Knowledge-Based Systems 259 (2023) 110015

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Zero-shot language extension for dialogue state tracking via pre-trained models and multi-auxiliary-tasks fine-tuning

Lu Xiang [a,b], Yang Zhao [a,b], Junnan Zhu [a,b], Yu Zhou [a,b,c,*], Chengqing Zong [a,b]

[a] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[b] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[c] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China

ARTICLE INFO

Article history:
Received 10 January 2022
Received in revised form 10 October 2022
Accepted 10 October 2022
Available online 17 October 2022

Keywords:
Dialogue state tracking
Zero-shot language extension
Multilingual DST
Pre-trained models
Multi-auxiliary-tasks fine-tuning

ABSTRACT

Dialogue state tracking (DST), a crucial component of the task-oriented dialogue system (TOD), is designed to track the user's goal. Existing DST models mainly focus on monolingual dialogue input, failing to meet the growing needs of a TOD to provide multilingual services. Therefore, this paper proposes a novel Zero-shot Language Extension scenario for DST, extending the monolingual DST to multilingual DST without extra high-cost dialogue data annotation. In this scenario, the multilingual DST only needs a single shared model to handle multilingual input and generate a unified dialogue state. This setting makes deploying a complete multilingual TOD easy since it could be reused by the downstream components from existing monolingual TOD. Specifically, we achieve the language extension by multi-auxiliary-tasks fine-tuning of multilingual pre-trained models, where five relevant auxiliary tasks are jointly designed, including monolingual DST, cross-lingual DST, forward word translation, utterance recovery, and semantic similarity. The extended multilingual DST model can be enhanced through joint optimization with all the auxiliary tasks by capturing multilingual context understanding and cross-lingual alignment characteristics. Comprehensive experiments on Multilingual WOZ dataset (English → German and English → Italian) and cross-lingual MultiWOZ dataset (English → Chinese and Chinese → English) demonstrate the effectiveness and superiority of the proposed method.

1. Introduction

Task-oriented dialogue systems (TOD) are designed to make some services, such as ticket booking and restaurant reservation, more convenient through interaction with users [1–6]. In the mainstream paradigm of TOD, dialogue state tracking (DST) is the first and key component, which can understand and track the belief of a user's goal through a dialogue history. The quality of DST heavily affects other downstream components like knowledge base retrieval, action selection, and response generation [4,5].

Mainly DST is monolingual. However, with the acceleration of globalization, there is an ever-growing demand for multilingual dialogue services. For example, a ticket booking dialogue in the international airport should simultaneously handle utterances in different countries. To satisfy the multilingual demand, various methods have been proposed and can be divided into the following two categories:

(1) Multilingual data collection. In this method, we first need to collect the multilingual DST data and then train the DST model for each language individually. However, it is quite costly and resource-intensive to collect high-quality dialogue data, especially for low-resource languages.

(2) Cross-lingual transfer. Given the DST training data for a high-resource language, cross-lingual transfer methods [7–11] can learn a new monolingual DST model by transferring the DST knowledge of a high-resource language (source language) into that of a low-resource language (target language), as shown in Fig. 1(a). Meanwhile, to bridge the gap between source and target languages, a machine translation engine or a multilingual pre-trained model is necessary for these methods. Remarkable progress has been made, while these methods still face the following deployment challenge: each language has to maintain its own DST model, which will raise the deployment difficulty with the increasing of languages.

Different from the existing methods, we focus on a more challenging multilingual DST scenario, namely **Zero-shot Language Extension Scenario**, in which the DST model can generate a unified dialogue state when the input utterances are in different languages, as illustrated in Fig. 1(b). We believe that it is worth exploring since it has the following advantages:

* Corresponding author at: National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China.
E-mail addresses: lu.xiang@nlpr.ia.ac.cn (L. Xiang), yang.zhao@nlpr.ia.ac.cn (Y. Zhao), junnan.zhu@nlpr.ia.ac.cn (J. Zhu), yzhou@nlpr.ia.ac.cn (Y. Zhou), cqzong@nlpr.ia.ac.cn (C. Zong).

https://doi.org/10.1016/j.knosys.2022.110015
0950-7051/

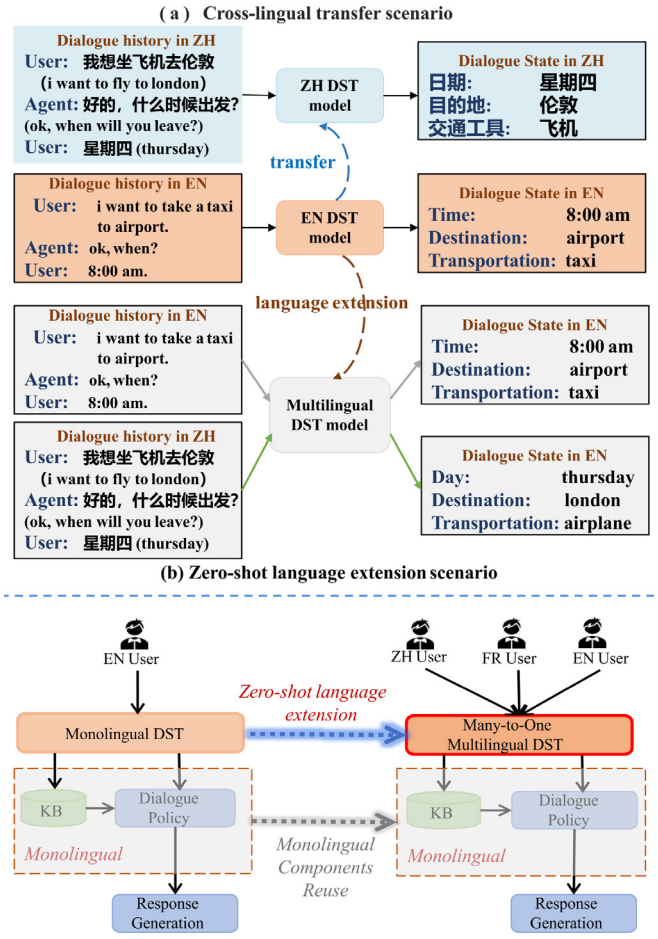


Fig. 1. A comparison of Cross-lingual transfer scenario (a) and Zero-shot language extension scenario (b).

- **Single model setting.** Different from the cross-lingual transfer, multilingual DST only needs a single shared model to handle different language inputs. So the deployment resources and difficulty can be reduced sharply.
- **Unified output state setting.** We hope that the model could generate a unified dialogue state when the input utterances are in different languages. Here the unified dialogue state is depicted in the source language. As shown in Fig. 1(b), when the inputs are English or Chinese, both the output dialogue states are in English. As the first component of the TOD, if the DST model can handle M languages (*e.g.*, French, Chinese, etc.) and generate a unified dialogue state (*e.g.*, English), some downstream components such as knowledge base retrieval and action selection can be reused without extra annotation and training, making the extension of TOD to other languages much more effortless.
- **Zero-shot setting.** Since collecting high-quality dialogue data for each language is unpractical, we focus on extending the DST model to a target language without any target language annotated dialogue training data.

Recent progress in multilingual pre-trained models enables many NLP applications for other languages [12–15]. To achieve the zero-shot language extension, it becomes natural to utilize the multilingual pre-trained models. However, the cross-lingual contextualized representations of multilingual pre-trained models are inconsistent, limiting the performance of zero-shot cross-lingual transfer [16]. In addition, the pre-trained models usually need to be fine-tuned on the corresponding task to achieve the capability. However, in our scenario, the DST training data of the new language is zero-shot, making the standard pre-trained models+fine-tuning paradigm unsuitable.

Therefore, to address the above problems, we propose a **pre-trained models+multi-auxiliary-tasks fine-tuning** method for the zero-shot language extension scenario. Our main idea is to design various relevant auxiliary tasks and fine-tune the multilingual pre-trained model with these auxiliary tasks to realize the multilingual DST. Concretely, we design five auxiliary tasks from DST ability and cross-lingual alignment perspectives, namely **monolingual DST** task, **cross-lingual DST** task, **forward word translation** task, **utterance recovery** task, and **semantic similarity** task. Specially, as shown in Fig. 2, (1) the monolingual DST task aims to generate the dialogue states in the source language given the dialogue history context of the source language. Similarly, (2) the cross-lingual DST task is to leverage the code-switching dialogue context, which is generated by replacing some words in the original dialogue context to the words in the target language through an automatically extracted dictionary, to generate the dialogue states in the source language. Besides the above two tasks, (3) forward word translation task is to predict the replaced words in the source language dialogue context, and (4) utterance recovery task is to recover the code-switching dialogue context into the source language dialogue context. The goal of the two tasks is to enhance the cross-lingual representations across different languages at the word-level. Moreover, (5) we design the semantic similarity task to make the hidden representations of code-switching and original dialogue history undistinguishable, guiding the encoder to learn language-independent hidden representations at sentence-level. By learning with the

