Mixture-of-languages Routing for Multilingual Dialogues

JIAHUAN PEI[*], Vrije Universiteit Amsterdam, The Netherlands
GUOJUN YAN[†], China Academy of Engineering Physics, China
MAARTEN DE RIJKE, University of Amsterdam, The Netherlands
PENGJIE REN[‡], Shandong University, China

We consider multilingual dialogue systems and ask how the performance of a dialogue system can be improved by using information that is available in other languages than the language in which a conversation is being conducted. We adopt a collaborative chair-experts framework, where each expert agent can be either monolingual or cross-lingual, and a chair agent follows a mixture-of-experts procedure for globally optimizing multilingual task-oriented dialogue systems. We propose a mixture-of-languages routing framework that includes four functional components, i.e., input embeddings of multilingual dialogues, language model, pairwise alignment between the representation of every two languages, and mixture-of-languages. We quantify language characteristics of unity and diversity using a number of similarity metrics, i.e., genetic similarity, and word and sentence similarity based on embeddings. Our main finding is that the performance of multilingual task-oriented dialogue systems can be greatly impacted by three key aspects, i.e., data sufficiency, language characteristics, and model design in a mixture-of-languages routing framework.

CCS Concepts: • **Computing methodologies → Discourse, dialogue and pragmatics**.

Additional Key Words and Phrases: multilingual systems, task-oriented dialogue systems, collaborative agents, mixture-of-experts

1 INTRODUCTION

How many human languages are there in the world? As of 2019, Ethnologue summarized the most extensive catalog of human languages in the world.[1] It covers 6,909 distinct languages, out of which 230 are spoken in Europe, while 2,197 are spoken in Asia.[2] The ability to retrieve information across language boundaries is a long-standing ambition in the information retrieval community [33]. Substantial progress has been made over the years, with practical systems in place to help overcome language barriers [87].

[*] This work was partially done while the author was working at the University of Amsterdam and Centrum Wiskunde en Informatica.
[†] This work was partially done while the author was working at Shandong University.
[‡] Corresponding author.
[1] https://www.ethnologue.com/
[2] https://www.linguisticsociety.org/content/how-many-languages-are-there-world

Authors' addresses: Jiahuan Pei, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, The Netherlands, ppsunrise99@gmail.com; Guojun Yan, China Academy of Engineering Physics, Mianyang, China, yan_gj@qq.com; Maarten de Rijke, University of Amsterdam, Science Park 900, 1098 XH, Amsterdam, The Netherlands, m.derijke@uva.nl; Pengjie Ren, Shandong University, Qingdao, China, renpengjie@sdu.edu.cn.

ACM 1558-2868/2024/8-ART
https://doi.org/10.1145/3676956

Table 1. Hierarchical classification based on the Ethnologue catalog[3] for English, German, Italian, Spanish, and Thai. The Code column shows the unique identification by ISO 639-3 standards. Classification is the path to a language in the language family trees in Ethnologue.

Language	Code	Classification
English	eng	Indo-European>Germanic>West>English
German	deu	Indo-European>Germanic>West>High German>German>Middle German>East Middle German
Italian	ita	Indo-European>Italic>Romance>Italo-Western>Italo-Dalmatian
Spanish	spa	Indo-European>Italic>Romance>Italo-Western>Western>Gallo-Iberian>Ibero-Romance>West Iberian>Castilian
Thai	tha	Kra-Dai>Kam-Tai>Tai>Southwestern

The need for multilingual access has not disappeared. Roughly 80% of the world population does not speak English [14]. As the ways in which we interact with information evolve, e.g., from documents to questions-and-answers and from single-shot to multi-turn interactions, fundamental questions about multilingual information retrieval resurface. How can we develop multilingual dialogue models to support multiple communities with multiple languages as input and output [26, 91, 92]?

Significant challenges remain before we have effective multilingual dialogue systems. First, multilingual dialogue datasets are quite scarce and face an acquisition challenge. For example, a survey [99] from several years ago reports 63 available dialogue corpora and only 2 of them contain multilingual dialogues (i.e., Verbmobil [5] and DSTC5 [41]), until March 2017. Since then, several publications have released dialogue datasets for training multilingual chitchat [15, 52], and both bilingual [53] and multilingual [21, 36, 48, 71, 98, 103, 107, 108] task-oriented dialogue systems (TDSs). Furthermore, a lack of language experts makes the acquisition of non-English data challenging [26]. For example, in the multilingual natural language understanding (NLU) dataset [98], only 11.7% and 20.0% of the utterances are obtained for Thai and Spanish, respectively, due to a lack of bilingual speakers.

Second, language commonalities and peculiarities are very important. On the one hand, languages have genetic relationships through language evolution. We list the Ethnologue catalog entries of five languages (i.e., English, German, Italian, Spanish, and Thai) in Table 1. English is neither always the best nor the only pivot language to bridge the language gap [17, 83]. On the other hand, the unity and diversity of languages have been encoded into high-dimensional vectors in recent computational linguistics. Figure 2 visualizes the mT5 [120] embeddings of words from two multilingual dialogue benchmark datasets [71, 98], covering the five languages mentioned above. Thai words are clustering independently, while words from European languages are mixed up. We further conduct pairwise comparisons of the European languages in Figure 2. We find that intersecting areas (representing commonalities between languages) and disjoint areas (representing peculiarities of languages) can be preserved at the same time, but their proportions can be very different for different language pairs. For example, English and Spanish are not as clearly separated as the other three language pairs, so the proportion of intersecting areas is larger (see Figure 2).

Last but not least, the majority of TDS models focuses on either multiple language-specific optima [70, 98] or cross-lingual adaptation from English to non-English towards multilingual TDSs (see Table 12 and 13). Very few publications consider improving multilingual performance simultaneously, but simply training models using multilingual data does not always lead to improvements, e.g., multilingual NBT [71] and bilingual mBART [53]

