XL-NBT: A Cross-lingual Neural Belief Tracking Framework

Wenhu Chen[1], **Jianshu Chen**[2], **Yu Su**[3], **Xin Wang**[1], **Dong Yu**[2]
Xifeng Yan[1] **and William Wang**[1]

University of California, Santa Barbara, CA, USA[1]
Tencent AI Lab, Bellevue, WA, USA[2]
The Ohio State University, Columbus, Ohio[3]
{wenhuchen,xwang,xyan,william}@cs.ucsb.edu, su.809@osu.edu
jianshuchen@tencent.com, dongyu@ieee.org

Abstract

Task-oriented dialog systems are becoming pervasive, and many companies heavily rely on them to complement human agents for customer service in call centers. With globalization, the need for providing cross-lingual customer support becomes more urgent than ever. However, cross-lingual support poses great challenges—it requires a large amount of additional annotated data from native speakers. In order to bypass the expensive human annotation and achieve the first step towards the ultimate goal of building a universal dialog system, we set out to build a cross-lingual state tracking framework. Specifically, we assume that there exists a source language with dialog belief tracking annotations while the target languages have no annotated dialog data of any form. Then, we pre-train a state tracker for the source language as a teacher, which is able to exploit easy-to-access parallel data. We then distill and transfer its own knowledge to the student state tracker in target languages. We specifically discuss two types of common parallel resources: bilingual corpus and bilingual dictionary, and design different transfer learning strategies accordingly. Experimentally, we successfully use English state tracker as the teacher to transfer its knowledge to both Italian and German trackers and achieve promising results.

1 Introduction

Over the past few years, we have witnessed the burgeoning of real-world applications of dialog systems, with many academic, industrial, and startup efforts racing to lead the widely-believed next-generation human-machine interfaces. As a result, numerous task-oriented dialog systems such as virtual assistants and customer conversation services were developed (Wen et al., 2015; Rojas-Barahona et al., 2017; Bordes and Weston,

2017; Williams et al., 2017; Li et al., 2017), with Google Duplexbeing the most recent example.

With the rapid process of globalization, more countries have observed growing populations of immigrants, and more companies have moved forward to develop their overseas business sectors. To provide better customer service and bring down the cost of labor at call centers, the development of universal dialog systems has become a practical issue. A straightforward strategy is to separately collect training data and train dialog systems for each language. However, it is not only tedious but also expensive. Two settings naturally arise for more efficient usage of the training data: (1) Multi-lingual setting: we annotate data for multiple languages and train a single model, with possible innovations on joint training. (2) Cross-lingual setting: we annotate data and train a model for only one (popular) language, and transfer the learned knowledge to other languages. Here we are interested in the second case, and the important research question we ask is: *How can we build cross-lingual dialog systems that can support less popular, low- or even zero-resource languages*?

As an initial step towards cross-lingual dialog systems, we focus on the cornerstone of dialog systems – *dialog state tracking* (DST), or *belief tracking*, a key component for understanding user inputs and updating *belief state*, i.e., a system's internal representation of the state of conversation (Young et al., 2010). Based on the perceived belief state, the dialog manager can decide which action to take, and what verbal response to generate (Precup and Teh, 2017; Bordes and Weston, 2017).

DST models require a considerable amount of annotated data for training (Henderson et al., 2014b; Mrksic et al., 2015, 2017). For a common dialog shown in Figure 1, a typical data acquisition process (Rojas-Barahona et al., 2017) not only requires two human users to converse for multiple

turns but also requires annotators to identify user's intention in each turn. Such two-step annotation is very expensive, especially for rare languages.

We study the novel problem of *cross-lingual DST*, where one leverages the annotated data of a source language to train DST for a target language with *zero* annotated data (Figure 1); no conversation dialog or dialog state annotation is available for the target language. In order to deal with this zero-resource challenging scenario, we first decouple the state-of-the-art neural belief tracker framework (Mrksic et al., 2017) into sub-modules, namely *utterance encoder*, *context gate*, and *slot-value decoder*. By introducing a teacher-student framework, we are able to transfer knowledge across languages module by module, following the divide-and-conquer philosophy. Requiring no target-side dialog data, our method relies on other easy-to-access parallel resources to understand the connection between languages. Depending on the popularity and availability of target language resources, we study two kinds of parallel data: bilingual corpus and bilingual dictionary, and we respectively design two transfer learning strategies.

We use the popular *Wizard-of-Oz* (Rojas-Barahona et al., 2017) dataset as our DST benchmark to evaluate the effectiveness of our cross-lingual transfer learning. We specify English as the source (primary) language and two different European languages (German and Italian) as our zero-annotation target languages. Compared with an array of alternative transfer learning strategies, our cross-lingual DST models consistently achieve promising results in both scenarios for both zero-annotation languages. To ensure reproducibility, we release our code, training data and parallel resources in the github[1]. Our main contributions are three-fold:

- Towards building cross-lingual dialog systems, we are the first to study the cross-lingual dialog state tracking problem.
- We systematically study different scenarios for this problem based on the availability of parallel data and propose novel transfer learning methods to tackle the problem.
- We empirically demonstrate the efficacy of the proposed methods, showing that our methods can accurately track dialog states for

languages with zero annotated data.

