to capture language-specific knowledge. The formalization defines as follows:

$$r_i^s = e_{x_i}^c \oplus e_{x_i}^w, \forall i = 1, \ldots, L^s \quad (1)$$

$$r_j^t = e_{x_j}^c \oplus e_{x_j}^w, \forall j = 1, \ldots, L^t \quad (2)$$

where the final word representation r_i^s of English is the concatenation of shared character embeddings $e_{x_i}^c$ and monolingual pre-trained word embeddings $e_{x_i}^w$. L^s is the length of input sequence. We could obtain the Spanish word representation r_j^t in the same way. For the language-shared feature extractor, we employ the multilingual word embeddings aligned to a common vector space. The embeddings are pre-computed in an offline fashion and are not adapted while training the whole model. In our experiments, we use MUSE [18] cross-lingual embeddings.

Since the word embeddings for source and target language share a common vector space, the shared parts of the target low-resource language model are capable of processing data samples from the completely unseen target language and perform accurate prediction, i.e. enabling zero-shot cross-lingual slot tagging.

D. Multi-Layer Language Shared and Specific RNN Encoders with Neural Adapters

Since we get char-level and word-level representations for each word in addition to shared word embeddings, the proposed model wishes to incorporate language-shared knowledge

to enhance the slot tagging of low-resource language via non-linear neural adapters [19], [20].

Specifically, we use gate mechanism as the implementation of neural adapters. Suppose the language-specific representation for source language $r_i^s \in \mathbf{R}^E$ and the language-shared representation $r_i \in \mathbf{R}^E$, where E is the embedding size, the output of neural adapter is calculated as follows:

$$g = \sigma(\boldsymbol{W}_g \cdot [r_i^s; r_i] + b_g) \tag{3}$$

$$r_i^{out} = g \otimes r_i^s + (1 - g) \otimes r_i \tag{4}$$

where $\boldsymbol{W}_g \in \mathbf{R}^{E \times 2E}$ and $b_g \in \mathbf{R}^E$ are trainable parameters in each neural adapter, $\sigma(\cdot)$ denotes *sigmoid* operation, and $\otimes$ denotes the element-wise multiplication. Here we only give formula of neural adapter for source language, but it is absolutely the same when it comes to target language.

In our work, we aim to transfer cross-lingual knowledge in an incremental and progressive way. To mitigate the discrepancies between the feature distributions of language specific and shared knowledge, we propose the neural adapters to fuse them. Besides, the neural adapters could automatically adjust the appropriate scales of the different knowledge inputs and avoid catastrophic forgetting.

The previous works [10]–[12], [20], [21] usually take care of knowledge transfer of the top layer, immune to the inner feature hierarchy of multi-layer RNN encoders. We assume that each layer of RNN encoders comprises coarse-to-fine granularity complementary information, facilitating gains of sentence-level semantics. Experiment results show that the incorporation of multi-layer sentence-level semantics substantially boosts the performance of natural language understanding under low-resource settings.

E. Shared CRF Layer

In our experiment setting, English corpus D_s and Spanish corpus D_t share the same tag set. Therefore, we design the cross-lingual shared CRF layer to unearth the tag-level correlation. For example, our model corrects Spanish phrase *[B-datetime veinte] [B-datetime minutos]* (means twenty minutes in English) to *[B-datetime veinte] [I-datetime minutos]* because the CRF layer trained on plenty of English sentences assigns a low score to the rare transition *(B-datetime, B-datetime)* and promotes *(B-datetime, I-datetime)*. We expect this tag-level correlation could help learn shared transition knowledge from high-resource languages.

In the shared CRF layer, given an input sentence $\boldsymbol{x}$ of length L and the tag scores $\boldsymbol{y}$, the final score of a sequence of tags $\boldsymbol{z}$ is defined as:

$$S(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \sum_{t=1}^{L} \left(\boldsymbol{A}_{z_{t-1}, z_t} + \boldsymbol{y}_{t, z_t} \right) \tag{5}$$

where $\boldsymbol{A}$ is a transition matrix in which $\boldsymbol{A}_{p,q}$ represents the binary score of transitioning from tag p to tag q, and $y_{t,z}$ represents the unary score of assigning tag z to the t-th word.