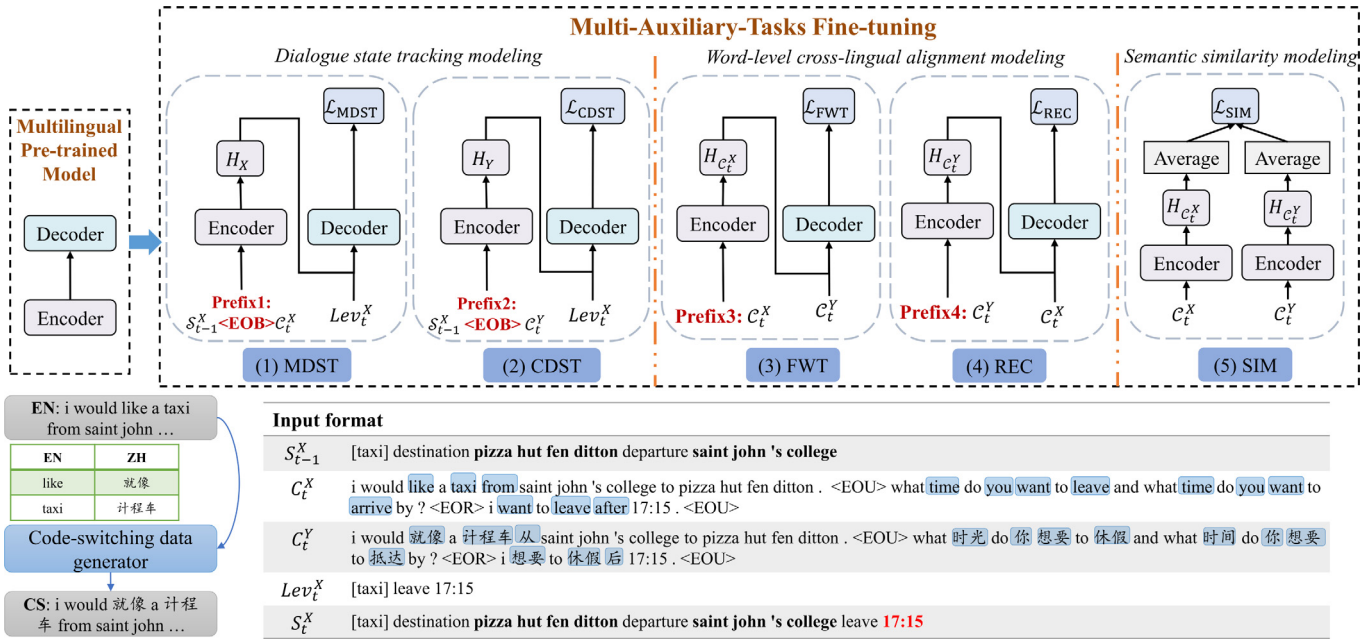


Fig. 2. Overview of *pre-trained models+multi-auxiliary-tasks fine-tuning* method. The encoder and decoder of auxiliary tasks are shared. *Prefix1* to *Prefix4* are task-specific prefixes added to the input sequence. The S^X_{t-1}, C^X_t, and C^Y_t denote the previous dialogue state in source language X, the history dialogue context in X, and the generated code-switching dialogue context in target language Y, respectively. It leverages the *multi-auxiliary tasks* to fine-tune the pre-trained model for multilingual DST.

joint objectives of all these auxiliary tasks, the model is enhanced to understand the dialogue context across languages, thus improving the multilingual DST's performance.

We conduct extensive experiments in zero-shot language extension scenario on Multilingual WOZ 2.0 dataset [17] and the cross-lingual MultiWOZ dataset [18] using two multilingual pre-trained models. Our proposed five auxiliary tasks lead to an impressive performance on the zero-shot extended target languages, which demonstrates the effectiveness of our proposed method. Furthermore, a surprising finding can be made from the experiments that the proposed method has an ability of **back transfer**, in which after expanding a new language, the performance of the new language could be sharply improved, the source language is also increased. In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to explore the zero-shot language extension for multilingual DST, which can utilize a single DST model to handle multilingual input utterances and generate a unified dialogue state.
- We propose a **pre-trained models+multi-auxiliary-tasks fine-tuning** method for the zero-shot language extension scenario, in which multi-auxiliary tasks are designed and then used to fine-tune the multilingual pre-trained models.
- Extensive experiments on two multilingual dialogue datasets (including four language directions: English → German, English → Italian, English → Chinese, and Chinese → English) with two different multilingual pre-trained models demonstrate that the proposed method consistently achieves substantial improvements on the target languages and maintains or even improves the performance of the source language.

This paper is organized as follows. Section 2 reviews related researches, including dialogue state tracking, cross-lingual dialogue state tracking, and multilingual pre-trained models. Section 3 introduces the monolingual DST model and formally defines the problem to be solved. In Section 4, we introduce the methodology proposed in the paper. Section 5 conducts the experiments and analyzes the results. Section 6 summarizes the work of this paper.

2. Related work

Dialogue state tracking. Dialogue state tracking (DST) is a hot research topic in task-oriented dialogue systems (TOD) [19–21]. Typical monolingual DST models can be divided into classification-based and generation-based methods, where the main difference is how the slot values are inferred. The former category [19,21–23] uses predefined dialogue ontology and simplifies the DST modeling into a classification problem. In contrast, generation-based methods treat dialogue state tracking as a generation task and directly generate slot values sequentially [5,24–26]. Since the generation-based methods have the potential to handle unseen values, much attention has been attracted from researchers. Recently, generation-based models that are built on large-scale pre-trained language models have achieved promising results on MultiWOZ 2.0 and 2.1 [4,5,27–29]. In this study, we follow the generation-based DST method using a pre-trained sequence-to-sequence model as our kick-off to realize the language extension of DST model.

Cross-lingual dialogue state tracking. The study of cross-lingual dialogue systems has gained much attention, and it studies how to adapt a dialogue system into the target language. The current researches mainly concentrate on the following two directions: cross-lingual natural language understanding [8–10,17,19,30] and cross-lingual dialogue state tracking [7–9,11,31]. As stated before, monolingual DST has been explored extensively, but there is limited work for a multilingual scenario.

Multilingual WOZ 2.0 dataset [17] is a popular cross-lingual DST benchmark, where a DST model is trained only using English data and evaluated directly for German and Italian. Most of the existing cross-lingual DST studies are evaluated on this benchmark. Chen et al. [7] propose to use a teacher–student network to perform cross-lingual transfer learning for DST. The

teacher model transfers its own knowledge to the student model of the target language by using bilingual corpus and bilingual dictionary. Lin et al. [8] propose an attention-informed mixed-language training method, which leverages the code-switching data to build cross-lingual DST. The code-switching sentences are generated by replacing the selected words that receive the highest attention score with their translations in an existing bilingual dictionary. Another code-switching method [9] focuses on generating multilingual code-switching data dynamically for better fine-tuning, where the randomly selected words are replaced with their translations in different languages. Lin and Chen [11] explore the transferability of a cross-lingual generative DST using a multilingual pre-trained model. Besides, one recent study [31] uses parallel conversational data for cross-lingual intermediate fine-tuning of multilingual pre-trained models. It facilitates the performance in the zero-shot cross-lingual transfer of the DST task, while our work addresses the problem of zero-shot language extension for DST.

Despite the progress in cross-lingual DST, previous studies of cross-lingual DST focus on transferring monolingual DST of source language to monolingual DST of target language. In this paper, we concentrate on a more realistic scenario that the DST model could generate a unified dialogue state when the input utterances are in different languages. In such setting, some downstream components can be reused when deploying multilingual TOD services. The two scenarios have been illustrated in Fig. 1. To realize this goal, we focus on a zero-shot language extension for DST, which can handle input from multiple languages and output a unified dialogue state.

Multilingual pre-trained models. Multilingual pre-trained models, such as mBERT [12], XLM-R [32], and mBART [33], have been applied to zero-shot cross-lingual transfer for various NLP tasks. The models are pre-trained on large size of corpora, some of which are pre-trained with masked language modeling objective using only monolingual data, like mBERT and XLM-R, while the multilingual pre-trained models with sequence-to-sequence architecture, mBART [33] and mT5 [34], are pre-trained on large-scale monolingual corpora across many languages. After pre-trained, they are fine-tuned on downstream tasks in one language and directly tested in other languages. Due to the imperfect alignments of cross-lingual embeddings, the performance of zero-shot cross-lingual transfer is limited [16]. For better cross-lingual representations, some studies use auxiliary pre-training tasks or extra parallel resources as explicit signals to encourage the alignment between source and target language space [12,35,36]. Conneau and Lample [35] propose two methods for learning cross-lingual representations, one is unsupervised using cross-lingual language modeling and the other is supervised method leveraging parallel data. VECO [37] and ERNIE-M [38] are proposed to leverage sentence-level parallel data to capture cross-lingual information. Hu et al. [39] propose to use two explicit alignment objectives that align the multilingual representations at the word and sentence level. Notably, these learned representations are often irrelevant to the downstream tasks, and the sentence-level parallel data is also expensive for low-resource languages.

Another line of work similar to ours is the zero-shot cross-lingual transfer of neural machine translation (NMT) [15,40,41]. These studies mainly consider mapping different source languages into the same semantic space, thus enabling the NMT model to translate sentences in source languages unseen during supervised training into the target language. Unlike the cross-lingual transfer of NMT task, the DST needs to consider the characteristic of dialogue and understand the dialogue history to obtain the user's goals and intents.

In this work, we are the first to focus on the zero-shot language extension for DST. We adopt mBART and mT5 as the backbone, and explore how to fine-tune the multilingual pre-trained model to realize the extension of monolingual DST into multilingual DST in a zero-shot setting.

3. Background and problem definition

3.1. Monolingual DST model

The goal of DST is to predict the dialogue state $\mathcal{S}_t = \{(d_i, s_i, v_i) | 1 \leq i \leq I\}$ by utilizing the history dialogue context $\mathcal{C}_t$ at the tth turn, where (d_j, s_j, v_j) is the (*domain, slot, value*) tuple, I is the number of states to be tracked, history dialogue context $\mathcal{C}_t = \{U_1, R_1, \ldots, R_{t-1}, U_t\}$ is composed of user utterance U_i and system response R_i before tth turn.

The state-of-the-art generative method for DST follows the encoder–decoder architecture, which treats the DST task as a sequence generation task. Specifically, the efficient *Lev* (*Levenshtein belief spans*) DST model proposed in Lin et al. [5] is adopted as our monolingual DST model. The model is implemented with a general encoder–decoder architecture. The idea of *Lev* is to generate minimal belief spans at each turn for editing previous dialogue states for current dialogue states. Given $\mathcal{S}_{t-1}$, $\mathcal{S}_t$, and a pair of (d, s), there are three slot level edit operation condition, including insertion, deletion, and substitution.