2 Related Work

2.1 Dialog State Tracking

Broadly speaking, the dialog belief tracking algorithms can be divided into three families: 1) hand-crafted rules 2) generative models, and 3) maximum-entropy model (Metallinou et al., 2013). Later on, many deep learning based discriminative models have surged to replace the traditional strategies (Henderson et al., 2014a; Mrksic et al., 2017; Williams et al., 2016) and achieved state-of-the-art results on various datasets. Though the discriminative models are reported to achieve fairly high accuracy, their applications are heavily restricted by the domain, ontology, and language. Recently, a pointer network based algorithm (Xu and Hu, 2018) and another multi-domain algorithm (Rastogi et al., 2017) have been proposed to break the ontology and domain boundary. Besides, (Mrkšić et al., 2017) has proposed an algorithm to train a unified framework to deal with multiple languages with annotated datasets. In contrast, our paper focuses on breaking the language boundary and transfer DST knowledge from one language into other zero-annotation languages.

2.2 Cross-Lingual Transfer Learning

Cross-lingual transfer learning has been a very popular topic during the years, which can be seen as a transductive process. In such process, the input domains of the source and target are different (Pan and Yang, 2010) since each language has its own distinct lexicon. By discovering the underlying connections between the source and target domain, we could design transfer algorithms for different tasks. Recently, algorithms have been successfully designed for POS tagging (Zhang et al., 2016; Kim et al., 2017), NER (Pan et al., 2017; Ni et al., 2017) as well as image captioning (Miyazaki and Shimizu, 2016). These methods first aim at discovering the relatedness between two languages and separate language-common modules from language-specific modules, then resort to external resources to transfer the knowledge across the language boundary. Our method addresses the transfer learning using a teacher-student framework and proposes to use the teacher to gradually guide the student to make more proper decisions.
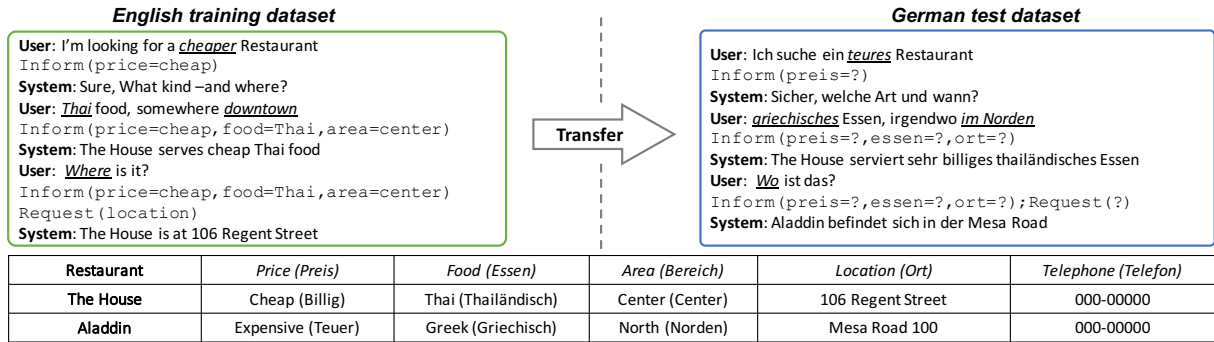
Figure 1: Cross-lingual transfer learning for dialog state tracking, where the underlying database (the table above) is shared across languages. The source language has annotated dialogs and the ground truth states, but the target language has neither dialogs nor ground truth states (only a testing dataset for evaluation).
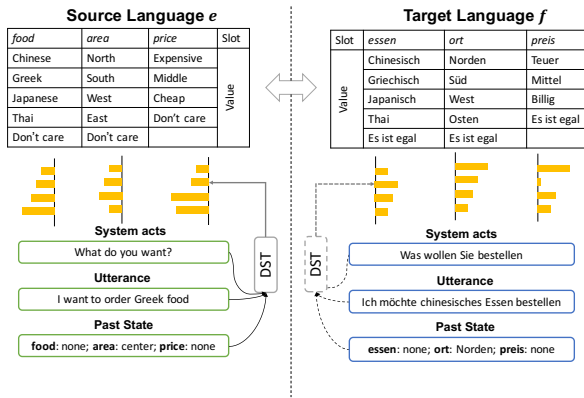
3 Problem Definition

Figure 2: Cross-lingual DST structure, the ontology and database between multiple languages are shared.