TABLE I
SUMMARY STATISTICS OF THE DATASET WE USE. THE THREE VALUES IN TABLE CELLS CORRESPOND TO THE NUMBER OF UTTERANCES IN THE TRAINING, DEVELOPMENT, AND TEST SPLITS.

Domain	Number of utterances		Slot types
	English	Spanish	
Alarm	9,282/1,309/2,621	1,184/691/1,011	2
Reminder	6,900/943/1,960	1,207/647/1,005	6
Weather	14,339/1,929/4,040	1,226/645/1,027	5
Total	30,521/4,181/8,621	3,617/1,983/3,043	11

Given the ground truth sequence of tags $\boldsymbol{z}$, we maximize the following objective function during the training phase:

$$\begin{aligned} \mathcal{O} &= \log P(\boldsymbol{z}|\boldsymbol{x}) \\ &= S(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) - \log \sum_{\tilde{\boldsymbol{z}} \in Z} e^{S(\boldsymbol{x}, \boldsymbol{y}, \tilde{z})} \end{aligned} \tag{6}$$

where $\boldsymbol{Z}$ is the set of all possible tagging paths.

F. The Language Discriminator

The language discriminator aims at facilitating shared cross-lingual features of the source language and target language through adversarial learning. We would like to differentiate the language-shared features from language-specific features to improve the generalization capability of transfer learning. When a well-trained discriminator can't classify the language of the input sequence properly, we can think the shared knowledge is language invariant [17].

As the previous works [22], [23] applied CNN to some related classification tasks, we construct our discriminator with CNN in a homogeneous manner. First, the input to the CNN is the context-aware representation $\boldsymbol{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_L\}$ of the input sentence. We introduce a convolutional layer upon $\boldsymbol{h}_i$:

$$\boldsymbol{f}_i = \mathit{Relu}\left(\boldsymbol{W}\boldsymbol{h}_{i-k:i+k} + b\right) \tag{7}$$

where $\boldsymbol{h}_{i-k:i+k}$ is the concatenation of vectors from $\boldsymbol{h}_{i-k}$ to $\boldsymbol{h}_{i+k}$, $\boldsymbol{W}$ and b are convolutional kernel and the bias term respectively. k is the (one-side) window size of convolutional layer. A number of different kinds of kernels with different windows sizes are used in our work to extract different features at different scales. Next, we apply a max-over-time pooling operation over the feature maps to get a new feature map $\boldsymbol{f}$. Finally, the feature map is fed into a fully connected network with a sigmoid activation function to make the final predictions:

$$p(d) \propto \exp\left(\mathbf{W}_d \cdot \mathbf{f} + b_d\right) \tag{8}$$

where d is the language label of source language English or target language Spanish.

IV. EXPERIMENTS

A. Dataset

In our experiments, we use the cross-lingual slot tagging dataset presented in [8]. The dataset contains English, Spanish and Thailand sentences annotated according to the same annotation scheme, from which we select English as source language and Spanish as target language. It consists of three

TABLE II
PERFORMANCE COMPARISON BETWEEN MONOLINGUAL AND CROSS-LINGUAL MODELS WITH DIFFERENT COMPONENTS.

Language	Setting	Model	F1-score
Spanish	Monolingual	Baseline without CRF	84.77
	Monolingual	Baseline(full model)	86.05
	Cross-lingual	Our method without CRF	86.12
	Cross-lingual	Our method without neural adapters	86.58
	Cross-lingual	Our method without language discriminator	86.98
	Cross-lingual	Our method(full model)	**88.14***
English	Monolingual	Baseline without CRF	94.85
	Monolingual	Baseline(full model)	95.49
	Cross-lingual	Our method without CRF	95.35
	Cross-lingual	Our method without neural adapters	95.69
	Cross-lingual	Our method without language discriminator	95.93
	Cross-lingual	Our method(full model)	**96.05**

domains: Alarm, Reminder and Weather. Note that we report the scores on the overall dataset rather than single domain.

Table I contains several summary statistics of the dataset. Note that the percentage of training examples as compared to development and test examples is much higher for the English data than for the Thai and Spanish data. We decided for a more even split for the latter two languages so that we had a sufficiently large data set for model selection and evaluation.