[3] https://www.ethnologue.com/browse/names

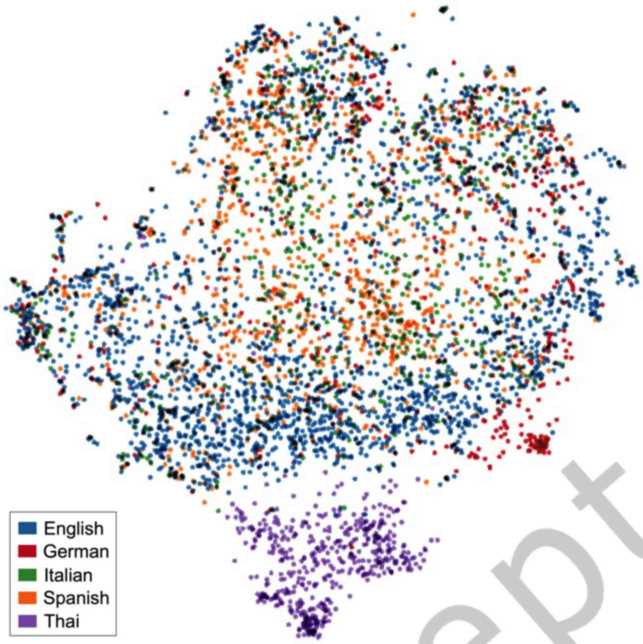


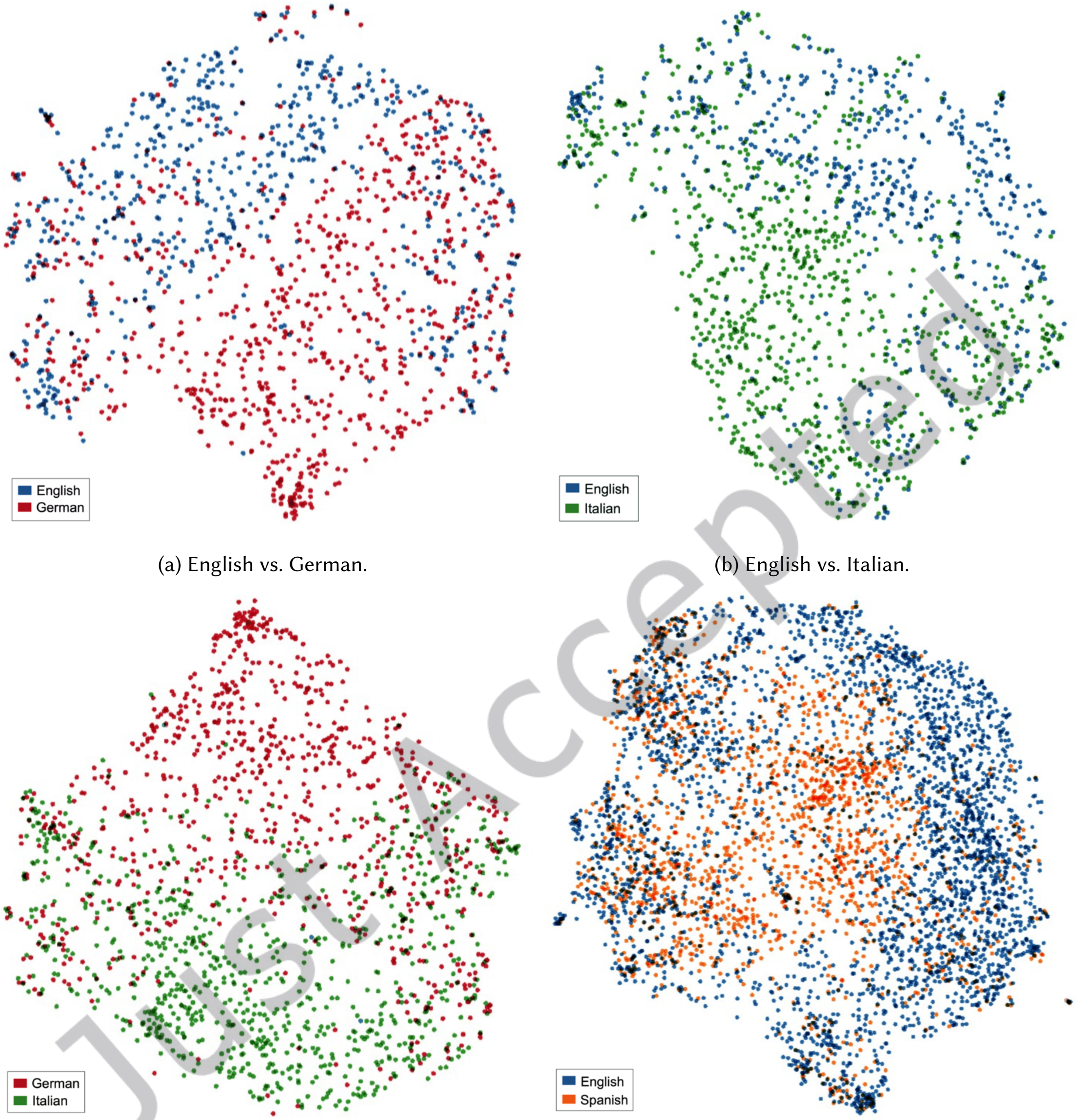
Fig. 1. Visualization of the embeddings of words from two benchmark multilingual dialogue datasets, covering English, German, Italian, Spanish, and Thai. We conduct dimension reduction using the UMAP algorithm [67] and plot all scatter in 2D coordinates using the Tensorflow embedding projector.[4]

are outperformed by their monolingual settings in terms of multiple evaluation metrics. Besides, optimizing all pipeline tasks requires all language-specific annotations, which makes global optimization challenging [91].

In this work, we propose a multilingual dialogue framework that: (i) fully makes use of multilingual data; (ii) captures commonalities between, and peculiarities of, languages; and (iii) improves multilingual performance simultaneously. Figure 3 displays the framework.

We recast the multilingual TDS problem in a collaborative TDS framework [84, 86]: k expert agents account for monolingual and cross-lingual dialogues, and a chair agent conducts a mixture-of-experts for globally optimizing multilingual dialogues. To be more precise, we unify TDS tasks as a standard dialogue generation task and implement a mixture-of-languages routing (MOLR) framework with four functional components, i.e., (i) input embeddings, (ii) a language model, (iii) pairwise alignment between the representations of every two languages, and (iv) mixture-of-languages. For the former two components, we choose mT5 [120] as the backbone of our base model after comparing with pre-trained language baselines [8, 128]. Note that each base model can be either a monolingual or cross-lingual expert agent, and it can flexibly be replaced by other popular multilingual language models such as mBERT [20], mBART [55], and XLM-R [11], etc. Next, we introduce pairwise alignment between the representations of every two languages to bridge the relationship between every two language routes. Here, a *language route* is a path commencing from a source language as the starting point, passing through a pivot language, to a target language as its destination. Language commonalities and peculiarities can be embedded into pairwise alignment states. After that, we conduct global optimization by mixture of languages routing with two collaboration policies, i.e., route-addressing and parameter-sharing. By *mixture of languages routing* we mean

[4] https://projector.tensorflow.org/



(a) English vs. German.

(b) English vs. Italian.

(c) German vs. Italian.

(d) English vs. Spanish.

Fig. 2. Pairwise comparison of the embeddings of words in dialogues from European languages.

the process of learning a combination of routes in the proposed model between or across multiple languages. This setup enables the multilingual dialogue model to automatically learn the pivot languages, rather than fixing