The dialog states are defined as a set of search constraints (i.e. informable slots or goals) that the user specified through the dialog and a set of attribute questions regarding the search results (i.e. requestable slots or requests). The objective of dialog state tracking (DST) is to predict and track the user intention (i.e., the values of the aforementioned slots) at each time step based on the current user utterance and the entire dialog history. As shown in Figure 2, for each slot, the DST computes an output distribution of the candidate values using three inputs: (i) *system response* a_t, which is the sentence generated by the system, (ii) *utterance* u_t, which is the sentence from the user, and (iii) *previous state*, which denotes the selected slot-value pairs. We define the ontology of the dialog system to be the set of all the possible words the dialog slot and value can take. In this paper, we are interested in learning a cross-lingual DST. Specifically, we assume that the DST for the *source* language has access to a human-annotated training dataset D while the DSTs for the *target*

languages do not have access to annotated data in other languages except for testing data. We here mainly consider two different types of parallel resources to assist the transfer learning:

(1) Bilingual Corpus, where abundant bilingual corpora exist between the source and the target languages. This is often the case for common language pairs like German, Italian, and French, etc.
(2) Bilingual Dictionary, where public bilingual dictionaries exist between the source and the target languages, but large-scaled parallel corpus are harder to obtain. This can be the case for rarer languages like Finnish, Bulgarian, etc.

Furthermore, we assume that all the languages share a common multi-lingual database, whose column/row names and entry values are stored via multiple languages (see the database in Figure 1). That is, the ontology of dialog among different languages is known with a one-to-one mapping between them (e.g., greek=griechisch=greco, food=essen=cibo). Based on that, we could construct a mapping function M to associate the ontology terms from different languages with predesigned language-agnostic concepts: for example, $M(foods) = M(Essen) = M(Cibo) =$ FOOD. We illustrate our problem definition in Figure 2.

4 Decoupled Neural Belief Tracker

We design our cross-lingual DST on top of the state-of-the-art Neural Belief Tracker (NBT) (Mrksic et al., 2017), which demonstrates many advantages (no hand-crafted lexicons, no linguistic knowledge required, etc). These nice properties are essential for our cross-lingual DST design because we are pursuing a general and simple framework regardless of the language properties. In short, NBT consists of a neural

network that computes the matching score for every candidate slot-value pair (c_s, c_v) based on the following three inputs: (i) the system dialog acts $a_t = (t_q, t_s, t_v)$,[2] (ii) the user utterance u_t, and (iii) the candidate slot-value pair. And it identifies the user intents by evaluating the scores for all the slot-value pairs (see Figure 3). With a slight abuse of notation, we still use $c_s, c_v, t_s, t_v, t_q \in \mathbb{R}^H$ to denote the vector representations of themselves, where H is the embedding dimension. We will use pre-trained embedding vectors in our cross-lingual NBT, just like the original NBT and they will be fixed during training. To enable cross-lingual transfer learning, we first re-interpret the architecture of the original NBT by decomposing it into three components:

Utterance Encoding The first component is an utterance encoder, which maps the utterance $u_t = \{w_1, w_2, \cdots, w_N\}$ of a particular language into a semantic representation vector $r(u_t) \in \mathbb{R}^H$, where $w_i \in \mathbb{R}^H$ is the word vector for the i-th token and N is the length of the utterance. Note that the dimension of the semantic vector $r(u_t)$ is the same as that of the word vector. We implement
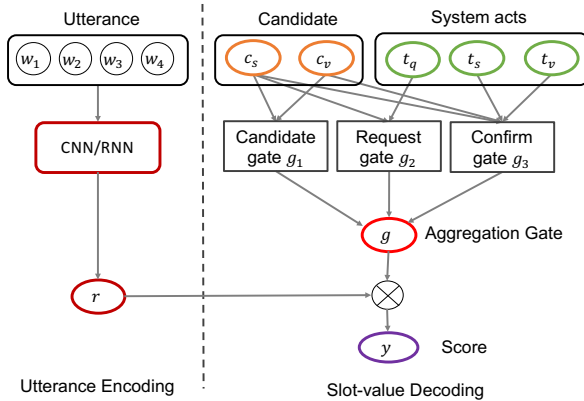

Figure 3: Our implementation of baseline NBT, slightly modified from (Mrksic et al., 2017).

the encoder using the same convolutional neural network (CNN) as the original NBT, with a slight modification of adding a top batch normalization layer. We will explain this change in section 5.

Context Gate The second part is the context gate, which takes the system acts $a_t = (t_q, t_s, t_v)$

---

[2] t_q represents the system request, t_s, t_v represents the system confirmation. If the system wants to request some information from the user by asking "what's your favorite area?", then NBT sets t_q="AREA". If the system wants to confirm some information from a user by asking "should I try Persian restaurants in the north?" then NBT sets t_s, t_v="area, north".

and the candidate slot-value pair (c_s, c_v) as its inputs and filter out the desired information from the encoded utterance. The context gate g is a sum of three separate gates:

$$g(c_s, c_v, a_t) = g_1 + g_2 + g_3 \quad (1)$$

where the individual gates are defined as:

$$
\begin{aligned}
g_1 &= \sigma(W_c^s(c_s + c_v) + b_c^s) \\
g_2 &= (c_s \cdot W_t^q t_q) \odot [1, \cdots, 1]^H \quad (2) \\
g_3 &= (c_s \cdot W_t^s t_s)(c_v \cdot W_t^v t_v) \odot [1, \cdots, 1]^H
\end{aligned}
$$

where $W_c^s, W_t^q, W_t^s, W_t^v \in \mathbb{R}^{H \times H}$ are the weight matrices, and $\odot$ and $\cdot$ denote the Hadamard product and the inner product, respectively. The three gates $g_1 \in \mathbb{R}^H, g_2 \in \mathbb{R}^H, g_3 \in \mathbb{R}^H$ model the relevance between the candidate slot and value, the system request and the system confirms, respectively. The transformation matrices W_t^q, W_t^s, W_t^v are added to the original NBT to increase the model flexibility of the gates.