B. Implementation Details

We implement all slot tagging models within the AllenNLP framework [24]. We train models for 500 epochs and select the model that performs best on the development set via early stopping. We use the Adam optimizer [25] with a learning rate of 0.0001. We train our models with a batch size of 256 in the cross-lingual slot tagging dataset. We use dropout of 0.2 in the BiLSTM and set the size of the BiLSTM layers to 300 dimensions. For multilingual word embeddings in the experiments, we use Multilingual Unsupervised or Supervised word Embeddings(MUSE) to transfer word-level shared knowledge.

C. Comparison of Different Models

In this part, we present the empirical results under different experiment settings. In Table II, we compare our method with the monolingual LSTM-CNNs model(denoted as *baseline*) with/without CRF layer, our method without CRF, our method without neural adapters and our method without language discriminator for both target language Spanish and source language English. All of the experiments above are performed on the full dataset. We leave zero-shot transfer learning to the next section. The numbers with * indicate that the improvement of our model over all baselines is statistically significant with $p < 0.05$ under t-test.

From Table II, we can see that our method substantially outperforms the monolingual baseline model up to 2.09% in target language Spanish. Meanwhile, the components we propose, neural adapter and language discriminator, also consistently achieve statistically significant improvements. Furthermore, whether the CRF layer is applied to the slot tagging model has little effect on performance improvement, which reveals changing the capacity of the model(such as adding more RNN layers, changing different activation function, etc) does not exacerbate the effect of our method. In other words, our method obtains consistent performance gain from cross-lingual knowledge transfer, whatever architecture of the basic monolingual slot tagging model differs. Note that we employ the same basic architecture of slot tagging model, except for the progressive connections in all experiments. Therefore, we could replace the baseline architecture with other slot tagging models. For instance, the RNN layers could be substituted by Transformer layers.

TABLE III
COMPARISON WITH STATE-OF-THE-ART MODELS. *monolingual baseline* IS THE BASIC ARCHITECTURE WE EXPLAIN IN MONOLINGUAL BASELINE ARCHITECTURE SECTION. *our method* IS THE OVERALL ARCHITECTURE OF MULTI-LEVEL CROSS-LINGUAL MODEL WE PROPOSE.

Language	Model	F1-score
Spanish	monolingual baseline	86.05
	NeuroNER [26]	86.72
	multi-lingual multi-task [12]	87.90
	ELMo [27]	88.01
	Bert [28]	87.10
	our method	**88.14**

Results of source language English show that although we design this method for low-resource settings, it also achieves good performance in high-resource settings. The empirical results substantiate shared knowledge across different languages facilitates natural language understanding, especially for low-resource languages.

D. Comparison with State-of-the-Art Models

In Table III, we compare our model with more state-of-the-art sequence tagging models using all Spanish and English data. Both ELMo [27] and Bert [28] are recent methods of contextual word representations, which have been proved to drastically improve a large portion of tasks of natural language understanding. In our experiments, we typically employ the same setting as the original paper did. For emphasis, Spanish represents the low-resource language as the source language while English is referred to the target language. Since the

TABLE IV

SLOT TAGGING EXAMPLES OF SPANISH FROM THE MONOLINGUAL BASELINE MODEL AND OUR METHOD, EACH OF WHICH CONTAINS A SPANISH TEXT, A CORRESPONDING ENGLISH TRANSLATION, RESULTS OF MONOLINGUAL BASELINE AND OUR METHOD. THE [GREEN] ([RED]) HIGHLIGHT INDICATES A CORRECT (INCORRECT) TAG. FOR SIMPLICITY, WE OMIT CORRECT **O** TAGS.

#1 Spanish text: *¿ Mi àrea es propensa a tornados*
English translation: *is my area prone to tornadoes*
⋆ monolingual baseline: O [O] [O] O O O [B-weather/attribute]
⋆ our method: O [B-location] [I-location] O O O [B-weather/attribute]
#2 Spanish text: *cual es el pronóstico para la proxima semana*
English translation: *what is the forecast for next week*
⋆ monolingual baseline: O O O [O] O O [B-datetime] [I-datetime]
⋆ our method: O O O [B-weather/noun] O O [B-datetime] [I-datetime]
#3 Spanish text: *¿ Va a llover en Atlanta este fin de semana*
English translation: *is it going to rain in Atlanta this weekend*
⋆ monolingual baseline: O O O [B-weather/attribute] O [O] [B-datetime] [I-datetime] [B-datetime] [I-datetime]
⋆ our method: O O O [B-weather/attribute] O [B-location] [B-datetime] [I-datetime] [I-datetime] [I-datetime]

main purpose of our method is to improve the performance of low-resource languages, we only conduct our experiments on Spanish for computation cost.