The encoder takes the concatenation of the dialogue context $\mathcal{C}_t$ and previous dialogue state $\mathcal{S}_{t-1}$ as input, and the decoder decodes Lev_t, which records the difference between old states and new states. All sub-sequences are concatenated with special segment tokens as input to the encoder.

$$\begin{aligned}\mathcal{C}_t = U_1 \oplus \langle \texttt{EOU} \rangle \oplus R_1 \oplus \langle \texttt{EOR} \rangle \oplus \ldots \oplus \\ R_{t-1} \oplus \langle \texttt{EOR} \rangle \oplus U_t \oplus \langle \texttt{EOU} \rangle \end{aligned} \tag{1}$$

where <EOU> and <EOR> are special tokens used to mark the boundaries of user utterance and system response, respectively.

$$\mathbf{H} = Encoder(\mathcal{S}_{t-1} \oplus \langle \texttt{EOB} \rangle \oplus \mathcal{C}_t) \tag{2}$$

$$Lev_t = Decoder(\mathbf{H}) \tag{3}$$

$$\mathcal{S}_t = f(Lev_t, \mathcal{S}_{t-1}) \tag{4}$$

<EOB> is a special token added between previous dialogue state $\mathcal{S}_{t-1}$ and history dialogue context $\mathcal{C}_t$. $\mathbf{H}$ is the hidden states of the encoder. The decoder attends to $\mathbf{H}$ and decodes Lev_t. The Lev_t is then used for editing $\mathcal{S}_{t-1}$ through function f, which updates the $\mathcal{S}_{t-1}$ when new slot-value pairs appear in Lev_t, and deletes the corresponding slot-value when the NULL symbol is generated. Given the training data $\mathcal{D} = \{(\mathcal{C}_t^X, \mathcal{S}_t^X)\}$ of DST in language X, the DST model is trained by minimizing the following objective function:

$$\mathcal{L}_\theta = -\log p(Lev_t | \mathcal{C}_t^X, \mathcal{S}_{t-1}^X; \theta) \tag{5}$$

The DST model is easily set up with pre-trained language models by initializing the model with pre-trained weights.

3.2. Problem definition

Our goal is to train a multilingual DST model $\theta_{(X, Y_1, \ldots, Y_n)}$ for languages $(X, Y_1, \ldots, Y_n)$, where X is the source language and $(Y_1, \ldots, Y_n)$ are the n different target languages. We only utilize the following resources:

(1) **DST training data in source language** X: $\mathcal{D}_X = \left\{(\mathcal{C}_t^X, \mathcal{S}_t^X)\right\}$, where $\mathcal{C}_t^X$ and $\mathcal{S}_t^X$ denote the dialogue context and corresponding dialogue state in X, respectively.

(2) **multilingual pre-trained model** $\Theta_{(X,Y_1,...,Y_n)}$: A multilingual pre-trained language model which contains the knowledge in source and target languages is also needed.

With the above two resources, we hope to train a single multilingual DST model, whose input is the dialogue context in any language from $(X, Y_1, .., Y_n)$, and the output is a unified dialogue state in language X.

4. Our method

Our goal is to learn a multilingual DST model $\theta_{(X,Y_1,...,Y_n)}$ with training data in source language X and a multilingual pre-trained model $\Theta_{(X,Y_1,...,Y_n)}$. To achieve this, we propose a *pre-trained models+multi-auxiliary-tasks fine-tuning* method in the zero-shot language extension scenario, which contains three steps: (1) multilingual pre-trained model setup, (2) multi-auxiliary-tasks design, and (3) fine-tuning.

4.1. Multilingual pre-trained model setup

Instead of pre-training a new multilingual model from scratch, we directly select an exiting multilingual pre-trained model since there have been various multilingual pre-trained models. As introduced before, the current DST task is treated as a sequence generation task and follows the encoder–decoder architecture. Thus we select sequence-to-sequence multilingual pre-trained models to train the multilingual DST model in zero-shot language extension scenario. Specifically, here we select **mBART** [33] and **mT5** [34] as the multilingual pre-trained models. Note that other sequence-to-sequence multilingual pre-trained models can also be applied in our method.

mBART is a complete autoregressive sequence-to-sequence model which is pre-trained on a subset of 25 languages using the BART objective [42]. Two types of noises are used to generate the corrected text. The first is to remove spans of text and replace them with a mask token, and the second is to permute the order of sentences within each instance.

mT5 is a multilingual variant of T5 [43], which is a sequence-to-sequence model. The mT5 is pre-trained on the mC4 dataset covering natural text in 101 languages using a span-corruption version of masked language modeling task. The model is trained to reconstruct all the masked spans in the inputs, using a standard cross-entropy loss.

4.2. Multi-auxiliary-tasks design

We elaborately design five auxiliary tasks to improve the performance of the extended multilingual DST model. As the basis of our method, we first introduce two preparations used in our auxiliary-task design.

(1) **Bilingual Lexicon Extraction.** Some recent work successfully extracted translation lexicons from two monolingual corpus [44,45]. By using the methods, we can build bilingual dictionaries $Dic_{(X,Y_i)} = \{(x, y_i)\}$ between source language X and each target language $Y_i \in (Y_1, \ldots, Y_n)$ without using any parallel corpora, where (x, y_i) denotes the source word x and its translation equivalence y_i in target language Y_i. Since we mainly concentrate on the zero-shot language extension of DST, we directly adopt the bilingual dictionaries released in facebookresearch/MUSE repository.[1]

(2) **Code-Switching Dialogue Context Generation.** After extracting bilingual lexicons for each target languages $Dic_{(X,Y_1)}, \ldots, Dic_{(X,Y_n)}$, we need to generate the code-switching dialogue context. Given each dialogue context $\mathcal{C}_t^X$ in training data of language X, we generate the code-switching dialogue context $\mathcal{C}_t^{Y_i}$ in language Y_i by replacing the source words with the target words as follows:

$$\mathcal{C}_t^{Y_i} \leftarrow \text{replace}(\mathcal{C}_t^X, Dic_{(X,Y_i)}, \delta) \tag{6}$$

where δ is the proportion of replaced words.[2] We generate the code-switching DST training data $\mathcal{D}_{Y_i} = \left\{(\mathcal{C}_t^{Y_i}, \mathcal{S}_t^X)\right\}$ for each target language Y_i. For example, consider the following source dialogue context:

`i want to take a taxi to airport .`

Given bilingual lexicon pairs (`taxi`, 出 租 车) and (`airport`, 机 场), we get the following code-switching dialogue context:

`i want to take a` 出租车 `to` 机场 `.`

We now introduce our designed **multi-auxiliary-tasks**. We design five auxiliary tasks to model the DST generation and explicitly make use of cross-lingual alignment information. As shown in Fig. 2, these tasks can be divided into three groups. The first group is for DST modeling, the second for word-level cross-lingual alignment modeling, and the third for sentence-level semantic similarity modeling.

Monolingual DST task in source language X (MDST). This task is used for DST modeling. As illustrated in Fig. 2 (1), given the dialogue history context $\mathcal{C}_t^X$ and previous dialogue state $\mathcal{S}_{t-1}^X$ in source language X, the MDST task forces the model to generate the corresponding Lev_t^X, which is used to edit the $\mathcal{S}_{t-1}^X$ to generate current dialogue state S_t^X. The goal of this task is to let the model obtain the ability of DST generation from clean training data. The training objective of this task can be formulated as:

$$\mathcal{L}_{\text{MDST}} = -\log p(Lev_t | \mathcal{C}_t^X, \mathcal{S}_{t-1}^X) \tag{7}$$

Cross-lingual DST task (CDST). The CDST task is similar to the MDST task, as shown in Fig. 2 (2), which predicts the dialogue state $\mathcal{S}_t^X$ given the code-switching dialogue context $\mathcal{C}_t^{Y_i}$ and previous dialogue state $\mathcal{S}_{t-1}^X$. The code-switching dialogue context contains partial information about the target languages, which can teach the model to generate source language dialogue state $\mathcal{S}_t^X$ from the history dialogue context in the target language. The training objective of this task can be formulated as follows:

$$\mathcal{L}_{\text{CDST}} = -\log p(Lev_t | \mathcal{C}_t^{Y_i}, \mathcal{S}_{t-1}^X) \tag{8}$$

Forward word translation task (FWT). The FWT task is to translate part of the words in the source language dialogue context $\mathcal{C}_t^X$ into words in target language Y_i, as illustrated in Fig. 2 (3). It takes $\mathcal{C}_t^X$ as input and $\mathcal{C}_t^{Y_i}$ as output. Formally, the training objective of FWT is defined as follows:

$$\mathcal{L}_{\text{FWT}} = -\log p(\mathcal{C}_t^{Y_i} | \mathcal{C}_t^X) \tag{9}$$

Utterance recovery task (REC). The direction of REC task is the opposite of the FWT task. Its goal is to recover the code-switching dialogue context $\mathcal{C}_t^{Y_i}$ to its original dialogue context $\mathcal{C}_t^X$, as shown in Fig. 2 (4). It takes $\mathcal{C}_t^{Y_i}$ as input and $\mathcal{C}_t^X$ as output. The FWT task and REC task are used to force the multilingual pre-trained model to learn better cross-lingual representations through explicitly utilizing the word-level cross-lingual alignment information. Similarly, the training objective of REC is defined as follows:

$$\mathcal{L}_{\text{REC}} = -\log p(\mathcal{C}_t^X | \mathcal{C}_t^{Y_i}) \tag{10}$$

[1] https://github.com/facebookresearch/MUSE

[2] We randomly choose any of the multiple translations as the target language word if multiple translations of the source word exist in the dictionary.