Slot-Value Decoding The final component is a slot-value decoder, which predicts the score y of a given slot-value pair using the filtered information from the utterance representation r as:

$$y(c_s, c_v, u_t, a_t) = W_y^T [r(u_t) \odot g(c_s, c_v, a_t)] \quad (3)$$

where $W_y \in \mathbb{R}^{H \times 1}$ is the weight vector. The above expression computes the score for the slot-value pair based on the information from the current turn. We combine it with the information from previous turns to get the final score:

$$
\begin{aligned}
\hat{y}(c_v | u_t, a_t, c_s) = &\lambda y(c_s, c_v, u_t, a_t) + \\
&(1 - \lambda)\hat{y}(c_s, c_v, u_{t-1}, a_{t-1})
\end{aligned} \quad (4)
$$

here λ is a combination weight. For each given slot c_s, NBT selects the single highest value for informable slots and selects all values above a certain threshold for request slots. Here we replace the multi-layer perceptron in the orginal NBT by a linear output layer (to be explained in section 5).

5 Cross-lingual Neural Belief Tracker

In this section, we develop a cross-lingual Neural Belief Tracker (XL-NBT) that distills knowledge from one NBT to another using a teacher-student framework. We assume the ontology mapping M is known a priori (see Figure 3). XL-NBT uses language-specific utterance encoder and context gate for each input language while sharing a common (language-agnostic) slot-value decoder across different languages (see Figure 3).

417

The key idea is to optimize the language-specific components of the student network (NBT of the target language) so that their outputs are language-agnostic. This is achieved by making these outputs close to that of the teacher network (NBT of the source language), as we detail below.

5.1 Teacher-Student Framework

We are given a well-trained NBT for a source language e, and we want to learn an NBT for a target language f without any annotated training data. Therefore, we cannot learn the target-side NBT from standard supervised learning. Instead, we use a teacher-student framework to distill the knowledge from the source-side NBT (teacher network) into the target-side NBT (student network) (see Figure 4). Let $x_e \triangleq (c_s^e, c_v^e, u_t^e, a_t^e)$ be the input to the teacher network and let $x_f \triangleq (c_s^f, c_v^f, u_t^f, a_t^f)$ be the associated input to the student network. The standard teacher-student framework trains the student network by minimizing

$$\hat{J}_1 = \sum_{x_e, x_f} ||y(c_s^e, c_v^e, u_t^e, a_t^e) - y(c_s^f, c_v^f, u_t^f, a_t^f)||^2 \quad (5)$$

where $y(c_s^e, c_v^e, u_t^e, a_t^e)$ and $y(c_s^f, c_v^f, u_t^f, a_t^f)$ denote the scores by the teacher and the student networks, respectively, and the slot-value pairs satisfy $M(c_v^f) = M(c_v^e)$ and $M(c_s^f) = M(c_s^e)$. However, the target-side inputs $(c_s^f, c_v^f, u_t^f, a_t^f)$ parallel to $(c_s^e, c_v^e, u_t^e, a_t^e)$ are usually not available in cross-lingual DST, and, even worse, the target-side utterance u_t^e is not available. We may have to generate synthetic input data for the student network or leverage external data sources. It is relatively easy to use the mapping $M(\cdot)$ to generate $(c_s^f, c_v^f, a_t^f))$ (i.e., the inputs of the target-side context gate) from the (c_s^e, c_v^e, a_t^e). But it is more challenging to obtain the parallel utterance data u_t^f from u_t^e). Therefore, we have to leverage external bilingual data sources to alleviate the problem. However, the external bilingual data are usually not in the same domain as the utterance, and hence they are not aligned with the slot-value pair and system acts (i.e., (c_s^e, c_v^e, a_t^e) or (c_s^f, c_v^f, a_t^f)). For this reason, we cannot perform the knowledge transfer by minimizing the cost (5). Instead, we need to develop a new cost function where the utterance is not required to be aligned with the slot-value pair and the system acts. To this end, let $g^e = g_e(c_s^e, c_v^e, a_t^e)$ and $g^f = g_f(c_s^f, c_v^f, a_t^f)$. And we substitute (3)

into (5) and get:

$$\hat{J}_1 \leq ||W_y||^2 \sum_{c_v^f, c_v^e} ||r_e \odot g^e - r_f \odot g^f||^2$$

$$= ||W_y||^2 \sum_{c_v^f, c_v^e} ||g^e \odot (r_e - r_f) + r_f \odot (g^e - g^f)||^2$$

$$\leq ||W_y||^2 \sum_{c_v^f, c_v^e} ||g^e||^2 ||r_e - r_f||^2 + ||r_f||^2 ||g^e - g^f||^2$$