Results show that our method outperforms these state-of-the-art models with a statistically significant margin, especially over recent ELMo and Bert. Since ELMo and Bert are trained on a large monolingual Spanish corpus to learn linguistic knowledge, we argue that our method could effectively learn better cross-lingual knowledge from the source language English.

V. QUALITATIVE ANALYSIS

A. Effect of Cross-Lingual Learning

In Table IV, we demonstrate some slot tagging examples from the monolingual baseline model and our method.

The first example of Table IV shows that shared character-level embeddings boost the performance of transfer learning. Rare words in Spanish may utilize shared lexical knowledge from similar words in English. For instance, monolingual baseline fails to recognize *"Mi àrea"*, extremely rare words in Spanish dataset, while our method is capable of understanding them. We assume that our method transfer shared morphological information from similar English words *"my area"*.

The next example proves that cross-lingual word embeddings also help identify rare words in low-resource languages via aligned embedding space. In example 2, *"pronóstico"* only occurs six times in Spanish dataset, so monolingual baseline cannot recognize it. However, *"pronóstico"* is intimately associated with *"forecast"* which is frequent in English dataset, enabling our method to transfer knowledge from high-resource languages.

The final example shows that the shared CRF layer learns the transition relationship among slot tags from English and applies it to Spanish. For instance, the monolingual baseline mistakenly assigns *"B-datetime I-datetime B-datetime I-datetime"* to *"este fin de semana"*. By contrast, our method adopts knowledge from English dataset where *"B-datetime I-datetime I-datetime I-datetime"* occurs more frequently and gets a higher transition score.

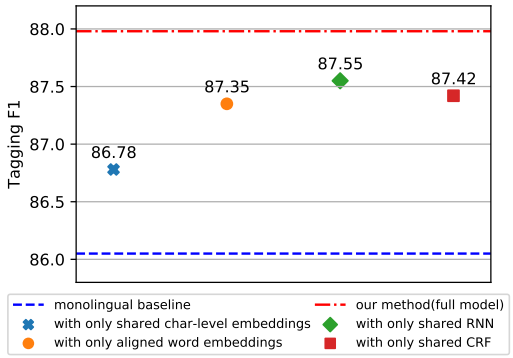


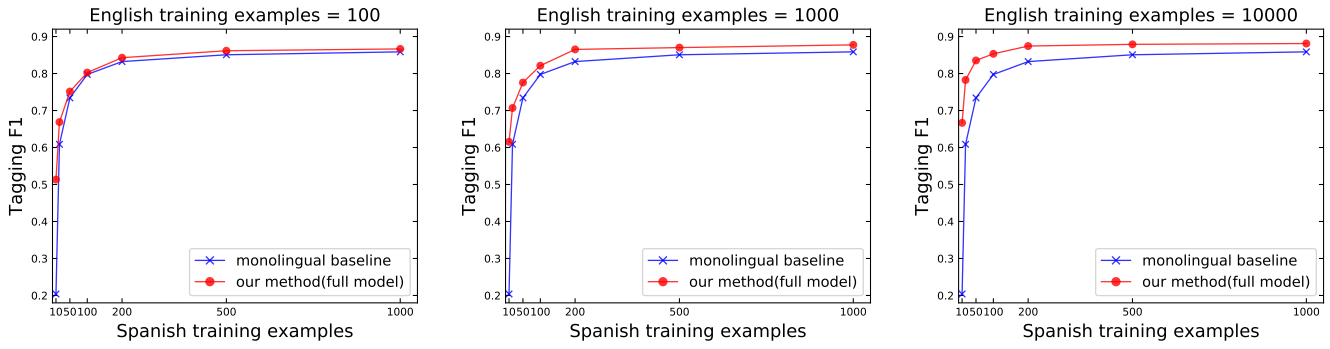
Fig. 3. Effects of different shared layer for cross-lingual transfer learning.

All of the examples substantiate that our cross-lingual method could effectively alleviate data scarcity of target languages by transferring shared knowledge from high-resource languages to low-resource ones.