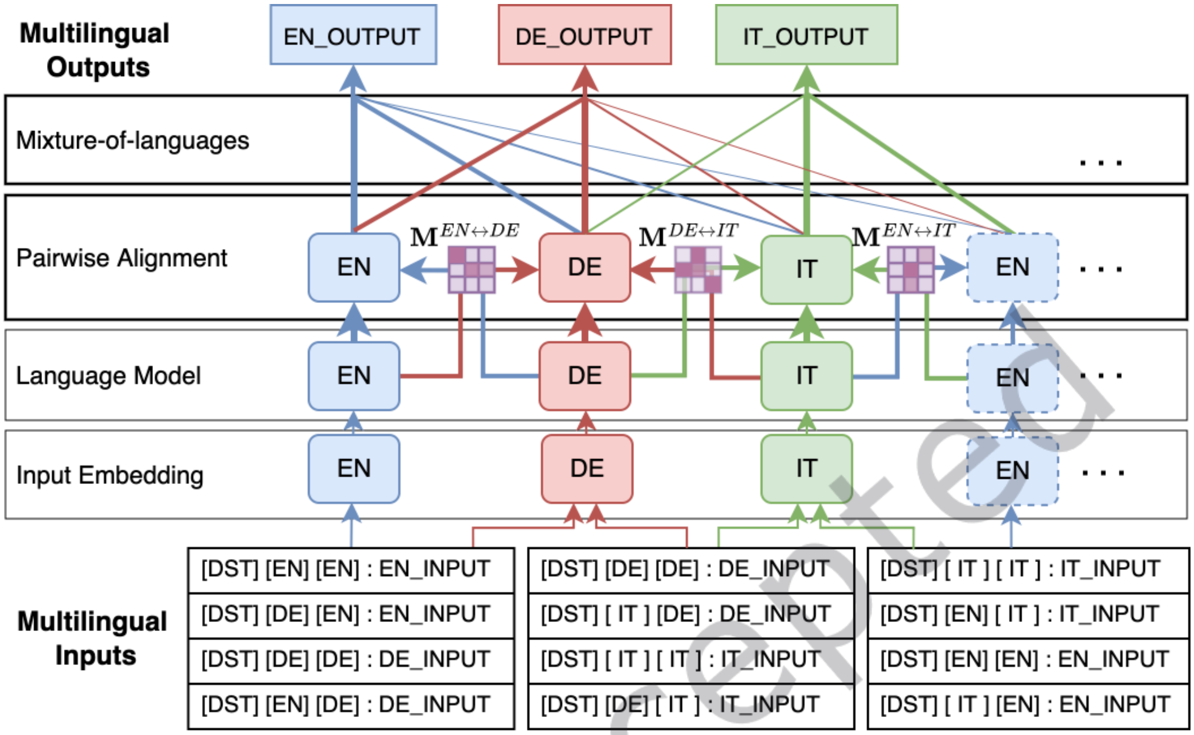

Multilingual Outputs
EN_OUTPUT
DE_OUTPUT
IT_OUTPUT
Mixture-of-languages
. . .
Pairwise Alignment
EN
$\mathbf{M}^{EN\leftrightarrow DE}$
DE
$\mathbf{M}^{DE\leftrightarrow IT}$
IT
$\mathbf{M}^{EN\leftrightarrow IT}$
EN
. . .
Language Model
EN
DE
IT
EN
. . .
Input Embedding
EN
DE
IT
EN
. . .
Multilingual Inputs
[DST] [EN] [EN] : EN_INPUT
[DST] [DE] [EN] : EN_INPUT
[DST] [DE] [DE] : DE_INPUT
[DST] [EN] [DE] : DE_INPUT
[DST] [DE] [DE] : DE_INPUT
[DST] [IT] [DE] : DE_INPUT
[DST] [IT] [IT] : IT_INPUT
[DST] [DE] [IT] : IT_INPUT
[DST] [IT] [IT] : IT_INPUT
[DST] [EN] [IT] : IT_INPUT
[DST] [EN] [EN] : EN_INPUT
[DST] [IT] [EN] : EN_INPUT

Fig. 3. The framework of mixture-of-languages routing (MOLR) in multilingual TDSs. Taking the dialogue state tracking (DST) task as an example, the raw inputs are extended with prefixes and processed into monolingual and cross-lingual data, respectively. The rounded rectangles represent intermediate language-specific representations. A learnable matrix $\mathbf{M}^{a\leftrightarrow b}$ is used to transform the presentations between any two languages l_a, l_b. The blue, red, and green arrows represent the routes of English (EN), German (DE), and Italian (IT), respectively.

English as the only pivot language. Moreover, the unified generation framework equips the proposed model with the ability to optimize multiple subtasks, simultaneously.

To assess the effectiveness of the proposed mixture-of-languages routing (MOLR) framework, we conduct extensive experiments on two benchmark datasets, i.e., the multilingual DST dataset [71] and the NLU dataset [98]. We find that bilingual and multilingual MOLR models are on par with, and even outperform, state-of-the-art baselines for both multilingual DST and NLU tasks. At best, compared with mT5, the proposed MOLR models improve 2.31%/2.56%/0.67% of joint goal accuracy for English/German/Italian on the DST task, and 0.13%/1.89%/5.53% of slot F1 for English/Spanish/Thai on the NLU task. Note that most of the baselines conduct classification over the predefined task-related label space; in contrast, we generate all the labels from the vocabulary space.

The larger prediction space increases the difficulty of tasks, but the benefits are obvious: our framework is able to predict values that are not predefined and is applicable to all dialogue tasks in a unified way.

The main contributions of this work are as follows:

- We propose a mixture-of-languages routing (MOLR) framework that is able to globally and simultaneously optimize the multilingual task-oriented dialogue system (TDS) performance. MOLR benefits from multilingual data argumentation, language characteristic modeling, mixture-of-language routing.

- We develop generation baselines that are at least on par with the-state-of-the-art classification baselines.
- We carry out a large number of contrastive experiments and deep-dive analyses, which reveal the effectiveness of the MOLR framework and help understand its effectiveness.
- We find that it is better to gradually cross the language chasm: a larger degree of similarity between source language and pivot language is usually helpful for the overall performance.

2 RELATED WORK

Given the challenges of multilingual TDSs, we summarize related work from three points of view: (i) data, (ii) language, and (iii) model.

2.1 Multilingual data augmentation

Data augmentation has been widely used for alleviating data scarcity problems in multilingual dialogues [93]. On the one hand, data augmentation targets better representation of dialogues. Zhao et al. [126] use atomic templates to produce exemplars from dialogue acts, followed by a sentence generator to complete the whole utterance. Louvan and Magnini [60] involve simple text and syntax substitutions, and combine them with pre-trained language models. Yin et al. [123] replace text spans with paraphrases and use reinforcement learning to control the quality of the augmented data. Pei et al. [85] search nearest neighbor dialogues as supplements to a current dialogue to alleviate the scarcity of user preferences. Yan et al. [121] introduce heuristic approaches to generate data and adopt contrastive learning to further improve the overall performance. Most recent work usually benefits from monolingual pre-trained language models (e.g., BERT [20] and GPT-2 [89]). Researchers have also explored retrieval-based approaches to expand the scale of dialogues [30]. Pei et al. [85] utilize retrieved neighbor dialogues to enrich user profiles to improve the performance of dialogue response selection. Xu et al. [118] learn the relation of neighboring elements and phrasal patterns to extend long-range dependencies in dialogues. Li et al. [49] extract personalized wording from user-specific dialogue history as extra matching information to improve retrieval-based dialogue systems. Ren et al. [95] involve search engine result pages to generate conversational responses for answering complex information needs. Yan et al. [122] consider multiple responses to enhance diversity of retrieval-based conversations by dynamic representation learning. Ling et al. [54] generate diverse relevant and informative questions for improving interactiveness and persistence of human-machine interactions.

On the other hand, data augmentation aims to bridge language gaps. Dominant code-switching methods [43, 88] translate sentences in English into randomly selected target languages, which enables them to fine-tune multilingual transformers with generalization ability across languages. XeroAlign [27] introduces an auxiliary loss function based on machine translation and jointly optimizes the overall performance with the primary task. Kaliamoorthi et al. [40] conduct knowledge transfer during distillation from a pre-trained mBERT teacher to a tiny student model. Mrkšić et al. [71] learn specialized cross-lingual vector spaces by multilingual data training enhanced with semantic relations from lexical resources. Most recent work crosses the language chasm using multilingual pre-trained language models (e.g., mBERT [20, 21, 36], XLM-R [11, 36, 48, 128], mT5 [108, 129]).

Similar to most recent work [108, 129], we choose a state-of-the-art multilingual pre-trained language model (i.e., mT5) as our backbone for both better dialogue representation and language transfer. But unlike the above approaches, (i) we generate pairwise language routes and focus on how to learn the relationships between language pairs, and (ii) we aggregate language routes for global optimization of multilingual TDSs.

2.2 Unity and diversity of languages

In bioscience [24] and linguistic studies [22, 25, 105, 109], both unity and diversity play key roles for cross-linguistic variation in human languages.

Over time, languages generate biological or genetic relationships [29]. Linguists and language institutions have conducted large-scale studies on language affinity[5] and the Ethnologue catalog.[6] Generally, a language family tree is a common way to interpret genetic relationships that can reveal the unity and diversity of languages [104]. Their basic assumption is that two languages belong to the same language family if they are from a common ancestor, or one is descended from the other.

In modern linguistics, important research topics include universal grammar [18, 75] and linguistic typology [79, 109, 114]. The former focuses on unity, in which all languages are treated as universal components of the language faculty [18]. This is the theoretical basis of research on part-of-speech tagging [6, 73], chunking [61, 64], and syntactic parsing [62, 68]. The latter emphasizes diversity, which captures the structural differences of languages, as the principal bridge, to discover universals [19]. Morphology is usually diverse across languages, and it is hard to find universals for traditional linguistic typology [97]. The world language tree is constructed based on Levenshtein distances, which define the average number of edits needed to convert a source language to a target language [72].

Language similarity has been a commonly used metric to quantitatively measure unity and diversity in recent computational linguistics [2, 7, 112]. One branch of work measures language similarity by their structural properties [16]. Bjerva et al. [4] define language similarity based on language structures, i.e., phrase structure trees and dependency relations. Oco et al. [78] compute Dice's coefficient to measure the similarity of eight Philippine languages based on the language family tree in the Ethnologue. However, these approaches do not apply when the structure is not available. Another branch of work measures language similarity as lexical overlap between languages based on handcrafted cognates [77] or automatically extracted cognates [100]. Beinborn et al. [3] identify cognates based on character-based machine translation. However, their methods cannot compare the similarity of cognates without translation relationship (e.g., English "father" and the Italian "padre") [2]. To this end, most recent work encodes natural languages into high dimensional vectors namely embeddings, e.g., word embeddings [96, 102] and word-based syntax embeddings [50] and pretrained language models [42, 81]. Therefore, unity and diversity of languages can be measured using similarity and dissimilarity of embeddings.