where $r_e = r_e(u_t^e)$ and $r_f = r_f(u_t^f)$. As we mentioned earlier, the weight W_y in the slot-value decoder is shared between the student and the teacher networks and will not be updated. The teacher-student optimization only adjusts the weights related to the language-specific parts in Figure 3 (i.e., utterance encoding and context gating). Therefore, the shared weight $||W_y||$ is seen as a constant. Furthermore, $\sum_{c_v^f, c_v^e} ||g^e||^2$ can be seen as a constant since the teacher gate is fixed. Since we use batch normalization layer to normalize the encoder output (described in Figure 3), $||r_f(u_t^f)||^2$ can also be treated as a constant C_2. Therefore, we formally write the upper bound of $\hat{J}_1$ as our surrogate cost function J:

$$J = C_1 ||r_e(u_t^e) - r_f(u_t^f)||^2 + C_2 \sum_{c_v^f, c_v^e} ||g^e - g^f||^2 \quad (6)$$

The surrogate cost has successfully decoupled utterance encoder with context gate, and we use J_r and J_g to measure the encoder matching cost and the gate matching cost, respectively.

$$J_r = ||r_e(u_t^e) - r_f(u_t^f)||^2$$
$$J_g = \sum_{c_v^f, c_v^e} ||g^e - g^f||^2 \quad (7)$$

The encoder cost J_r is optimized to distill the knowledge from the teacher encoder to student encoder while gate cost J_g is optimized to distill the knowledge from teacher gate to student gate. This objective function successfully *decouples* the optimization of encoder and gate, thus we are able to optimize J_r and J_g separately from different data sources. Recall that we can easily simulate the target-side system acts, slot-value pairs (c_s^f, c_v^f, a^f) by using the ontology mapping M. Therefore, optimizing J_g is relatively easy. Formally, we write the gate matching cost as follows:

$$J_g = \sum_{\substack{a_t^e, c_s^e, c_v^e \\ a_t^f, c_s^f, c_v^f}} ||g_e(c_s^e, c_v^e, a_t^e) - g_f(c_s^f, c_v^f, a_t^f)||^2 \quad (8)$$
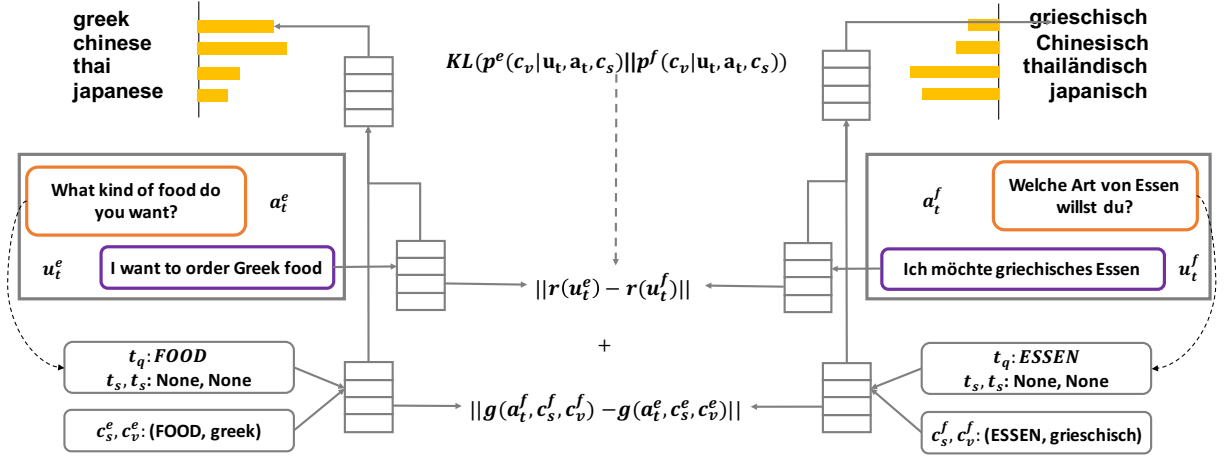
418

Figure 4: Teacher-Student Framework for cross-lingual transfer learning. The dotted line denotes the imaginary utterances, which expresses the same intention as the source side.
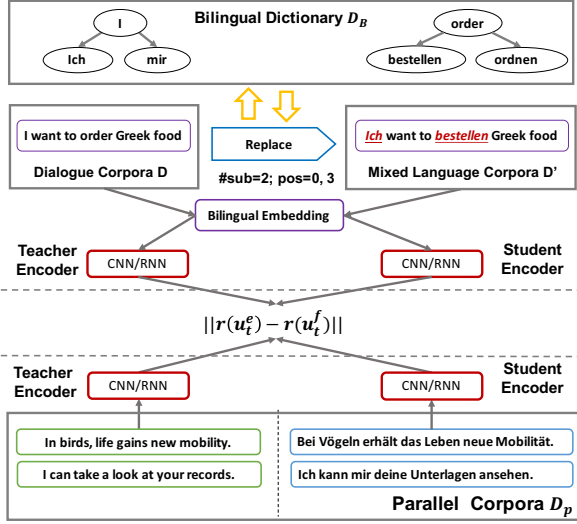

Figure 5: XL-NBT-C and XL-NBT-D for two scenarios

However, exact optimization of J_r is difficult and we have to approximate it using external parallel data. We consider two kinds of external resources (bilingual corpus and bilingual dictionary) in the sections 5.2-5.3 (see Figure 5 for the main idea).