B. Multi-Level Knowledge Transferability

In this part, we shed light on the transferability of multi-level knowledge and unearth effects of different shared layer for cross-lingual transfer learning. Fig 3 shows the performance gain of our models with different variants compared to the monolingual baseline. We perform 4 model variants, each of which only contains one shared layer, such as char-level embeddings, aligned word-level embeddings, sentence-level RNN layer and tag-level CRF layer.

Experiment Setting The cross-lingual model with only shared char-level embeddings represents the submodules of the left and right columns in Fig 2 only share char-level embeddings without aligned word embeddings, language-shared encoders, while the other parts are the same as the monolingual baseline. For clarity, we eliminate the language discriminator component among all of the models in Fig 3. For the model with only shared RNN encoders, we employ monolingual word embeddings as the input of shared RNN encoders. The only difference between baseline and model with only shared CRF



(a) (b) (c)

Fig. 4. Results of few-shot learning for Spanish with different sizes of English training examples. We report all of these F1-scores across 5 training runs. Figure (a) (b) (c) represent the learning curves when the size of English data is 100, 1000, and 10000 respectively.

is whether the CRF layer is shared by the source language and the target language.

Results and Analysis Results of Fig 3 convincingly substantiate shared RNN layers make a notable difference compared to the shared layers of other levels. We assume that the performance gain is predominantly attributed to sentence-level language-shared semantics. Besides, aligned cross-lingual word embeddings and shared CRF layer also make exceptional effects. Overall, all levels of shared knowledge facilitate to recognize slot tags and work in a complementary way.

C. Zero-shot Transfer Learning

As mentioned in the previous section, just from the results on the full dataset, it is not entirely clear whether there is a significant advantage of our multi-level cross-lingual method. Therefore, we conduct additional experiments with plentiful smaller training sets of the target language Spanish: the case where no annotated data in the target language exists (zero-shot) and the case where a very limited amount of training data in the target language exists(few-shot). We would expect our multi-level cross-lingual model with language shared and specific knowledge to facilitate much better performance gain in the zero-shot and very low-resource scenarios than monolingual baseline.

Experiment Setting We use the same models with the same hyperparameters in this section, except the number of Spanish training utterances. In the zero-shot case of Table V, we only use English data for training. For the learning curve experiments, we sample 10, 20, 50, 100, 200, 500, or 1000 utterances from the target language Spanish for training and upsample the target language data so that it roughly matches the size of the English data. For these experiments, we report the average F1-scores of slot tagging across 5 runs. Fig 4(a) (b) (c) represent the learning curves respectively when the size of English data is 100, 1000, and 10000.

Results and Analysis Table V shows the results of zero-shot learning for Spanish with different sizes of English training examples. These results indicate that our multi-level cross-lingual method consistently outperforms monolingual baseline with a statistically significant margin up to 12.21%. Multi-level cross-lingual shared knowledge can significantly boost the performance of slot tagging for low-resource languages on the zero-shot setting. Moreover, the size of the English data also makes a difference. More data from source languages results in better performance gains. We assume that sufficient source data demonstrates requisite linguistic knowledge.

TABLE V
RESULTS OF ZERO-SHOT LEARNING FOR SPANISH WITH DIFFERENT SIZES OF ENGLISH TRAINING EXAMPLES WHILE THE SIZE OF SPANISH TRAINING EXAMPLES IS 0.

Target Language	Model	Training Examples of English	F1-score
Spanish	Monolingual baseline	0	6.48
Spanish	Our method	100	18.69
Spanish	Our method	1000	25.87
Spanish	Our method	10000	**35.01**

The results for different Spanish training set sizes are shown in Fig 4. We observe that cross-lingual training improves the results over training only on the target language (to a much bigger extent when there is much less target language training data available). We further observe that cross-lingual learning leads to much more stable training which can be seen in the much smaller ranges of results as compared to the models trained only on the target language. Considering Fig 4(a) (b) (c) together, we find more source data of high-resource languages facilitates to stabilize the training process and results in faster convergence.

VI. RELATED WORK

A number of transfer learning approaches have been proposed for solving the data-lacking issue. Early approaches [7], [29] mainly rely on machine translation systems to translate either training or testing utterances, together with token-level alignment to align tags with tokens. Several works such as [9] employ cross-lingual word embeddings to utilize shared knowledge in word embeddings layer. [8] uses extra features derived from pre-trained machine translation model to help transfer common knowledge. They pre-trained several types of machine translation model between low-resource language and high-resource language, using its encoder to provide cross-lingual features when training target model.