In this work, we conduct an analysis of multilingual TDS results from the point of the view of language characteristics, i.e., the unity and diversity of languages. We compare commonalities and specifications of languages using multiple aspects, including visualization of word embeddings, as well as genetic and embedding-based similarity metrics.

2.3 Multilingual TDS models

Monolingual TDSs have made considerable progress as reported in a large number of recent publications [9, 74, 99, 124]. Many recent studies have built new datasets and/or tasks to advance research on multilingual TDSs [21, 36, 80, 119]. However, it is hard to fairly compare with the majority of approaches because they do not report results on those all datasets [91]. Besides, from a technical perspective, understanding semantics is the heart of any dialogue systems [35]. Thus, in this work, we mainly focus on a comparison of cross-lingual and multilingual models on two commonly-used semantic understanding tasks (i.e., DST [71] and NLU [98]).

2.3.1 Cross-lingual models. Existing cross-lingual models mainly consider two key factors: dialogue representation and cross-lingual transfer. To conduct better language modeling, previous studies utilize variants of sequential models. Upadhyay et al. [107] jointly train bilingual embeddings with a biRNN model for few-shot cross-lingual NLU. Liu et al. [56] equip a biLSTM model with latent variables and word pairs to refine the aligned cross-lingual word embeddings. Schuster et al. [98] deploy a biLSTM-CRF model, where the cross-lingual transfer comes from sharing between the biLSTM and CRF layer across languages. Liu et al. [57] develop a biLSTM, transformer, and

[5] http://www.linguaechristi.org/people-groups/
[6] https://www.ethnologue.com/browse/names

mBERT for sequence labeling tasks, and find that removing the word order can improve cross-lingual performance. Several researchers generate code-switching sentences to enable cross-lingual capabilities, by either replacing words [37, 58] or sentences [88] in target languages. MultiATIS++ [119] learns slot alignment based on an mBERT encoder, machine translation, and label projection. GlobalWoZ [21] introduces several data augmentation baselines for zero-shot and few-shot cross-lingual learning on the proposed dataset. Siddhant et al. [101] gain cross-lingual transfer capabilities by representations from a multilingual neural machine translation encoder. Gritta and Iacobacci [27] use an auxiliary translation-based loss function to jointly learn with the primary task. Xiang et al. [116] inject multi-granularity translation-based noise to improve the robustness of cross-lingual task-oriented dialogues. Hung et al. [36] finetune the XLM-R model with English and target languages in zero-shot and few-shot transfer settings. Li et al. [48] provide several multilingual pretrained benchmarks such as XLM-R and mBAR, and evaluate them on the multilingual ATIS and MTOP datasets for task-oriented semantic parsing. Very recently, Zuo et al. [129] have applied mT5 with meta-learning on their unpublished dataset and Van et al. [108] simply use mT5 as a state-of-the-art benchmark in their Vietnamese task-oriented dialogue dataset.

To sum up, the proposed methods enable transfer across languages using a variety of techniques, including cross-lingual word embeddings [107], multilingual knowledge distillation [10], transferable latent variables [56], code-swtching [37, 58, 88], word alignment [58, 119], and machine translation [27, 101, 116]. Most recent work benefits from these techniques and from pre-trained multilingual language models such as mBERT [20, 21], XLM-R [36, 48, 128], and mT5 [108, 129]. However, many terms in low resource languages are not in the vocabulary of pre-trained language models [63]. So cross-lingual transfer techniques (e.g., pairwise alignment of language states) are still necessary in the presence of pre-trained language models.

2.3.2 Multilingual models. Only few previous studies target multilingual TDS models. An intuitive solution is to train a single model on combined multilingual datasets and evaluate the model on test data for all languages, respectively. Mrkšić et al. [71] use constraints from monolingual and cross-lingual synonymy and antonymy to finetune multilingual word embedding spaces and apply them to the DST task. Schuster et al. [98] use a multilingual translation-based biLSTM encoder to learn contextual word representations, evaluating on multiple languages. GlobalWoZ [21] introduces several data augmentation baselines for zero-shot and few-shot cross-lingual learning on the proposed dataset. Ding et al. [21] monolingual and cross-lingual use cases are parts of multilingual TDSs and optimize for each use case separately. Recent work by Zuo et al. [129] reports a benchmark of mT5 with meta-learning [23], however, neither the dataset nor the source code of the model is publicly accessible.

None of the released models has modeled language relationships or conducted global optimization for multilingual TDSs on public datasets, to the best of our knowledge. Unlike the majority of classification models, the proposed generation model (i.e., MOLR) achieves competitive performance and is able to predict out-of-ontology slot values as in [44, 115, 117].

Large language models (LLMs) have become more and more prevailing in the past few months. We have been aware of this and are actively working towards extending MOLR by integrating open-sourced decoder-only LLMs. This initiative facilitates the seamless adoption of the most recent LLMs. OpenChat3.5 [111] is a multilingual chat model fine-tuned with the C-RLFT strategy on mixed-quality data, achieving performance comparable to larger models like ChatGPT. BLOOM [46] is pretrained on the multilingual ROOTS corpus, offering multilingual capabilities for various natural language processing tasks. Llama2-Chat [106] is a pretrained and fine-tuned generative text model optimized specifically for multilingual dialogue tasks, ensuring high-quality conversational responses. The sparse mixture-of-experts (MOE) language model Mixtral 8x7B [38] showcases notable enhancements over its predecessor, Mistral 7B. Notably, the number of learnable parameters inevitably increases 8-fold compared to its predecessor. We extend MOLR with decoder-only LLM backbones, however, the increasing number of parameters is only from pairwise-align layers, not growing exponentially.

3 COLLABORATIVE MULTILINGUAL DIALOGUE FRAMEWORK

3.1 A unified task-oriented dialogue system

A dialogue consists of multiple turns between a user and a system. At the t-th turn, the user provides an utterance U_t, and the system produces a response R_t as a reply. To get high-quality responses R_t, a TDS is usually decomposed into four subtasks: natural language understanding (NLU), dialogue state tracking (DST), dialogue policy learning (DPL) and natural language generation (NLG).

In this work, we unify a TDS with a neural network $f_\theta(\cdot)$ parameterized by θ, which generally contains (i) an input embedding, (ii) hidden states encoding, and (iii) output projection layers. This neural network works with all subtasks in an end-to-end fashion. Specifically, we formulate the four subtasks as follows.

3.1.1 *Natural language understanding (NLU).*

The NLU task is the key component of a task-oriented dialogue system responsible for extracting the meaning or intent from user utterances expressed in natural language. It involves parsing and interpreting user inputs to identify the user's intention and extract relevant entities necessary for completing the task. The types of the relevant entities are slots. Given a current user utterance U_t as input, the model outputs intents I_t and slots S_t by:

$$I_t, S_t = f_\theta(U_t). \tag{1}$$

3.1.2 *Dialogue state tracking (DST).*

The DST task serves to maintain an internal representation of the current state, i.e., belief state, of the dialogue based on the information exchanged between the user and the system. It involves tracking the user's goals, preferences, constraints, and other relevant information throughout the conversation. Given a dialogue history $C_t = [U_1, S_1, \ldots, U_t]$ as input, the model outputs a belief state B_t by:

$$B_t = f_\theta(C_t), \tag{2}$$

which can be denoted as a set of triples representing slot-value pairs for a specific domain: (domain, slot_name, value).

3.1.3 *Dialogue policy learning (DPL).*

The DPL task aims to the process of learning optimal policies for managing the dialogue flow and making decisions on system actions. It involves learning a policy that maps the current dialogue state to the most appropriate system action, considering the system's goals and user preferences. Given dialogue history C_t, belief states B_t, and retrieval records from database D_t as input, the DPL outputs system actions by:

$$A_t = f_\theta([C_t; B_t; D_t]), \tag{3}$$

which is a list of triples representing as (domain, action_type, slot_name).