5.2 Bilingual Corpus (XL-NBT-C)

In our first scenario, we assume there exists a parallel corpus D_p consisting of sentence pairs from the source language and the target language. In this case, the cost function (6) is approximated by

$$J = \mathbb{E}_{(m_e, m_f) \in D_p} ||r_e(m_e) - r_f(m_f)||^2 + \alpha J_g \quad (9)$$

where α is the balancing factor and J_g is defined in (6). The cost function (9) is minimized by stochastic gradient descent. At test time, we switch the encoder to receive target language inputs.

5.3 Bilingual Dictionary (XL-NBT-D)

In the second scenario, we assume there exists no parallel corpus but a bilingual dictionary D_B that defines the correspondence between source words and target words (a one-to-many mapping $\{w : M_D(w)\}$). Likewise, it is infeasible to optimize the exact encoder cost J_r due to the lack of target-side utterances. We propose a word replacement strategy (to be described later) to generate synthetic parallel sentence $\hat{u}_t^f$ of "mixed" language. Then, we use the generated target parallel sentences to approximate the cost (6) by

$$J_r = \mathbb{E}_{u_t \in D} ||r_e(u_t^e) - r_f(\hat{u}_t^f)||^2 + \alpha J_g \quad (10)$$

where α is the balancing factor. For word replacement, we first decide the number of words N_w to be replaced, then we draw N_w positions randomly from the source utterance and substitute the corresponding word w_i with their target word synonym from $M_D(w)$ based on the context as follows:

$$jp(N_w = i) = \frac{\exp(-i/\tau)}{\sum_{i' < N} \exp(-i'/\tau)}$$
$$p(\hat{w}) = \frac{\hat{w} \cdot h_{\hat{w}}}{\sum_{w' \in M(w_i)} w' \cdot h_{\hat{w}}} \quad (11)$$

where $h_{\hat{w}} = \sum_{k=-2:k\neq0}^{2} w_{i+k}$ represents the context vector and N denotes the utterance length. The context similarity of context and the target-side synonym can better help us in choosing the most appropriate candidate from the list. In our following experiments, we adjust the temperature of τ to control the aggressiveness of replacement.

6 Experiments

6.1 Dataset

The Wizard of Oz (WOZ) (Rojas-Barahona et al., 2017) dataset is used for training and evaluation, which consists of user conversations with task-oriented dialog systems designed to help users find suitable restaurants around Cambridge, UK. The corpus contains three informable (i.e. goal-tracking) slots: FOOD, AREA, and PRICE. The users can specify values for these slots in order to find which best meet their criteria. Once the system suggests a restaurant, the users can ask about the values of up to eight requestable slots (PHONE NUMBER, ADDRESS, etc.). Multilingual WOZ 2.0 (Mrksic et al., 2017) has expanded this dataset to include more dialogs and more languages. The train, valid and test datasets for three different languages (English, German, Italian) are available online[3]. We use the English as source language where 600 dialogs are used for training, 200 for validation and 400 for testing. We use the German and Italian as the target language to transfer our knowledge from English DST system. In the experiments, we do not have access to any training or validation dataset for German and Italian, and we only have access to their testing dataset which is composed of 400 dialogs.

For external resource, we use the IWSLT2014 Ted Talk parallel corpus (Mauro et al., 2012) from the official website[4] for bilingual corpus scenario. In the IWSLT2014 parallel corpus, we only keep the sentences between 4 and 40 words and decrease the sentence pairs to around 150K. We use Panlex (Kamholz et al., 2014) as our data source and crawl translations for all the words appearing in the dialog datasets to build our bilingual dictionary. We specifically investigate two kinds of pre-trained embedding, and we use Glove (Pennington et al., 2014) as the monolingual embedding and MUSE (Conneau et al., 2017) as the bilingual embedding to see their impacts on the DST performance.

We split the raw DST corpus into turn-level examples. During training, we use the ground truth previous state V_{t-1} as inputs. At test time, we use the model searched states as the previous state to continue tracking intention until the end of the

dialog. When the dialog terminates, we use two evaluation metrics introduced in Henderson et al. (2014a) to evaluate the DST performance: (1) *Goals*: the proportion of dialog turns where all the users search goal constraints were correctly identified. (2) *Requests*: similarly, the proportion of dialog turns where users requests for information were identified correctly. Our implementation is based on the NBT[5], the details of our system setting are described in the appendix.