Semantic similarity task (SIM). Besides the word-level cross-lingual alignment information, we further design the SIM task to model the sentence-level semantic similarity. Fig. 2 (5) depicts the SIM task in detail. This task is to encourage the encoder to learn language-independent hidden representations. Since $\mathcal{C}_t^{Y_i}$ is converted from $\mathcal{C}_t^X$ through bilingual dictionary, the semantic information contained in the two dialogue context should be similar. Hence, we adopt a similarity loss over $\mathcal{C}_t^X$ and $\mathcal{C}_t^{Y_i}$ to learn a language-invariant encoder. The similarity loss is defined as follows:

$$\mathcal{L}_{\text{SIM}} = \text{L1Loss}(E_{\mathcal{C}_t^X}, E_{\mathcal{C}_t^{Y_i}}) \tag{11}$$

where $E_{\mathcal{C}_t^X}$ and $E_{\mathcal{C}_t^{Y_i}}$ are the sentence embeddings for $\mathcal{C}_t^X$ and $\mathcal{C}_t^{Y_i}$, respectively. The sentence embeddings are obtained by averaging the last hidden states of the encoder.

4.3. Fine-tuning

Fig. 2 describes the workflow of the *pre-trained models+multi-auxiliary-tasks fine-tuning* method with a general encoder–decoder architecture. We leverage the multi-task learning framework to incorporate the five tasks. To specify the task, we add task-specific prefixes to the input sequence, such as `Monolingual dst` for MDST task, `Cross-lingual dst` for CDST task, `EN to CS MT` for FWT task, and `CS to EN MT` for REC task. During fine-tuning, the total training objective is finally formulated as:

$$\begin{aligned}\mathcal{L}_{\theta_{(X,Y_1,\ldots,Y_n)}} &= \mathcal{L}_{\text{MDST}} + \mathcal{L}_{\text{CDST}} \\ &+\alpha(\mathcal{L}_{\text{FWT}} + \mathcal{L}_{\text{REC}}) + \gamma\mathcal{L}_{\text{SIM}}\end{aligned} \tag{12}$$

where α and γ are the hyper-parameters to control the balance among different tasks. The detailed training process is shown in Algorithm 1.

Algorithm 1 Multi-Auxiliary-Tasks Fine-tuning

Require: Initial parameters from multilingual **pre-trained** model $\Theta_{(X,Y_1,\ldots,Y_n)}$, initial auxiliary task weight α and γ
Require: Samples from different tasks $\mathcal{T} = \{MDST, CDST, FWT, REC, SIM\}$, initial learning rate η
Ensure: multilingual DST model $\theta_{(X,Y_1,\ldots,Y_n)}$

```
while Algorithm Not converge do
    for min-batch {C_t^X, C_t^{Y_i}, S_{t-1}^X, Lev_t} from task T do
        Compute the MDST objective L_MDST by Eq. (7)
        Compute the CDST objective L_CDST by Eq. (8)
        Compute the FWT objective L_FWT by Eq. (9)
        Compute the REC objective L_REC by Eq. (10)
        Compute the SIM objective L_SIM by Eq. (11)
        The total training objective L_{θ(X,Y_1,...,Y_n)} is computed by Eq. (12)
        Update the network parameters using AdamW with learning rate η
    end for
end while
```

5. Experiments

5.1. Experimental settings

Dataset. We evaluate the proposed framework with two dialogue datasets: the Multilingual WOZ 2.0 dataset [17] and the cross-lingual MultiWOZ dataset [18].

- **Multilingual WOZ 2.0:** The Multilingual WOZ 2.0 contains train, valid, and test datasets for three languages (English, German and Italian) and is expanded from the original WOZ 2.0 dataset [2], which is a restaurant reservation dataset in English consisting of three informable slots and seven requestable slots. We use English as the source language and German and Italian as the target languages. Our goal is to extend the DST in English into a multilingual DST only using the English training data, which can accept the German/Italian utterances and generate English dialogue states. To conduct the experiments, we make some modifications to the German and Italian test sets, in which we replace the dialogue states in German/Italian with their corresponding English dialogue states.
- **Cross-lingual MultiWOZ dataset:** The original dataset MultiWOZ 2.1 [46] is a large-scale multi-domain task-oriented dialogue benchmark containing over 10,000 dialogs. It contains dialogues between tourists and clerks at an information center across seven domains, including restaurant, hotel, attraction, etc. Both the ontology of the dialogue states and the dialogues were translated from English to Chinese using Google Translate and then corrected manually by expert annotators. The cross-lingual MultiWOZ dataset is released in the DSTC-9, and more details on the dataset creation can be referred to Gunasekara et al. [18]. During training and evaluation, the language of the dialogue state is consistent with the source language. For example, for the extension from English to Chinese, regardless of whether the input utterances are English or Chinese, the dialogue state language is English.

Evaluation metrics. We use joint goal accuracy (JGA) and slot F_1 as metrics to evaluate the performance of DST.

- **Joint Goal Accuracy (JGA)**: It calculates the proportion of dialogue turns where the predicted dialogue states exactly match the ground-truth dialogue states.
- **Slot** F_1: The macro-averaged F_1 score over all slots for every turn.

We report JGA and slot F_1 for both Multilingual WOZ 2.0 and cross-lingual MultiWOZ datasets. To be specific, the JGA and slot F_1 for Multilingual WOZ 2.0 only count on informable slots.

Implementation details. We set up our framework with two multilingual pre-trained models: (1) mT5-small[3]; (2) mBART-large-cc25[4] In all experiments, the models are optimized with AdamW [47] with learning rate of $6e^{-4}$ for mT5 and $1e^{-5}$ mBART, respectively. The window size of the dialogue history is 3, and the replacement proportion δ is 1.0. In fine-tuning, we select the best hyper-parameters by searching a combination of auxiliary task weight α, and γ with the following range: α: {0.001, 0.01, 0.1, 1.0}; γ: {0.01, 0.1, 1.0, 10.0}. All the models are fine-tuned with a batch size of 64 and early stop according to the performance on the validation set. Our implementation is based on HuggingFace Transformers library [48]. Each model is trained on 1 NVIDIA Tesla V100 GPU.

Baselines. Our goal is to extend the monolingual DST to multilingual DST, which can handle input from multiple languages and output a unified dialogue state in the source language. As there are no models evaluated on this zero-shot language extension scenario, we use the following methods as baselines.

- **Direct fine-tuning (Direct FT)**. Directly fine-tune mT5 and mBART on the training data in the source language and then apply the model to target languages.
- **Code-switching fine-tuning (CSFT)**. We replace the words in the source language with their target counterparts in the bilingual dictionary to generate code-switching training data. Then we fine-tune the multilingual pre-trained models with the generated code-switching data.

[3] https://huggingface.co/google/mt5-small
[4] https://huggingface.co/facebook/mbart-large-cc25

Table 1
Experimental results of zero-shot language extension on **Multilingual WOZ 2.0**. *EN → DE* and *EN → IT* denote expanding the English monolingual DST to German and Italian, respectively. The best results within each column are marked in **Bold**. **JGA**: Joint goal accuracy. The last two columns indicate average gain over *Direct FT* for target languages.

#	Model	EN → DE				EN → IT				Target language	
		English		German		English		Italian		Avg Gain	
		JGA	Slot F1	JGA	Slot F1	JGA	Slot F1	JGA	Slot F1	JGA	Slot F1
						mT5					
1	Direct FT	72.78	88.60	17.19	47.89	58.51	82.87	9.23	35.60	0.00	0.00
2	CSFT	59.90	81.70	54.01	78.57	80.26	91.92	56.93	80.77	42.26	37.93
3	Our Method	**84.93**	**94.43**	**65.43**	**84.92**	**88.46**	**95.89**	**71.45**	**88.26**	**55.23**	**44.85**
						mBART					
4	Direct FT	76.67	89.99	5.71	29.07	78.37	91.27	4.62	31.09	0.00	0.00
5	CSFT	63.43	84.16	53.40	77.27	67.25	86.56	60.09	82.78	51.58	49.95
6	Our Method	**84.99**	**93.93**	**67.01**	**84.70**	**81.83**	**92.69**	**72.78**	**88.38**	**64.73**	**56.46**

Table 2
Experimental results of zero-shot language extension on **cross-lingual MultiWOZ dataset**. *EN → ZH* and *ZH → EN* denote expanding the English monolingual DST to Chinese and reverse. The best results within each column are marked in **Bold**. **JGA**: Joint goal accuracy. The last two columns indicate average gain over *Direct FT* for target languages.

#	Model	EN → ZH				ZH → EN				Target language	
		English		Chinese		Chinese		English		Avg Gain	
		JGA	Slot F1	JGA	Slot F1	JGA	Slot F1	JGA	Slot F1	JGA	Slot F1
						mT5					
1	Direct FT	**51.28**	**90.29**	1.77	25.25	44.90	87.78	5.58	39.74	0.00	0.00
2	CSFT	31.96	81.59	14.96	68.07	33.79	82.11	25.85	78.14	16.73	40.61
3	Our Method	49.65	90.04	**16.28**	**69.25**	**47.50**	**89.24**	**27.62**	**79.36**	**18.28**	**41.81**
						mBART					
4	Direct FT	**44.47**	**88.00**	5.03	37.07	38.81	85.59	6.46	43.76	0.00	0.00
5	CSFT	35.01	83.00	14.39	**65.47**	29.44	76.17	22.78	65.88	12.84	25.26
6	Our Method	41.88	86.37	**14.71**	64.68	**41.07**	**86.39**	**27.59**	**79.25**	**15.41**	**31.55**

5.2. Experimental results

Tables 1 and 2 show the experimental results of zero-shot language extension for DST on Multilingual WOZ 2.0 and parallel MultiWOZ dataset, respectively. Table 1 reports the performance between two language pairs, including English to German (EN → DE) and English to Italian (EN → IT). Table 2 shows the results for extending English to Chinese (EN → ZH) and Chinese to English (ZH → EN). In both tables, Line 1–3 and Line 4–6 report the performance using two different multilingual pre-trained models, mT5 and mBART, respectively. Line 1 and Line 4 report the direct fine-tuning (FT) performance, where the DST model is obtained by fine-tuning the pre-trained models using the English training data and then directly applying to target languages. Line 2 and Line 5 report the results of code-switching fine-tuning. Line 3 and Line 6 report the results of our proposed pre-trained models+multi-auxiliary-task fine-tuning method. From the results, we can make the following conclusions:

(i) *Direct fine-tuning* cannot work well on the target languages (line 1 and line 4). Compared to the performance on the source language, the DST performance degrades drastically on target languages, including German, Italian, Chinese, and English. This implies that the multilingual contextual embedding spaces for different languages are not perfect in the multilingual pre-trained model, thus limiting the performance of direct fine-tuning.