3.1.4 *Natural language generation (NLG).*

NLG task refers to the process of generating natural language responses or utterances based on the dialogue context, such as dialogue history, dialogue state, and system actions. It aims to generate coherent and fluent natural languages that can be communicated to the user. Given dialogue history C_t, belief states B_t, retrieval records from database D_t, and system actions A_t as input, the model outputs a response R_t by:

$$R_t = f_\theta([C_t; B_t; D_t; A_t]). \tag{4}$$

To unify the above subtasks, we tackle them as a sequence-to-sequence generation task [32]. The input of all tasks is a sequence of tokens that are aggregated from the concatenation of input sources, i.e., $[U_t]$, $[C_t]$, $[H_t; B_t; D_t]$ $[H_t; B_t; D_t; A_t]$ for NLU, DST, DPL, NLG, respectively.

3.2 Monolingual and cross-lingual expert agents

We use mT5 [120] as our backbone following conditional causal language modeling [121], which adopts a transformer-based encoder-decoder model to learn a mapping f from an input sequence $\mathbf{X}_{1:n} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ to a target sequence $\mathbf{Y}_{1:m} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m)$, i.e., $f_{\theta_{\text{enc}},\theta_{\text{dec}}} : \mathbf{X}_{1:n} \rightarrow \mathbf{Y}_{1:m}$, by the following conditional probability distribution:

$$p_{\theta_{\text{enc}},\theta_{\text{dec}}}(\mathbf{Y}_{1:m} \mid \mathbf{X}_{1:n}). \tag{5}$$

For each input sequence, the encoder converts $X_{1:n}$ to the corresponding hidden states $\tilde{\mathbf{X}}_{1:n} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_n)$, the encoder is represented as $f_{\theta_{\text{enc}}} : \mathbf{X}_{1:n} \longrightarrow \tilde{\mathbf{X}}_{1:n}$, formally, the probability can be computed as:

$$p_{\theta_{\text{enc}}}(\tilde{\mathbf{X}}_{1:n} \mid \mathbf{X}_{1:n}). \tag{6}$$

Mathematically, the decoder learns the probability distribution of $Y_{1:m}$ given $H_{1:n}$, i.e., $p_{\theta}(\mathbf{Y}_{1:m} \mid \tilde{\mathbf{X}}_{1:n})$. Using Bayes's rule, the distribution can be decomposed into conditional distribution over the vocabulary $\mathcal{V}$ at the j-th timestamp token in the target sequence by:

$$p_{\theta_{dec}}(\mathbf{Y}_{1:m} \mid \tilde{\mathbf{X}}_{1:n}) = \prod_{j=1}^{m} p_{\theta_{dec}}(\mathbf{y}_j \mid \mathbf{Y}_{0:j-1}, \tilde{\mathbf{X}}_{1:n}), \tag{7}$$

where $\mathbf{y}_0$ denotes the 0-th target vector that represents the vector of the special "begin-of-sentence" token [BOS]. The model can be learned by minimizing the cross-entropy loss as follows:

$$\mathcal{L}_{expert} = -\sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{y}_j^i \log p_{\theta}(\mathbf{y}_j^i \mid \mathbf{Y}_{0:j-1}^i, \tilde{\mathbf{X}}_{1:n}^i), \tag{8}$$

where N denotes the batch size and n_i denotes the length of the i-th target sequence.

For a monolingual agent, both the input sequence $\mathbf{X}_{1:n}$ and the target sequence $\mathbf{Y}_{1:m}$ are in the same language. For a cross-lingual agent, the input sequence $\mathbf{X}_{1:n}$ and the target sequence $\mathbf{Y}_{1:m}$ are from two different languages.

3.3 Multilingual agents with mixture-of-languages routing

We introduce the workflow of the mixture-of-languages routing (MOLR) model as shown in Figure 3, considering the DST task as an example. First, we follow T5's modeling of prefix and use "[TASK]" as the class label [90] and extend each raw input with a task-specific prefix in the following format:

[TASK] [Pivot-language] [Target-language]: [Source-language-input]

Note that if [Pivot-language] and [Target-language] are identical, then the processed data is monolingual data, otherwise it is cross-lingual data. Then, the processed inputs pass through the input embedding layers followed by a language model, and they are transformed into language-specific hidden states. Next, MOLR uses a learnable matrix to conduct pairwise alignment for every two language-specific hidden states. Last, MOLR adopts mixture-of-languages policies to integrate all states from multiple routes between or across multiple languages. To be more specific, we implement the input embeddings layers and the language model based on mT5, the state-of-the-art pretrained language model and introduce pairwise alignment and mixture-of-languages routing as follows.

3.3.1 Pairwise alignment. Recall that in Eq. 7, the k-th monolingual model outputs the probability over the vocabulary $\mathcal{V}$ at the j-th timestamp by:

$$\begin{aligned} p^{k}_{\theta_{dec}}(\mathbf{y}_j \mid \mathbf{Y}_{0:j-1}, \tilde{\mathbf{X}}_{1:n}) &= \mathrm{softmax}(f_{\theta_{\mathrm{dec}}}(\mathbf{Y}_{0:j-1}, \tilde{\mathbf{X}}_{1:n})) \\ &= \mathrm{softmax}(\psi_{\theta_{\mathrm{task}}}(\tilde{\mathbf{y}}^{k}_{j})) \\ &= \mathrm{softmax}(\mathbf{W}^{\top}_{\mathrm{emb}}\tilde{\mathbf{y}}^{k}_{j}) \\ &= \mathrm{softmax}([\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{|\mathcal{V}|}]^{\top}\tilde{\mathbf{y}}^{k}_{j}), \end{aligned} \tag{9}$$

where softmax(·) is the activation function that scales logits into probabilities, $\tilde{\mathbf{y}}^{k}_{j} \in \mathbb{R}^{d}$ represents the decoded hidden state at the j-th timestamp from a language model. Here, we use a pre-trained language model mT5. $\psi_{\theta_{\mathrm{task}}}$ denotes the task layer and $\mathbf{W}_{\mathrm{emb}} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{|\mathcal{V}|}] \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the word embedding matrix.

Given any two monolingual models for languages l_a, l_b, the hidden states can be denoted as $\tilde{\mathbf{y}}'^{a}_{j}$ and $\tilde{\mathbf{y}}'^{b}_{j}$. Given a learnable matrix $\mathbf{M}^{a \to b}_{j} \in \mathbb{R}^{d \times d}$ that transforms the decoded hidden states from $\mathbf{y}^{a}_{j}$ to $\tilde{\mathbf{y}}'^{b}_{j}$, and vice versa, formally, we can denote the pairwise alignment as:

$$\begin{aligned} \tilde{\mathbf{y}}'^{b}_{j} &= \mathbf{M}^{a \to b}_{j}\tilde{\mathbf{y}}^{a}_{j} \in \mathbb{R}^{d}, \\ \tilde{\mathbf{y}}'^{a}_{j} &= \mathbf{M}^{b \to a}_{j}\tilde{\mathbf{y}}^{b}_{j} \in \mathbb{R}^{d}. \end{aligned} \tag{10}$$

The benefit of this transition is that we can learn the hidden state of language b even though we only have the training data of language a and vice versa.

3.3.2 Mixture-of-languages routing. To learn from a mixture of language routes, we utilize two collaboration policies, i.e., route-addressing and parameter-sharing.

Route-addressing. Let $H = [\tilde{\mathbf{y}}^{a}_{j}; \tilde{\mathbf{y}}'^{a}_{j}; \tilde{\mathbf{y}}^{b}_{j}; \tilde{\mathbf{y}}'^{b}_{j}; \dots] \in \mathbb{R}^{l \times d}$ ($\frac{l}{2}$ is the number of languages and d is the dimension), and $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$ be the matrices for query Q, key K, and value V. Each H is associated with a query Q and a key-value pair (K, V). The computation of an attentive representation A of $\mathbf{y}_j$ in the self-attention is:

$$\begin{aligned} &Q = \mathbf{W}_q H \in \mathbb{R}^{l \times d}, K = \mathbf{W}_k H \in \mathbb{R}^{l \times d}, V = \mathbf{W}_v H \in \mathbb{R}^{l \times d}, \\ &A = \mathrm{softmax}(\alpha^{-1} Q K^{\top}) \in \mathbb{R}^{l \times l}, \\ &\tilde{\mathbf{y}}_j = \phi(AV) \in \mathbb{R}^{d}, \end{aligned} \tag{11}$$

where H is the attended output and A is the attention distribution that attends to V, α is a scaling factor, and ϕ is a linear layer followed by the accumulation of attended values, parameterized by θ.