Besides those model-driven approaches, transferring parameters and then fine-tuning has also proved to be useful. [26] breaks slot tagging model into several layers, and studies different impact when transferring parameters of different layers. In [30], the authors find that transferring parameters of character embeddings and character RNN is most useful, since many languages share the same alphabet, and may have similar affixes. [10] focuses on transfer learning from English to Japanese, proposing the method called romanization to help dissimilar languages share a common character embedding space. Other approaches include [31], which encodes slot description to vectors and employs an attention layer to obtain slot-aware representations of user input, and [20], which uses features derived from the source model.

VII. Conclusion and Future Work

In this paper, we introduce a novel multi-level cross-lingual model with language shared and specific knowledge for slot tagging to transfer shared linguistic features from high-resource languages to low-resource languages. Experiment results show that our method effectively alleviates issues of data scarcity and performs significantly better than state-of-the-art monolingual baseline by a large margin of 2.09% in absolute F1-score when training with the full dataset, even higher improvement of 12.21% in a zero-shot learning setting where no example of the target language is used. We provide extensive analysis of the results to shed light on future work. We plan to extend our method to more low-resource languages, especially for Chinese and Japanese with no similar alphabet as English. Besides, our method could also be applied to other cross-lingual NLP tasks, such as sentiment classification and dependency parsing.

Acknowledgment

We thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[2] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding." in *Interspeech*, 2013, pp. 3771–3775.

[3] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," *ArXiv*, vol. abs/1609.01454, 2016.

[4] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *HLT-NAACL*, 2016.

[5] N. T. Vu, "Sequential convolutional neural networks for slot filling in spoken language understanding," *Interspeech 2016*, Sep 2016. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-395

[6] E. A. Stepanov, I. Kashkarev, A. O. Bayer, G. Riccardi, and A. Ghosh, "Language style and domain adaptation for cross-language slu porting," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 144–149.

[7] X. He, L. Deng, D. Hakkani-Tur, and G. Tur, "Multi-style adaptive training for robust cross-lingual spoken language understanding," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8342–8346.

[8] S. Schuster, S. Gupta, R. Shah, and M. Lewis, "Cross-lingual transfer learning for multilingual task oriented dialog," 2018.

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

[10] A. Johnson, P. Karanasou, J. Gaspers, and D. Klakow, "Cross-lingual transfer learning for japanese named entity recognition," in *NAACL-HLT*, 2019.

[11] L. Huang, H. Ji, and J. May, "Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging," in *NAACL-HLT*, 2019.

[12] Y. Lin, S. Yang, V. Stoyanov, and H. Ji, "A multi-lingual multi-task architecture for low-resource sequence labeling," in *ACL*, 2018.

[13] J. D. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.

[14] L.-A. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *CoNLL*, 2009.

[15] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016. [Online]. Available: http://dx.doi.org/10.18653/v1/P16-1101

[16] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357–370, 2016.

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[18] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.

[19] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *ArXiv*, vol. abs/1606.04671, 2016.

[20] L. Chen and A. Moschitti, "Transfer learning for sequence labeling using source model and target data," *ArXiv*, vol. abs/1902.05309, 2019.

[21] S. Gu, Y. Feng, and Q. Liu, "Improving domain adaptation translation with domain invariant and specific information," *ArXiv*, vol. abs/1904.03879, 2019.

[22] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[23] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[24] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "Allennlp: A deep semantic natural language processing platform," *arXiv preprint arXiv:1803.07640*, 2018.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[26] J. Y. Lee, F. Dernoncourt, and P. Szolovits, "Transfer learning for named-entity recognition with neural networks," *arXiv preprint arXiv:1705.06273*, 2017.

[27] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018. [Online]. Available: http://dx.doi.org/10.18653/v1/N18-1202

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2018.

[29] F. García, L. F. Hurtado, E. Segarra, E. Sanchis, and G. Riccardi, "Combining multiple translation systems for spoken language understanding portability," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 194–198.

[30] Z. Yang, R. Salakhutdinov, and W. W. Cohen, "Transfer learning for sequence tagging with hierarchical recurrent networks," *arXiv preprint arXiv:1703.06345*, 2017.

[31] S. Lee and R. Jha, "Zero-shot adaptive transfer for conversational language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6642–6649.