(ii) *Code-switching fine-tuning*, which fine-tunes the mT5 and mBART using generated code-switching data, boosts the extended DST performance for the target languages (line 2 and line 4). The code-switching data is generated by replacing words in source language data with their translations in the bilingual dictionary, and it contains explicit lexicon knowledge about target languages. Fine-tuning on such data improves the performance of all the target languages. From Tables 1 and 2, we also observe that the improvements between EN → DE and EN → IT are much more obvious than the improvements between EN → ZH and ZH → EN. However, such fine-tuning leads to obvious performance

Table 3
Ablation study of different auxiliary tasks. *MDST*, *CDST*, *FWT*, *REC*, and *SIM* denote the monolingual DST task, cross-lingual DST task, forward word translation task, utterance recovery task, and semantic similarity task, respectively. *+FWT+REC* denotes combining *MDST*, *CDST*, *FWT*, and *REC* during fine-tuning. *+SIM* means using *MDST*, *CDST*, and *SIM* during fine-tuning. *full tasks* denotes jointly utilizing the above five tasks together.

Model	EN → DE				EN → IT			
	English		German		English		Italian	
	JGA	Slot F1	JGA	Slot F1	JGA	Slot F1	JGA	Slot F1
				mT5, Multilingual WOZ 2.0				
MDST+CDST	64.95	84.63	58.81	80.73	84.99	94.27	65.92	85.96
+FWT+REC	84.81	**94.78**	64.09	83.70	87.23	94.95	67.98	86.24
+SIM	83.35	93.83	60.15	81.67	85.12	94.63	67.44	86.37
Full tasks	**84.93**	94.43	**65.43**	**84.92**	**88.46**	**95.89**	**71.45**	**88.26**
				mBART, Multilingual WOZ 2.0				
MDST+CDST	74.30	88.74	56.20	78.71	75.76	89.58	59.84	81.63
+FWT+REC	84.99	93.93	67.01	84.70	76.79	90.46	67.86	86.47
+SIM	75.70	89.93	58.02	80.93	74.67	89.65	66.16	86.24
Full tasks	**84.99**	**93.93**	**67.01**	**84.70**	**81.83**	**92.69**	**72.78**	**88.38**
Model	EN → ZH				ZH → EN			
	English		Chinese		Chinese		English	
	JGA	Slot F1	JGA	Slot F1	JGA	Slot F1	JGA	Slot F1
				mT5, Cross-lingual MultiWOZ				
MDST+CDST	50.48	90.19	14.85	68.72	46.64	88.64	26.60	78.63
+FWT+REC	49.65	90.04	16.28	69.25	47.50	89.34	27.51	79.88
+SIM	**51.24**	**90.38**	15.75	69.00	47.50	89.24	27.62	79.36
Full tasks	49.65	90.04	**16.28**	**69.25**	**47.50**	**89.24**	**27.62**	**79.36**

Table 4
Ablation study (JGA/Slot F1 score (%)) for different replacement proportion δ. The best results for each language pair are marked in Bold.

δ	MDST+CDST				Full tasks			
	EN → DE		EN → IT		EN → DE		EN → IT	
	English	German	English	Italian	English	German	English	Italian
0.1	77.28/90.92	33.41/62.69	**82.87/93.07**	50.97/75.61	79.65/92.11	36.63/64.76	77.04/90.35	47.87/74.10
0.3	73.45/88.99	50.67/76.69	76.67/90.28	58.38/81.31	79.04/91.93	54.31/78.31	78.68/91.23	59.11/81.25
0.5	**79.59/91.77**	57.23/79.73	71.87/88.59	60.69/82.63	83.05/93.36	63.30/**83.45**	76.43/91.22	64.34/85.02
0.7	75.70/89.78	**59.66/80.87**	78.74/91.31	**66.71/85.08**	**84.02/94.29**	**63.37**/83.22	77.34/90.12	64.46/84.57
1.0	74.30/88.74	56.20/78.71	75.76/89.58	59.84/81.63	78.55/91.40	57.84/80.98	**80.44/92.20**	**70.05/87.52**

degradation for the source language (except for EN → IT when fine-tunes the mT5).

(iii) Our proposed *pre-trained models+multi-auxiliary-task fine-tuning* method, which fine-tunes the multilingual pre-trained model through multiple elaborately designed auxiliary tasks, improves the performance for all the extended languages significantly (line 3 and line 6). Furthermore, our proposed method can even improve the DST performance for the source language. This is an important speciality since it implies that we can extend the DST model to new languages without compromising the performance of the source language. Hence, we can maintain one unique DST model for different languages, reducing the maintenance costs of deploying multilingual TOD. Generally speaking, the experimental results suggest that our proposed method can effectively leverage multilingual pre-trained models with multiple auxiliary tasks to realize the zero-shot language extension for DST.

5.3. Model analysis

The effect of different auxiliary tasks. In this paper, due to the lack of annotated training data in target languages, we elaborately design five auxiliary tasks to achieve language extension. To investigate the effect of each auxiliary task, we conduct an ablation study by adding the objectives of the auxiliary tasks one at a time until the full learning objective. Since both FWT and REC are used for the word-level cross-lingual alignment modeling, we add the two tasks together (+FWT+REC). Table 3 reports the ablation results. First of all, all auxiliary tasks are helpful as adding any of them will bring a performance boost both for the source language and the extended target languages. The approach degenerates to direct fine-tuning when all the auxiliary tasks are removed. Without any explicit information on target languages, the DST performance on the target languages is much worse than that in the source language. After adding CDST (MDST+CDST), the DST performance of target languages improves a lot compared to using MDST or CDST alone. MDST+CDST archives comparable performance with mT5 while improves performance with mBART for English.

Secondly, the auxiliary tasks FWT, REC, and SIM help improve the DST performance further. Based on MDST+CDST, adding FWT and REC (+FWT+REC) or adding SIM (+SIM) improves the performance. We also notice that +FWT+REC performs better than +SIM on most extended languages. Utilizing the five tasks together achieves the best performance on the extended languages. Additionally, our approach can bring a windfall benefit, which is the DST performance of the source language is greatly improved at the same time. This is quite important since it implies that we only need to maintain one DST model in the multilingual scenario. These ablation results demonstrate that explicitly utilizing the word-level cross-lingual alignment and sentence-level semantic similarity is critical for the zero-shot language extension of DST. The improvements are consistent across both Multilingual WOZ 2.0 and cross-lingual MultiWOZ datasets, affirming our proposed method's superiority and general applicability.

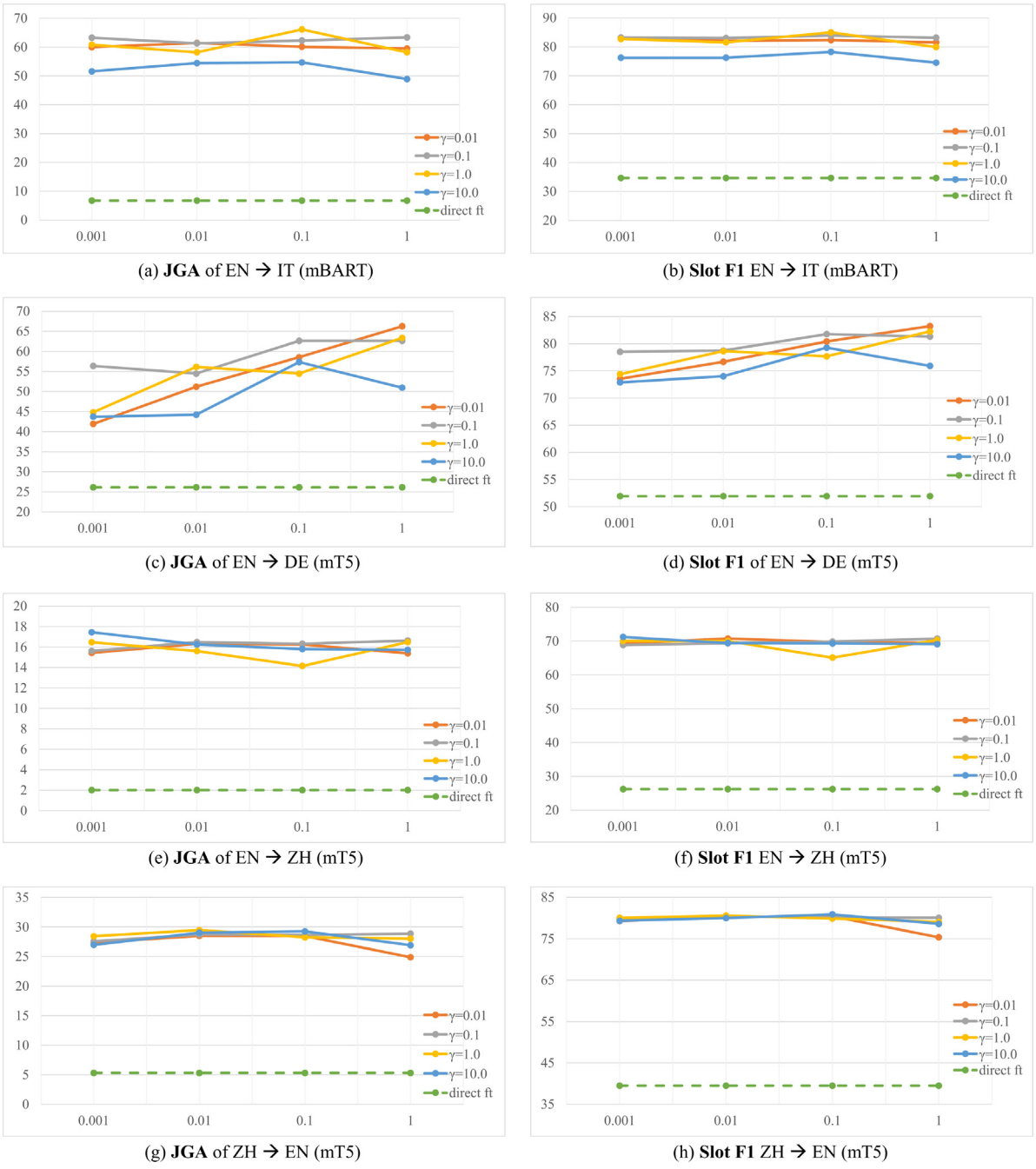
Table 5
The statistical information of the dictionaries. *#entries* denotes the total number of entries while *#uniq* denotes the total number of unique source language words in the dictionaries. *#words* denotes the number of unique words in the DST training data. *cover* counts the percentage of words in the DST training data that exist in the dictionaries.