Parameter-sharing. For the same task and the same language, all the model parameters are shared, otherwise only the parameters θ_{task} in a task layer ψ (see Eq. 9) are not shared, and the other parameters in the model are shared. In the shared modules, we aim to learn a common space representation for all tasks. This policy serves as regularization and alleviates the overfitting problem, as the model learns a representation that generalizes to all tasks.

3.4 Extension of decoder-only large language models

With the emergence of LLMs and the prevalence of applications such as ChatGPT, decoder-only architectures have become highly popular [28, 125]. We extend the proposed framework by incorporating decoder-only LLMs as backbones that adhere to causal language modeling. We follow the LLM's routine and treat it as an instruction-following task. First, we prepare the raw data with the following format:

Instruction: [Route-id][Pivot-language][Target-language][Task-prompt]
Input: [Raw-input]
Output: [Raw-output]

where "[Route-id]" serves as the special token used to identify each type of language route. It plays a crucial role in determining the information flow, ensuring effective learning by considering both specification and generalization of linguistic features." "[Task-prompt]" refers to an instance of a prompt designed to lead LLMs to accomplish a specific task. The details of the prompt are listed in Table 11, Appendix A. Similarly, the processed inputs pass through the input embedding layers followed by a decoder-only LLM, and they are transformed into language-specific hidden states. Last, MOLR employs a mixture-of-languages policies to integrate states from multiple routes, whether between or across multiple languages. More specifically, we implement the input embedding layers and a state-of-the-art decoder-only LLM (e.g., LLaMa2-Chat [106], Bloom [46], OpenChat3.5 [111]) capable of supporting the candidate languages. Subsequently, we introduce pairwise alignments and mixture-of-languages routing as enhancements. Then, we perform parameter-efficient fine-tuning (PEFT) with low-rank adapters (LoRAs) [34] on a LLM using the aforementioned data.

4 EXPERIMENTAL SETUP

4.1 Research questions

We seek to answer the following questions in the experiments:

(**RQ4.1**) Does the mixture-of-languages routing (MOLR) model improve the performance of monolingual and multilingual models?

(**RQ4.2**) How do language characteristics influence the performance of MOLR models? (i) How to qualitatively analyze language unity and diversity? (ii) How to quantify language unity and diversity? (iii) How do language unity and diversity influence the mixture-of-languages?

(**RQ4.3**) How do the key components influence the performance of MOLR models? (i) How do different combination policies influence the MOLR model? (ii) How do different number of layers of expert agents influence the gains of the MOLR model?

(**RQ4.4**) Can the MOLR framework be effectively adapted to the new era of decoder-only LLMs?

4.2 Datasets and evaluation

We conduct a large number of experiments on two benchmark datasets for the following multilingual TDS tasks to fairly compare with the majority of prior approaches [91].

4.2.1 Dialogue state tracking (DST) and natural language generation (NLG). The multilingual DST dataset [71] is extended from the WOZ 2.0 dataset [113] by manually translating English into Italian and German, respectively. For each language, the dataset contains 1200 multiple-turn dialogues in the restaurant domain, and it is split into 600, 200, 400 dialogues for training, validation, and testing. The dataset contains 4 types of goal-related slots: 3 informing slots (i.e., food, price range and area) to track a user's search constraints, and 1 request slot (i.e., request) to track a user's questions about the search results. The evaluation metrics for the DST task are:

- *Joint goal accuracy*, which measures the proportion of dialogue turns where all search constraints exactly match the ground truth on the test set.
- *Request accuracy*, which represents the proportion of dialogue turns where all the user questions are recognized correctly.

The evaluation metrics for the NLG task are:

- *BLEU-4*, which measures precision and calculates the ratio of 4-grams in the generated responses that match those in the reference responses.
- *ROUGE-L*, which measures recall and calculates the ratio of longest common subsequences in the reference responses that are captured by the generated responses.

4.2.2 Natural language understanding (NLU). The multilingual NLU dataset [98] consists of 43k, 8.6k, and 5k single-turn dialogues in English, Spanish, and Thai, respectively, covering 3 domains (weather, alarm, and reminder). The dataset has 12 types of intents and 11 types of slots. The evaluation metrics are:

- *Intent accuracy*, which indicates the proportion of the correctly identified intents.
- *Slot F1*, which is the geometric mean of the precision and recall for slot filling.

4.2.3 Language similarity metrics. We propose language similarity metrics to compare the similarity of any two languages α, β from a genetic and semantic point of view. A higher degree of similarity denotes a higher degree of language unity, while a smaller degree of similarity denotes a higher degree of language diversity. To compare phylogenetic relationships, the Robinson-Foulds distance is the most widely used metric [82]. To measure semantic similarity, word and sentence embeddings are widely used in modern NLP tasks [7].

- *Genetic similarity*, which defines the similarity of any two languages based on their Robinson-Foulds distance (RFD)[7] in language family trees. Here we define it as $\phi_{genetic}(\alpha, \beta) = \frac{1}{\text{RFD}(\alpha,\beta)+1}$ if they have at least one ancestor, otherwise $\phi_{genetic}(\alpha, \beta) = 0$. $\text{RFD}(\cdot, \cdot)$ counts the number of unique entries that are not in common in the classification based on the Ethnologue catalogue (see Table 1).
- *Word similarity*, which measures the parallel degree of two languages using the cosine similarity of the centroid word embeddings of the datasets, i.e., $\phi_{word}(\alpha, \beta) = \cos(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$. We compute the centroid of word embeddings $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ as the mean of all word embeddings.
- *Sentence similarity*, which measures the parallel degree of two languages as the cosine similarity of the centroid sentence embeddings in the datasets, that is, $\phi_{sentence}(\alpha, \beta) = \cos(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$. A language can be represented by the mean of all sentence embeddings in a dataset. We compute the centroid of word embeddings $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ as the mean of all sentence embeddings. Here we use the embedding of the "[TASK]" token at the beginning of in a sentence as its sentence embedding.

4.3 Language routes of mT5 and variants

Recall that a language route is a path starting from a source language to a target language, passing through a pivot language. We format all language routes of mT5 and its variants in Table 2. The notation $I(\cdot)$, $H(\cdot)$, $T(\cdot)$, and $M(\cdot)$ indicates input layers, hidden layers, task layers, and mapping layers (see Section 3.2 and 3.3). In this work, we develop $I(\cdot)$ and $H(\cdot)$ with a language model (Eq. 5) and $M(\cdot)$ for pairwise alignment (Eq. 10). To be more specific, we have the following types of language routes for mT5 and its variants:

- **mT5**: single language route for monolingual models.
- **mT5+bDA**: double language routes for training a single model with bilingual data.
- **mT5+bMOLR**: quadruple language routes for training a bilingual model with bilingual data. There are two monolingual routes and two cross-lingual routes.
- **mT5+bDA**: multiple language routes for training a single model with multilingual data. Here we use triple language routes.
- **mT5+bDA**: multi-hop quadruple language routes for training a multilingual model with multilingual data in multiple stages. Here we use two stages of quadruple language routes. The model from Hop1 is

[7] https://en.wikipedia.org/wiki/Robinson%E2%80%93Foulds_metric

Table 2. Language routes of the proposed models given any three languages α, β, γ with DST or NLU as a [TASK].