6.2 Results

Here we highlight the baselines we use to compare with our cross-lingual algorithm as follows:
(1) Supervised: this baseline algorithm assumes the existence of annotated dialog belief tracking datasets, and it determines the upper bound of the DST model.
(2) w/o Transfer: this algorithm trains an English NBT, and then directly feeds target language into the embedding level as inputs during test time to evaluate the performance.
(3) Ontology-match: this algorithm directly uses exact string matching against the utterance to discover the perceived slot-value pairs, it directly assigns a high score to the appearing candidates.
(4) Translation-based: this system pre-trains a translator on the external bilingual corpus and then translates the English dialog and ontology into target language as "annotated" data, which is used to train the NBT in the target language domain (more details about the implementation, performance and examples are listed in the appendix).
(5) Word-By-Word (WBW): this system transforms the English dialog corpus into target language word by word using the bilingual dictionary, which is used to train the NBT in target side.
We demonstrate the results for our proposed algorithms and other competing algorithms in Table 2, from which we can easily conclude that that (i) our Decoupled NBT does not affect the performance, and (ii) our cross-lingual NBT framework is able to achieve significantly better accuracy for both languages in both parallel-resource scenarios.

Compare with Translator/WBW. With bilingual corpus, XL-NBT-C with pre-trained bilingual embedding can significantly outperform our Translator baseline (Klein et al., 2017). This is intuitive because the translation model requires

---

Error Type	Examples
Modify Failure	Machine: I have two options that fit that description, golden wok Chinese restaurant and the Nirala which serves Indian food, do you have a preference? User: How about Nirala, whats the address and phone of that? Previous State: food=Chinese; Prediction: food=none; Groundtruth: food=Indian
Maintain Failure	Machine: there are $num places with a moderate price range. can you please tell me what kind of food you would like? User: well I want to eat in the north, whats up that way? Previous State: food=expensive; Prediction: food=none; Groundtruth: food=expensive
History Failure	Machine: Anatolia is located at $num bridge street city center. User: thank you goodbye! Previous State: food=Chinese; Prediction: food=Chinese;,Groundtruth: food=Turkish

Table 1: Here we show the frequent error types, the examples are translated to English for better understanding.

Language		German (student)		Italian (student)		English (teacher)	
Models		Goal	Request	Goal	Request	Goal	Request
w/ Supervised Dialog	NBT (Mrksic et al., 2017)	-	-	-	-	**0.84**	**0.91**
	Decoupled NBT (mono)	0.79	0.83	0.86	0.91	0.82	0.89
	Decoupled NBT (bilingual)	**0.80**	**0.84**	**0.88**	**0.91**	**0.84**	0.90
w/o Bilingual Data	w/o Transfer (mono)	0.15	0.10	0.15	0.11	-	-
	w/o Transfer (bilingual)	0.13	0.13	0.11	0.12	-	-
	Ontology Matching	0.24	0.21	0.23	0.21	-	-
w/ Bilingual Corpus	Translate (Klein et al., 2017)	0.41	0.42	0.48	0.51	-	-
	XL-NBT-C (mono)	0.48	0.54	0.65	0.60	-	-
	XL-NBT-C (bilingual)	**0.55**	**0.59**	**0.72**	**0.69**	-	-
w/ Bilingual Dictionary	Word-by-Word	0.22	0.25	0.25	0.27	-	-
	XL-NBT-D (mono)	0.14	0.15	0.23	0.22	-	-
	XL-NBT-D (bilingual)	**0.51**	**0.56**	**0.73**	**0.63**	-	-

Table 2: Experimental results for cross-lingual NBT and other baseline algorithms. All results are averaged over 5 runs. Here we use "mono" to refer to the experiments with pre-trained monolingual embedding, "bilingual" to refer to the experiments with pre-trained bilingual embedding.

both source-side encoding and target-side word-by-word decoding, while our XL-NBT only needs a bilingual source-encoding to align two vector space, which averts the compounded decoding errors. With the bilingual dictionary, the word-by-word translator is very weak and leading to many broken target sentences, which poses challenges for DST training. In comparison, our XL-NBT-D can control the replacement by adjusting its temperature to maintain the stability of utterance representation. Furthermore, for both cases, our teacher-student framework can make use of the knowledge learned in source-side NBT to assist its decision making, while translator-based methods learn from scratch.

Bilingual Corpus vs. Bilingual Dictionary. From the table, we can easily observe that bilingual corpus is obviously a more informative parallel resource to perform cross-lingual transfer learning. The accuracy of XL-NBT-D is lower than XL-NBT-C. We conjecture that our replace-ment strategy to generate "mixed" language utterance can sometimes break the semantic coherence and cause additional noises during the transfer process, which remarkably degrades the DST performance.

Monolingual vs. Bilingual embedding. From the table, we can observe that the bilingual embedding and monolingual embedding does not make much difference in supervised training. However, the gap in the bilingual corpus case is quite obvious. Monolingual embedding even causes the transfer to fail in a bilingual dictionary case. We conjecture that the bilingual word embedding already contain many alignment information between two languages, which largely eases the training of encoder matching objective.

German vs. Italian As can be seen, the transfer learning results for Italian are remarkably higher than German, especially for the "Goal" accuracy. We conjecture that it is due to German declension, which can produce many word forms. The

very diverse word forms present great challenges for DST to understand its intention behind. Especially for the bilingual dictionary, German tends to have much longer replacement candidate lists than Italian, which introduces more noises to the replacement procedure.