Dict	#entries	#uniq	#words	cover(%)
Multilingual WOZ 2.0				
EN-DE	53,097	29,464	2,041	32.29
EN-IT	41,219	30,592	2,041	28.81
Cross-lingual MultiWOZ				
EN-ZH	21,578	15,160	15,653	13.05
ZH-EN	21,549	13,713	18,109	12.40

The effect of replacement proportion δ. We adopt the MUSE dictionaries to generate the code-switching data. Since there are noises in the dictionaries, we use the following two rules to clean the dictionary: (1) Remove the entries that the source words and target words are the same, and (2) Remove the entries that the language of the words is not correct. To investigate the effect of dictionaries, we count the statistical information of the cleaned dictionaries from several aspects. The statistical information is given in Table 5. Combining these statistics with the experimental results in Tables 1 and 2, we can find that the dictionary coverage (*cover*) is related to the performance of both the code-switching fine-tuning and our method.

We further investigate the effect of replacement proportion δ on model fine-tuning. We set $\delta \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$ and fine-tune mBART with two settings: (1) only using MDST+CDST; and (2) using the five tasks together ($\alpha = 0.1, \gamma = 0.01$ when extending English to German, and $\alpha = 0.1, \gamma = 0.1$ when extending English to Italian). We evaluate the performances and report the results in Table 4. For MDST+CDST, when δ is greater than 0.0, the DST performance in German and Italian is superior to the performance when δ is 0.0. This demonstrates that code-switching does actually improve the DST performance for target languages. When the replacement proportion δ is greater than 0.7, the performance on target languages starts to degrade. Without carefully adjusting α and γ, the model with full tasks performs better than the model using only MDST+CDST for both the source and target languages in most of the replacement proportions, demonstrating the consistency of our method. Furthermore, replacing only 30% or even 10% tokens in the source language, our multi-auxiliary-tasks fine-tuning outperforms MDST+CDST, indicating that our proposed method is still useful when only a small-scale bilingual dictionary is available.

Impact of hyper-parameters α and γ. There are two hyper-parameters in the full learning objective, in which α controls the balance of bidirectional word translation (*FWT+REC*) and γ controls the weight of utterance semantic similarity (*SIM*). Fig. 3 presents validation results when training the model with replacement proportion δ be 1.0 and different combinations of α and γ, where α ranges in $\{0.001, 0.01, 0.1, 1.0\}$ and γ ranges in



(a) **JGA** of EN → IT (mBART) (b) **Slot F1** EN → IT (mBART)

(c) **JGA** of EN → DE (mT5) (d) **Slot F1** of EN → DE (mT5)

(e) **JGA** of EN → ZH (mT5) (f) **Slot F1** EN → ZH (mT5)

(g) **JGA** of ZH → EN (mT5) (h) **Slot F1** ZH → EN (mT5)

Fig. 3. Results of target languages with different α and γ. (a)–(d) are the results on Multilingual WOZ 2.0, where (a) and (b) present the results on EN → IT with mBART as the backbone, while (c) and (d) present the results on EN → DE with mT5 as backbone. (e)–(h) demonstrate the results with mT5 on cross-lingual MultiWOZ dataset, where (e) and (f) illustrate the results on EN → ZH, and (g) and (h) are the results on ZH → EN. The horizontal axis represents α. The solid lines with different colors denote performance with different γ. The dash line denotes the performance with direct fine-tuning.

{0.01, 0.1, 1.0, 10.0}. Fig. 3(a)–(d) present the results on Multilingual WOZ 2.0, and (e)–(h) show the results on cross-lingual MultiWOZ. With all the combinations of α and γ, our model consistently outperforms the baseline model by a large margin in JGA and slot F_1, which verifies the robustness of our proposed method. In addition, we have two interesting observations from Fig. 3(a)–(d). First, the performance of the model is worse than others regardless of the α value when γ is 10.0 on Multilingual WOZ 2.0 dataset. In Fig. 3(a) and (b), the solid line that denotes the performance of $\gamma = 10.0$ is under the other solid lines. Second, α and γ have different trends when using mT5 and mBART as the backbone. When fine-tuning mT5, the model favors bigger α and smaller γ, which indicates that FWT+REC is more important for mT5 while SIM is less helpful. When fine-tuning mBART, the performances are relatively more stable with different α and γ. Both the FWT+REC and SIM are useful for improving the performance of zero-shot language extension. For cross-lingual MultiWOZ, both the performance of extending English DST to Chinese and extending Chinese DST to English are relatively stable. From the trends of lines with different γ, the model benefits from smaller α and bigger γ when fine-tuning with mT5.

Context words vs. value words. Since DST tracks the slot values mentioned in the dialogue, we define two kinds of words. ***Value*** words are those words that mention the dialogue states,

Table 6
The JGA score of replacing different types of words (mBART as the backbone). *Random* denotes random replacement. *Context* denotes replacing the contextual words. *Value* denotes replacing the slot value words.

Model	EN → DE		EN → IT	
	English	German	English	Italian
Random	63.43	**53.40**	67.25	60.09
Context	64.28	12.70	68.96	12.03
Value	**69.02**	48.97	**71.32**	**63.30**

Table 7
The JGA score of replacing phrase-level slot values.

Model	EN → ZH		ZH → EN	
	English	Chinese	Chinese	English
Direct FT	**44.47**	5.03	38.81	6.46
CSFT	23.09	24.83	9.76	28.11
CSFT+MDST	42.78	**28.50**	**38.81**	**35.72**

while the rest are ***Context*** words. For example, considering the following utterance:

`i am looking for a place in the centre of town.`

centre is the Value words, and the others are Context words.

We conduct experiments to investigate the effect of each type of words. The experimental results are shown in Table 6. As shown in Table 6, Fine-tuning on ***Random*** code-switching data achieves the best performance in German, while fine-tuning on ***Value*** achieves the best in Italian. Both the two strategies perform much better than only replacing the context words. This implies that the generation of DST is relevant to both context and value words. And the translations of value words are more important than the context words when extending the DST to other languages. In this paper, we adopt the random substitution strategy.

The effect of replacement granularity. From Table 2, we observe that the improvements of our method are limited when extending the English DST to Chinese. After analyzing the English training data, we find that many slot values are phrases rather than words, such as `guest house`, `birmingham new street`, etc. However, we just randomly select one word and replace the word with the target word, which may destroy the semantic integrity of the slot values. To investigate the effect of replacement granularity, we add the English to Chinese translations of slot values provided in DSTC-9 to the English–Chinese MUSE dictionary. When generating code-switching data, we preferentially replace the phrase-level slot values. For the rest words, we use the approach described in Section 4.2 to deal with. Then, we fine-tune mBART with the generated code-switching data. The experimental results are shown in Table 7. As shown in Table 7, both CSFT and MDST+CDST can bring surprising improvements for the target languages, indicating that it will be much helpful to ensure the semantic integrity of slot values when generating the code-switching data.

Compared with translate-train and translate-test method. We also compare our proposed pre-trained models+multi-auxiliary-task fine-tuning method with translate-train and translate-test method. In translate-train method, we first translate the English training data into the target language using a machine translation system and pair the translated utterances with the original dialogue state in English. Then the synthetic translated training data is used to fine-tune the pre-trained model. In translate-test method, we first translate the target language test set into English and then input the translated test set into the English DST model. Here, we use *MBART-50* [49], which is a fine-tuned model for multilingual machine translation, as the translation model. The experimental results are given in Table 8.

Table 8
Comparison results of different methods (JGA/Slot F1 score (%), mBART as the backbone). *TL-Test* denotes the *translate-test* method and *TL-Train* denotes the *translate-train* method.