Model	**Setting**	**Language routes**
mT5	Single language route	$[\text{TASK}][\alpha][\alpha]: I(\alpha) \rightarrow H(\alpha) \rightarrow T(\alpha)$
mT5+bDA	Double language routes	$[\text{TASK}][\alpha][\alpha]: I(\alpha) \rightarrow H(\alpha) \rightarrow T(\alpha)$ $[\text{TASK}][\beta][\beta]: I(\beta) \rightarrow H(\beta) \rightarrow T(\beta)$
mT5+bMOLR	Quadruple language routes	$[\text{TASK}][\alpha][\alpha]: I(\alpha) \rightarrow H(\alpha) \rightarrow T(\alpha)$ $[\text{TASK}][\beta][\beta]: I(\beta) \rightarrow H(\beta) \rightarrow T(\beta)$ $[\text{TASK}][\beta][\alpha]: I(\alpha) \rightarrow H(\beta) \rightarrow M(\beta \rightarrow \alpha) \rightarrow T(\alpha)$ $[\text{TASK}][\alpha][\beta]: I(\beta) \rightarrow H(\alpha) \rightarrow M(\alpha \rightarrow \beta) \rightarrow T(\beta)$
mT5+mDA	Multiple language routes	$[\text{TASK}][\alpha][\alpha]: I(\alpha) \rightarrow H(\alpha) \rightarrow T(\alpha)$ $[\text{TASK}][\beta][\beta]: I(\beta) \rightarrow H(\beta) \rightarrow T(\beta)$ $[\text{TASK}][\gamma][\gamma]: I(\gamma) \rightarrow H(\gamma) \rightarrow T(\gamma)$
mT5+mMOLR	Multi-hop language routes	Hop 1: $[\text{TASK}][\gamma][\gamma]: I(\gamma) \rightarrow H(\gamma) \rightarrow T(\gamma)$ $[\text{TASK}][\beta][\beta]: I(\beta) \rightarrow H(\beta) \rightarrow T(\beta)$ $[\text{TASK}][\beta][\gamma]: I(\gamma) \rightarrow H(\beta) \rightarrow M(\beta \rightarrow \gamma) \rightarrow T(\gamma)$ $[\text{TASK}][\gamma][\beta]: I(\beta) \rightarrow H(\gamma) \rightarrow M(\gamma \rightarrow \beta) \rightarrow T(\beta)$ Hop 2: $[\text{TASK}][\alpha][\alpha]: I(\alpha) \rightarrow H(\alpha) \rightarrow T(\alpha)$ $[\text{TASK}][\beta][\beta]: I(\beta) \rightarrow H(\beta) \rightarrow T(\beta)$ $[\text{TASK}][\beta][\alpha]: I(\alpha) \rightarrow H(\beta) \rightarrow M(\beta \rightarrow \alpha) \rightarrow T(\alpha)$ $[\text{TASK}][\alpha][\beta]: I(\beta) \rightarrow H(\alpha) \rightarrow M(\alpha \rightarrow \beta) \rightarrow T(\beta)$

used as a pre-trained model for Hop2. A "hop" refers to a set of language routes between two languages, a stage in a multi-stage process of training a multilingual model using multilingual data (see Table 2).

4.4 Baselines

For the DST task, we consider four groups of baselines, depending on the base model that they use: (i) based on neural belief tracker (NBT), (ii) based on global-locally self-attentive dialogue state tracker (GLAD), (iii) based on bidirectional encoder representations from transformers (BERT), and (iv) based on cross-lingual language model pretraining (XLM). The selection is based on recent DST models that regard English, German, and Italian as target languages, and report comparable results on the multilingual DST dataset [71].

For the NLU task, we also consider four groups of baselines, depending on the base model that they use: (i) based on recurrent neural networks (RNNs), (ii) based on transformers, (iii) based on bidirectional encoder representations from transformers (BERT), and (iv) based on cross-lingual language model pretraining (XLM). The selection is based on recent NLU models that regard English, Spanish, and Thai as target languages, and report comparable results on the multilingual DST dataset [98].

4.5 Implementation details

We use a pre-trained model mT5-small from the Huggingface library.[8] It consists of 8 layers of transformer blocks for both encoders and decoders. Each attention module has 6 attention heads and the scaling factor α is 1. The total number of parameters is about 300 million. We set the training epochs to 60. We use the AdamW

[8] https://huggingface.co/google/mt5-small

optimizer [59] with the default learning rate of 1e-5. We use a linear scheduler with 2,000 warmup steps. In the DST task, we set the batch size to 6 and gradient accumulation to 2. In the NLU task, we set the batch size to 16 and gradient accumulation to 1.

The implementation of the extended framework uses decoder-only LLMs as backbones, which are open-resourced and capable of supporting the candidate languages. For example, LLAMA-2-7B-CHAT-HF,[9] BLOOM-7B1,[10] and OPENCHAT3.5[11] available through the Huggingface library. The total number of parameters is about 7 billion. We perform PEFT with LoRA [34] and integrate MOLR into the LLaMA Board [31]. Specifically, the maximum sequence length is 1024 and the learning rate is 1e-05. The MOLR models are trained for 10 epochs with a per-device batch size of 8, and accumulated gradients every 4 steps. A cosine learning rate scheduler is employed, with a maximum gradient norm of 1.0. We log results every 5 steps and save model checkpoints every 100 steps. Warm-up steps are set to 2000. LoRA is used with a rank of 8 and a dropout rate of 0.1 for regularization. The monolingual models are trained with the same settings excluding the number of epoch as 60. All experiments are run on NVIDIA GeForce RTX 4090 24GB and NVIDIA A100 SXM4 40GB GPUs.

5 RESULTS

We show experimental results to answer the research questions in Section 4.1.

5.1 Main results (RQ4.1)

We compare the performance of MOLR models with the existing monolingual models and multilingual models on both the DST (see Table 3) and the NLU (see Table 4) task.

5.1.1 MOLR improves both monolingual and multilingual DST. From the results for the DST task in Table 3, we have the following observations.

First, the mT5 models with MOLR outperform all monolingual and multilingual baselines for German and Italian. They also achieve higher scores than most of the scores reported for English. Specifically, mT5+bMOLR significantly outperforms mT5 by 1.89%/2.56%/0.67% of joint goal accuracy and 0.3%/0.73%/1.09% of request accuracy for English/German/Italian. The improvements demonstrate the effectiveness of MOLR. We believe the main reason is that MOLR is able to explore pairwise relationships between languages and to fully make use of multilingual data for global optimization. Although XQA-DST and DistilledBERT† achieve slightly higher results than mT5+bMOLR for English in terms of joint goal accuracy (+0.54%) and request accuracy (+0.38%), the predictive space is much smaller than MOLR models. This is mainly because QA-DST and DistilledBERT† are classification models; in contrast, the MOLR models are generation models.

Second, mT5 is the state-of-the-art base model compared with all types of base models. More precisely, mT5 (89.53%) achieves the highest joint goal accuracy for English, followed by XLM-R-DST (88.50%), GLAD (88.10%), BERT (87.70%), NBT (84.20%). mT5 dramatically improves the existing reported results on German and Italian. Specifically, it increases 10.96% and 9.48% over the best NBT results on German and Italian, respectively.

Third, pairwise alignment brings consistent improvement. MOLR improves over mT5 in all settings. However, NBT+mDA decreases the joint goal accuracy by 1.40% for English compared with monolingual NBT. The benefit of multilingual data for training appears to be limited without modeling the language relationships. When adding more languages, mT5+mMOLR achieves slight changes: Compared with mT5+bMOLR, mT5+mMOLR improves the joint goal accuracy by 0.42% in English and the request accuracy by 0.37% in German, but slightly drops in the remaining settings.

[9] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[10] https://huggingface.co/bigscience/bloom-7b1

[11] https://huggingface.co/openchat/openchat_3.5

Table 3. Comparison of dialogue state tracking (DST) models for supervised learning using English, German, and Italian as target languages. In the cells with results, the numbers before and after "/" denote joint the goal accuracy and request accuracy, respectively. The boldface indicates leading results. As multilingual settings are under-explored in the baselines models, we reproduce several competitive baselines for comparison. These results are marked using "*".