Error Analysis Here we showcase the most frequent error types in subsection 6.1. From our observation, these three types of errors distribute evenly in the test dialogs. The error mainly comes from the unaligned utterance space, which leads to failure in understanding the intention of human utterance in the target language. This can lead the system to fail in modifying the dialog state or maintaining the previous dialog states.

6.3   Discussion

Here we want to further highlight the comparison between our transfer learning algorithm with the MT-based approach. Though our approach outperforms the standard Translator trained on IWSLT-2014, it does not necessarily claim that our transfer algorithm outperforms any translation methods on any parallel corpus. In our further ablation studies, we found that using Google Translator [6] can actually achieve a better score than our transfer algorithm, which is understandable considering the complexity of Google Translator and the much larger parallel corpus it leverages. By leveraging more close-to-domain corpus and comprehensive entity recognition/replacement strategy, the translator model is able to achieve a higher score. Apparently, we need to trade off the efficiency for the accuracy. For DST problem, it is an overkill to introduce a more complex translation algorithm, what we pursue is a simple yet efficient algorithm to achieve promising scores. It is also worth mentioning that our XL-NBT algorithm only takes several hours to achieve the reported score, while the translator model takes much more time and memory to train depending on the complexity. Thus, the simplicity and efficiency makes our model a better fit for rare-language and limited-budget scenarios.

6.4   Ablation Test

Here we investigate the effect' of hyper-parameter α, τ on the evaluation results. The α is used to balance the optimization of encoder constraint and gate constraint, where larger α means more optimization on gate constraint. The temperature τ is used to control the aggressiveness of the replacement XL-NBT-D, where smaller τ means more source words are replaced by target synonyms. From the table Table 3, we can observe that the

α ablation (τ fixed to 0.1)			τ ablation (α fixed to 1)		
value	Goal	Request	value	Goal	Request
α=0	0.13	0.00	τ=0	0.14	0.08
α=0.1	0.46	0.54	τ=0.03	0.43	0.50
α=1	**0.51**	**0.56**	τ=0.1	**0.51**	**0.56**
α=5	0.46	0.54	τ=0.3	0.47	0.51
α=10	0.46	0.52	τ=1	0.44	0.52
α=100	0.44	0.50	τ=10	0.33	0.32

Table 3: Ablation test for hyper-parameter α and τ on English-to-German XL-NBT-D.

experimental results are not very sensitive to α, a dramatic change of α will not harm the final results too much, we simply choose $\alpha = 1$ as the hyper-parameter. In contrast, the system is more sensitive to temperature. Too conservative replacement will lead to weak transfer, while too aggressive replacement will destroy the utterance representation. Therefore, we choose the a moderate temperature of $\tau = 0.1$ throughout our experiments. We also draw the learning curve (Precision vs. Iteration) in the Appendix for both XL-NBT-C and XL-NBT-D. The learning curves show that our algorithm is stable and converges quickly, and the reported results are highly reproducible.

7   Conclusion

In our paper, we propose a novel teacher-student framework to perform cross-lingual transfer learning for DST. The key idea of our model is to decouple the current DST neural network into two separate modules and transfer them separately. We believe our method can be further extended into a general purpose multi-lingual transfer framework to resolve other NLP matching or classification problems.

8   Acknowledgement

---

[6]https://translate.google.com/

References

Antoine Bordes and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014a. Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 360–365. IEEE.

Matthew Henderson, Blaise Thomson, and Steve J. Young. 2014b. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA*, pages 292–299.

David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *LREC*, pages 3145–3150.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 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*, pages 2832–2838.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Çelikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 733–743.

Cettolo Mauro, Girardi Christian, and Federico Marcello. 2012. Wit3: Web inventory of transcribed and translated talks. In *Conference of European Association for Machine Translation*, pages 261–268.

Angeliki Metallinou, Dan Bohus, and Jason Williams. 2013. Discriminative state tracking for spoken dialog systems. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 466–475.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1780–1790.

Nikola Mrksic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Pei-hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2015. Multi-domain dialog state tracking using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 794–799.

Nikola Mrksic, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1777–1788.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *arXiv preprint arXiv:1706.00374*.

Jian Ni, Georgiana Dinu, and Radu Florian. 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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1470–1480.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 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, pages 1946–1958.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Doina Precup and Yee Whye Teh, editors. 2017. *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*. PMLR.

Abhinav Rastogi, Dilek Hakkani-Tür, and Larry P. Heck. 2017. Scalable multi-domain dialogue state tracking. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*, pages 561–568.

Lina Maria Rojas-Barahona, Milica Gasic, Nikola Mrksic, Pei-Hao Su, Stefan Ultes, Tsung-Hsien Wen, Steve J. Young, and David Vandyke. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 438–449.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1711–1721.

Jason Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 665–677.

Puyang Xu and Qi Hu. 2018. An end-to-end approach for handling unknown slot values in dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1448–1457.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi S. Jaakkola. 2016. Ten pairs to tag - multilingual POS tagging via coarse mapping between embeddings. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1307–1317.