#	Model	EN → DE				EN → IT			
		English		German		English		Italian	
		JGA	Slot F1	JGA	Slot F1	JGA	Slot F1	JGA	Slot F1
1	Direct FT	76.67	89.99	5.71	29.07	78.37	91.27	4.62	31.09
2	TL-Test	76.67	89.99	59.17	79.32	78.37	91.27	57.96	78.49
3	TL-Train	67.07	85.67	65.61	84.32	58.08	80.02	60.33	81.51
4	CSFT	63.43	84.16	53.40	77.27	67.25	86.56	60.09	82.78
5	Our Method	**84.99**	**93.93**	**67.01**	**84.70**	**81.83**	**92.69**	**72.78**	**88.38**

Both translate-train and translate-test improve the DST performance on target languages. On the target languages, the improvement of translate-train is higher than that of translate-test. However, the translate-train approach, which fine-tunes the model using the machine-translated target language training data, harms the performance of the source language. The code-switching fine-tuning method can achieve comparable performance with the translate-test method on German and Italian. Our proposed pre-trained models+multi-auxiliary-tasks fine-tuning method achieves the best performance for all languages, including the source language (English) and the extended languages (German and Italian). Compared with translate-train and translate-test which require machine translation systems as language bridges, our method only needs an automatically extracted bilingual dictionary to generate code-switching data and multiple elaborately designed tasks to realize multilingual DST. This also demonstrates the superiority of our proposed method.

5.4. Extending to other languages

To verify our proposed method on more language pairs, we automatically translate the MultiWOZ 2.1 test set into French, Spanish and Vietnamese using Google Translator. We only translate the utterances into the target languages and pair the utterances in target languages with the English dialogue states to compose the test datasets. The experimental results using machine-translated test sets are given in Table 9. From the results, we can observe that: (1) Our method outperforms the baseline systems by a large margin for all the target languages, demonstrating the effectiveness of our proposed method. (2) When extending English to other languages, compared to Chinese and Vietnamese, our method is more effective for French and Spanish. We believe there are two main reasons for this. On the one hand, French and Spanish are closer to English. It would be easier for the multilingual pre-trained models to map the three languages into the same semantic space. On the other hand, the dictionary coverage of English–French, English–Spanish, English–Vietnamese, and English–Chinese is 16.79%, 18.37%, 7.69%, and 13.05%, respectively. The dictionary coverage of the former two language pairs is higher than the latter two, which means more target language tokens can be introduced into the training process.

5.5. Extending to multiple languages simultaneously

The above experimental results have demonstrated that our proposed pre-trained models+ multi-auxiliary-tasks fine-tuning method effectively extends one additional language to the DST model. To prove the universality of our method, we conduct experiments to extend the English monolingual DST to multiple languages simultaneously on cross-lingual MultiWOZ dataset. The results are shown in Table 10. Compared to the direct fine-tuning procedure, the proposed method brings solid improvements for

Table 9
JGA/Slot F1 score (%) of zero-shot language extension for DST using machine-translated test set. *EN → FR*, *EN → ES*, and *EN → VI* denote expanding the English monolingual DST to French, Spanish, and Vietnamese, respectively.

#	Model	EN → FR		EN → ES		EN → VI	
		English	French	English	Spanish	English	Vietnamese
				mT5			
1	Direct FT	52.22/**90.07**	8.91/48.04	**52.22**/90.07	9.65/52.36	52.22/90.07	2.52/21.30
2	CSFT	47.57/88.78	21.10/72.62	49.21/89.14	37.73/84.28	48.76/88.17	6.02/41.83
3	Our method	**52.56**/89.97	23.54/73.74	51.35/**90.18**	40.96/85.54	**52.42/90.33**	6.51/50.15
				mBART			
4	Direct FT	42.44/86.29	9.60/49.25	42.44/**86.29**	9.79/53.90	42.44/86.29	5.13/34.47
5	CSFT	35.29/81.86	20.34/68.13	38.12/82.93	24.18/72.23	37.10/83.35	8.85/49.75
6	Our method	**45.96/87.09**	27.10/74.74	**43.55**/85.99	30.53/77.07	**45.08/87.59**	12.27/58.31

Input language	Dialogue Context
	Multilingual WOZ 2.0
Italian	**User:** vorrei un ristorante economico che serve cibo portoghese (*i would like a cheap restaurant that serves portuguese food*)
	Direct FT: [restaurant] food *portooccan*
	CSFT: [restaurant] food ***portuguese***
	Our Method: [restaurant] food ***portuguese*** pricerange ***cheap***
	Gold: [restaurant] food ***portuguese*** pricerange ***cheap***
German	**User:** ich suche ein restaurant im nördlichen teil der stadt (*i need to find a restaurant in the north side of town)* **Sys:** ok . did you have a certain price range of cuisine in mind? **User:** es muss keine bestimmte preisklasse sein . ich suche nach einem restaurant, das katalanisches essen serviert. (*i do not have a price range . i'm looking for a restaurant that serves catalan food.*)
	Direct FT: [restaurant] food kascan area dontcare pricerange expensive
	CSFT: [restaurant] area ***north*** food ***catalan***
	Our Method: [restaurant] area ***north*** food ***catalan*** pricerange ***dontcare***
	Gold: [restaurant] pricerange ***dontcare*** food ***catalan*** area ***north***
	Cross-lingual MultiWOZ
English	**User:** i need to find a train from cambridge to ely that leaves after 09:00 **Sys:** what day would you like to travel ? **User:** i will be traveling on monday .
	Direct FT: [出租车] 出发时间 ***09:00*** 出发地 剑桥桥梁宾馆 目的地 剑桥桥梁宾馆 ([taxi] leave *09:00* departure *bridge guest house* destination *bridge guest house*)
	CSFT: [列车] 出发时间 ***09:00*** 目的地 剑桥 出发地 伊利 日期 ***星期一*** ([train] leave *09:00* destination *cambridge* departure *ely* day *monday*)
	Our Method: [列车] 出发时间 ***09:00*** 目的地 ***伊利*** 出发地 ***剑桥*** 日期 ***星期一*** ([train] leave *09:00* destination *ely* departure *cambridge* day *monday*)
	Gold: [列车] 出发时间 ***09:00*** 目的地 ***伊利*** 出发地 ***剑桥*** 日期 ***星期一*** ([train] leave *09:00* destination *ely* departure *cambridge* day *monday*)
	Sys: i have tr8399 leaving cambridge at 09:50 and arriving at 10:07 in ely . would you like to book that ? **User:** what is the travel time and price ? **Sys:** it is a short trip , just 17 minutes and the cost is 4.40 pounds . **User:** i also need somewhere to sleep **Sys:** i would be happy to help . may i ask what price range and area of town you are looking for ? **User:** i need a 4 star hotel .
	Direct FT: [旅馆] 星级 ***4*** [出租车] 出发时间 ***09:00*** 出发地 剑桥桥梁宾馆 目的地 剑桥桥梁宾馆 ([hotel] stars *4* [taxi] leave *09:00* departure *bridge guest house* destination *bridge guest house*)
	CSFT: [旅馆] 星级 ***4*** [列车] 出发时间 09:00 目的地 剑桥 出发地 伊利 日期 星期一 (([hotel] stars *4* [train] leave *09:00* destination *cambridge* departure *ely* day *monday*)
	Our Method: [旅馆] 星级 ***4*** 类型 宾馆 [列车] 出发时间 09:00 目的地 伊利 出发地 剑桥 日期 星期一 ([hotel] stars *4* type *guest house* [train] leave *09:00* destination *ely* departure *cambridge* day *monday*)
	Gold: [旅馆] 星级 ***4*** 类型 ***酒店*** [列车] 出发时间 ***09:00*** 目的地 ***伊利*** 出发地 ***剑桥*** 日期 ***星期一*** ([hotel] stars *4* type *hotel* [train] leave *09:00* destination *ely* departure *cambridge* day *monday*)

Fig. 4. Three examples of multilingual DST. The upper part is two examples from The Multilingual WOZ 2.0, and the lower part is one dialogue from the cross-lingual MultiWOZ. The **Input language** denotes the target language that the DST model is extended to. The words marked in gray denote the error states compared with the gold states.

all the extended languages. What is more, compared to extending to one language, extending to multiple languages can further enhance multilingual DST's performance. Such improvements affirm the superiority and general applicability of our proposed method. It also proves that our method has the potential to maintain one DST model for multiple languages.

Table 10
Experimental results of extending English to another two languages simultaneously using mT5.

EN → FR, ES			
Model	English	French	Spanish
Direct FT	52.22/90.07	8.91/48.04	9.65/52.36
Our Method	52.00/89.98	25.47/74.80	38.46/84.56
EN → ES, ZH			
Model	English	Spanish	Chinese
Direct FT	52.22/90.07	9.65/52.36	1.77/25.25
Our Method	50.52/89.24	37.10/83.40	15.23/64.56

Table 11
The corresponding translations of the Chinese word in the Chinese–English MUSE dictionary.

Chinese word	English translation
宾馆	hotel, guesthouse
酒店	hotel, hotels
旅馆	hotel, hotels

5.6. Case study

We further demonstrate a qualitative analysis based on the case study. Fig. 4 shows three cases and the corresponding dialogue states generated by different models. As we can see, our proposed method generates more accurate dialogue states for the extended target languages. For the two dialogues in Multilingual WoZ 2.0, the extended model cannot work well for the target languages when extending the English DST to Italian or German with Direct FT. When tested with target languages, the model with direct fine-tuning generates correct slots, but the corresponding slot values are all wrong. For example, the model generates the correct slot "`food`" but the wrong slot value "`portooccan`" when the Italian utterance is given. After fine-tuning with code-switching data, the model can generate better dialogue states than direct FT. The CSFT generates partially correct dialogue states. Our proposed method, which fine-tunes the multilingual pre-trained model with multiple auxiliary tasks, generates completely correct dialogue states for Italian and German, showing its efficacy in capturing semantic information across languages.

The lower part in Fig. 4 shows the generated dialogue states as the conversation goes on. We can find that our model is much superior to the baselines. Notably, when generating new dialogue states given the user input "`i need a 4 star hotel`", our method generates the wrong slot value "类型 宾馆" (type guest house), while the correct one should be "类 型 酒店" (type hotel). This is because we use the MUSE dictionary to construct code-switching dialogue context. In the Chinese–English bilingual dictionary, the corresponding translations of 宾馆, 酒店, and 旅馆 are listed in Table 11. From Table 11, we can find that it is difficult for the model to generate the correct Chinese word for `hotel`. This experimental result also suggests that slot values should be uniformly translated when generating code-switching data.

6. Conclusion

This work explores a method for deploying a multilingual DST model, which accepts user input from different languages and outputs a unified dialogue state. Such a DST model can negate the need for multiple-language support of some downstream components in dialogue systems. Only a multilingual pre-trained model and the source language training data are needed. To alleviate the zero training data problem, we propose a *pre-trained models+multi-auxiliary-tasks fine-tuning* method, in which multiple auxiliary tasks are well-designed for the zero-shot language extension for DST. Experimental results have demonstrated the effectiveness and superiority of our method.

In this research, we conduct an extensive number of experiments in order to optimize hyper-parameters and explore the effects of each designed task, bilingual dictionary, and even the code-switching granularity. But there still exist some unexplored areas. In future work, we will analyze the differences between different language families and experiment with more language pairs with different morphological and syntactic structures. In addition, from our experimental results, we also find out that the performance of our proposed method is related to the dictionary coverage. The improvements for Vietnamese are limited. Thus, we are also interested in studying other techniques or more accessible resources to improve the DST performance for low-resource languages.

CRediT authorship contribution statement

Lu Xiang: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft. **Yang Zhao:** Conceptualization, Methodology, Resources, Writing – review & editing. **Junnan Zhu:** Methodology, Writing – review & editing. **Yu Zhou:** Project administration, Supervision, Methodology, Writing – review & editing. **Chengqing Zong:** Project administration, Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600.

References

[1] S. Young, M. Gašić, B. Thomson, J.D. Williams, Pomdp-based statistical spoken dialog systems: A review, Proc. IEEE 101 (5) (2013) 1160–1179.
[2] T.-H. Wen, D. Vandyke, N. Mrkšić, M. Gašić, L.M. Rojas-Barahona, P.-H. Su, S. Ultes, S. Young, 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: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 438–449.
[3] W. Lei, X. Jin, M.-Y. Kan, Z. Ren, X. He, D. Yin, Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, Melbourne, Australia, 2018, pp. 1437–1447.
[4] B. Peng, C. Li, J. Li, S. Shayandeh, L. Liden, J. Gao, SOLOIST: Few-shot task-oriented dialog with a single pre-trained auto-regressive model, in: Transactions of the Association for Computational Linguistics, 2020.
[5] Z. Lin, A. Madotto, G.I. Winata, P. Fung, MinTL: Minimalist transfer learning for task-oriented dialogue systems, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 3391–3405, Online.
[6] W. Liang, Y. Tian, C. Chen, Z. Yu, MOSS: end-to-end dialog system framework with modular supervision, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 8327–8335.
[7] W. Chen, J. Chen, Y. Su, X. Wang, D. Yu, X. Yan, W.Y. Wang, XL-NBT: A cross-lingual neural belief tracking framework, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018, pp. 414–424.

[8] Z. Liu, G.I. Winata, Z. Lin, P. Xu, P. Fung, Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 8433–8440.

[9] L. Qin, M. Ni, Y. Zhang, W. Che, CoSDA-ML: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, 2020, pp. 3853–3860.

[10] S. Schuster, S. Gupta, R. Shah, M. Lewis, 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, Vol. 1, Long and Short Papers, Minneapolis, Minnesota, 2019, pp. 3795–3805.

[11] Y.-T. Lin, Y.-N. Chen, An empirical study of cross-lingual transferability in generative dialogue state tracker, 2021, arXiv e-prints, arXiv:2101.11360.

[12] S. Wu, M. Dredze, Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT, 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, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 833–844.

[13] K. Song, X. Tan, T. Qin, J. Lu, T. Liu, MASS: masked sequence to sequence pre-training for language generation, in: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, PMLR, Long Beach, California, USA, 2019, pp. 5926–5936.

[14] B. Zhang, P. Williams, I. Titov, R. Sennrich, Improving massively multilingual neural machine translation and zero-shot translation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 1628–1639, Online.

[15] G. Chen, S. Ma, Y. Chen, L. Dong, D. Zhang, J. Pan, W. Wang, F. Wei, Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 15–26.

[16] S. Cao, N. Kitaev, D. Klein, Multilingual alignment of contextual word representations, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020, OpenReview.net.

[17] N. Mrkšić, I. Vulić, D. Ó Séaghdha, I. Leviant, R. Reichart, M. Gašić, A. Korhonen, S. Young, Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints, Trans. Assoc. Comput. Linguist. 5 (2017) 309–324.

[18] C. Gunasekara, S. Kim, L.F. D'Haro, A. Rastogi, Y.-N. Chen, M. Eric, B. Hedayatnia, K. Gopalakrishnan, Y. Liu, C.-W. Huang, D. Hakkani-Tür, J. Li, Q. Zhu, L. Luo, L. Liden, K. Huang, S. Shayandeh, R. Liang, B. Peng, Z. Zhang, S. Shukla, M. Huang, J. Gao, S. Mehri, Y. Feng, C. Gordon, S.H. Alavi, D. Traum, M. Eskenazi, A. Beirami, Eunjoon, Cho, P.A. Crook, A. De, A. Geramifard, S. Kottur, S. Moon, S. Poddar, R. Subba, Overview of the ninth dialog system technology challenge: DSTC9, 2020, arXiv e-prints, arXiv:2011.06486.

[19] N. Mrkšić, D. Ó Séaghdha, T.-H. Wen, B. Thomson, S. Young, 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, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1777–1788.

[20] L. Ren, K. Xie, L. Chen, K. Yu, Towards universal dialogue state tracking, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2780–2786.

[21] V. Zhong, C. Xiong, R. Socher, Global-locally self-attentive encoder for dialogue state tracking, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1458–1467.

[22] L. Chen, B. Lv, C. Wang, S. Zhu, B. Tan, K. Yu, Schema-guided multi-domain dialogue state tracking with graph attention neural networks, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 7521–7528.

[23] F. Ye, J. Manotumruksa, Q. Zhang, S. Li, E. Yilmaz, Slot self-attentive dialogue state tracking, in: Proceedings of the Web Conference 2021, WWW '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1598–1608.

[24] C.-S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, P. Fung, Transferable multi-domain state generator for task-oriented dialogue systems, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 808–819.

[25] S. Kim, S. Yang, G. Kim, S.-W. Lee, Efficient dialogue state tracking by selectively overwriting memory, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 567–582, Online.

[26] J. Hu, Y. Yang, C. Chen, L. He, Z. Yu, SAS: Dialogue state tracking via slot attention and slot information sharing, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 6366–6375, Online.

[27] D. Ham, J.-G. Lee, Y. Jang, K.-E. Kim, End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 583–592, Online.

[28] E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, R. Socher, A simple language model for task-oriented dialogue, 2020, arXiv e-prints, arXiv:2005.00796.

[29] M. Heck, C. van Niekerk, N. Lubis, C. Geishauser, H.-C. Lin, M. Moresi, M. Gasic, TripPy: A triple copy strategy for value independent neural dialog state tracking, in: Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, 1st virtual meeting, 2020, pp. 35–44.

[30] H. Bai, Y. Zhou, J. Zhang, L. Zhao, M.-Y. Hwang, C. Zong, Source critical reinforcement learning for transferring spoken language understanding to a new language, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3597–3607.

[31] N. Moghe, M. Steedman, A. Birch, Cross-lingual intermediate fine-tuning improves dialogue state tracking, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 1137–1150, Online and Punta Cana, Dominican Republic.

[32] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8440–8451, Online.

[33] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Trans. Assoc. Comput. Linguist. 8 (2020) 726–742.

[34] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 483–498, Online.

[35] A. Conneau, G. Lample, Cross-lingual language model pretraining, in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 2019, pp. 7057–7067.

[36] M. Artetxe, S. Ruder, D. Yogatama, On the cross-lingual transferability of monolingual representations, 2019, arXiv e-prints, arXiv:1910.11856.

[37] F. Luo, W. Wang, J. Liu, Y. Liu, B. Bi, S. Huang, F. Huang, L. Si, VECO: Variable and flexible cross-lingual pre-training for language understanding and generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Volume 1: Long Papers, Association for Computational Linguistics, 2021, pp. 3980–3994, Online.

[38] X. Ouyang, S. Wang, C. Pang, Y. Sun, H. Tian, H. Wu, H. Wang, ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 27–38.

[39] J. Hu, M. Johnson, O. Firat, A. Siddhant, G. Neubig, Explicit alignment objectives for multilingual bidirectional encoders, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 3633–3643, Online.

[40] G. Chen, S. Ma, Y. Chen, D. Zhang, J. Pan, W. Wang, F. Wei, Towards making the most of cross-lingual transfer for zero-shot neural machine translation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 142–157.

[41] B. Ji, Z. Zhang, X. Duan, M. Zhang, B. Chen, W. Luo, Cross-lingual pre-training based transfer for zero-shot neural machine translation, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 115–122.

[42] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020 pp. 7871–7880, Online.

[43] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (140) (2020) 1–67.

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

[45] M. Artetxe, G. Labaka, E. Agirre, A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 789–798.

[46] M. Eric, R. Goel, S. Paul, A. Sethi, S. Agarwal, S. Gao, A. Kumar, A.K. Goyal, P. Ku, D. Hakkani-Tür, MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines, in: Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, European Language Resources Association, 2020, pp. 422–428.

[47] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, la, USA, May 6-9, 2019, 2019, OpenReview.net.

[48] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, A.M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2020, pp. 38–45, Online.

[49] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, 2020, arXiv e-prints, arXiv:2008.00401.