		Joint Goal/Request Accuracy (%)		
DST Models	**Settings**	**English**	**German**	**Italian**
NBT				
NBT [69]	NBT w/CNN encoder	84.20/91.60	–	–
NBT-DNN [69]	NBT w/DNN encoder	84.40/91.20	–	–
NBT+SSU [70]	NBT+statistical state update	84.80/–	68.10/–	76.10/–
StateNet [94]	NBT+LSTM-based state update	88.90/–	–	–
NBT+Morph [110]	NBT+morphology fine-tuning	–	66.30/–	78.10/–
NBT+mDA [71]	NBT+multilingual data augmentation	82.80/–	57.70/–	77.10/–
GLAD				
GLAD [127]	Global-locally self-attentive DST	88.10/97.10	–	–
GCE [76]	GLAD+globally-conditioned encoder	88.51/97.38	–	–
GLAD+DA [123]	GLAD+paraphrase data augmentation	88.00/–	–	–
GLAD+RDA [123]	GLAD+DA+reinforcement learning	90.70/–	–	–
BERT				
BERT [8]	BERT context encoder	87.70/–	–	–
BERT+RNN [47]	BERT context encoder+RNN state decoder	89.20/–	–	–
BERT† [45]	BERT context & candidate encoder	90.50/97.60	–	–
DistilledBERT† [45]	Distilled variant of BERT†	90.40/**97.70**	53.28*/93.01*	67.92*/95.26*
SUMBT [47]	BERT+RNN+slot-utterance attention	91.00/–	–	–
XLM				
XLM-R-DST [128]	XLM-R context encoder with 270M parameters	88.50/–	–	–
XQA-DST [128]	XLM-R+value span extraction	**92.38**/–	–	–
T5 (Ours)				
mT5	Multilingual T5 with 300M parameters	89.53/97.02	79.06/95.92	87.58/95.44
mT5+bMOLR	mT5+bilingual mixture-of-languages routing	91.42/**97.32**	**81.62**/96.65	**88.25**/96.53
mT5+mMOLR	mT5+multilingual mixture-of-languages routing	**91.84**/97.02	81.56/**97.02**	87.77/96.41

Fourth, global optimization of multilingual DST is still underexplored. NBT+Morph finetunes German and Italian model with multilingual word embeddings. NBT+mDA uses multilingual data during training for global optimization. However, recent models ignore the performance on German and Italian. This might be because most research is English-centered: either English models or cross-lingual adaptation from English to other languages.

5.1.2 MOLR improves monolingual and multilingual NLU. From the results on the NLU task in Table 4, we have the following observations.

First, MOLR models outperform or are on par with all monolingual and multilingual baselines for English, Spanish, and Thai. Particularly, mT5+bMOLR and mT5+bMOLR improve over mT5 by 5.32%/1.14% and 5.13%/1.75% of slot F1 for Thai and Spanish. The improvements prove the effectiveness of MOLR. The gain of MOLR is limited in other settings, including evaluation results on English and intent accuracy. The general improvement is smaller than 0.5%. One reason is that the volume of data is already sufficient for good intent identification. For example, biLSTM-CRF achieves 99.11% of accuracy on intent identification and 94.81% of slot F1 for English. Thus, the

Table 4. Comparison of natural language understanding (NLU) models for supervised learning using English, Spanish, and Thai as target languages. In the cells with results, the numbers before and after "/" denote intent accuracy and slot F1 score, respectively. Boldface indicates leading results.

		Intent Accuracy/Slot F1 (%)		
NLU Models	**Settings**	**English**	**Spanish**	**Thai**
RNNs				
biLSTM [57]	biLSTM for only target language	–/94.87	–	–
biLSTM-CRF [98]	Monolingual biLSTM with CRF layer	99.11/94.81	97.26/80.95	95.13/87.26
CoVe [98]	biLSTM-CRF based NMT to English	–	97.81/82.55	96.87/90.60
mCoVe [98]	Multilingual CoVe [66]	–	97.82/82.49	96.98/91.22
mCoVe+Auto [98]	mCoVe with autoencoder objective	–	97.90/82.13	96.87/91.51
Transformers				
Transformer [57]	Transformer w/frozen word embeddings	–/94.93	–	–
BERT				
mBERT [57]	mBERT fine-tuning	–/95.97	–	–
mBERT+DA [93]	mBERT+ monolingual data augmentation	–	98.20/84.27	91.42/59.68
XLM				
XLM-R [27]	XLM-R encoder with 270M parameters	–	98.70/89.10	96.80/93.10
XLM-R+TA [27]	XLM-R+translation alignment loss	99.30/**96.60**	98.80/89.80	**97.80**/94.40
T5 (Ours)				
mT5	Multilingual T5 with 300M parameters	99.35/96.40	98.68/88.45	97.52/89.48
mT5+bMOLR	mT5+bilingual mixture-of-languages routing	99.29/96.49	**99.08**/89.59	97.28/**94.81**
mT5+mMOLR	mT5+multilingual mixture-of-languages routing	**99.40**/96.50	98.88/**90.21**	97.70/94.61

performance of slot filling leaves more room for improvement than intent identification. Another reason is that the extra information from low-resource languages (e.g., Spanish and Thai when used in combination with English) is quite limited. For example, in the NLU dataset, only 11.7% and 20.0% of the utterances are parallel with English, respectively, which is al that is available to help improve the English model. In contrast, the rest of the non-parallel English utterances can bring new information to the low-resource languages.

Second, mT5 is the state-of-the-art base model compared with all types of base models similar to DST (see Table 3). Specifically, mT5 (96.40%) obtains the highest slot F1 for English, followed by mBERT (95.97%), transformers (94.93%) and biLSTM (94.87%). XLM-R is as competitive as mT5, but we choose mT as our backbone considering both the DST performance and the generation benefit (i.e., out-of-ontology prediction).

Third, pairwise alignment brings consistent improvements for slot filling. Compared with mT5, mT5+bMOLR improves 5.32%, 1.14%, 0.09% of slot F1 for Thai, Spanish and English, respectively. Adding more languages to mT5+bMOLR, mT5+mMOLR brings a small increase for most settings, except for a small decrease in slot F1 for Thai (-0.2%). Similarly, it depends on how much meaningful information a new language can bring to learn better relationships.

Fourth, global optimization of multilingual NLU is still underexplored. Even for language-specific optimization, biLSTM-CRF and XLM-R+TA are the only approaches for which results are reported on all languages, to the best of our knowledge. This might be because most prior research focuses on cross-lingual adaptation from English to other low-resource languages in the NLU dataset.

5.2 Analysis of language characteristics (RQ4.2)

In Section 5.1, we observe that the overall performance varies a lot for different languages. In this section, we first analyze the language characteristics (i.e., unity and diversity) in depth by visualizing the word embeddings, as well as genetic, word, and sentence similarities of different languages. Then we analyze the gains on different languages in different settings.

5.2.1 Qualitative analysis of the unity and diversity of languages. See Figure 4 for visualizations of the word embeddings of mT5+mMOLR in the DST and NLU datasets, before and after fine-tuning, respectively. We aim to understand how MOLR influences the unity and diversity of languages qualitatively.

First, different languages have both similar and dissimilar words in the semantic embedding space. Specifically, some data points from different languages are very close to each other while other data points are far away and located in an isolated cluster. For example, parts of English and German points are mixed up while other sets of German data points are concentrated in an isolated area.

Second, similarities between languages are very different for different language pairs. For example, the boundaries between English and German, and between English and Italian are not obvious; in contrast, the boundaries between Thai and English, and Thai and Spanish are quite clear. This indicates that English, German, and Italian are quite similar to each other, while Thai is an independent and distinct language that is not similar to English and Spanish. Besides, Thai is closer to English rather than Spanish; the English cluster seems to separate Thai and Spanish.

Third, the relative relationships are not changed before and after fine-tuning. In DST, English, German, and Italian points are mixed together with a small isolated cluster of German points. In NLU, English and Spanish have both shared and non-shared areas, while the majority of Thai points are in an isolated cluster. This shows that English, German, and Italian have more unity while Thai has more diversity compared with English and Spanish.

5.2.2 Quantitative analysis of the unity and diversity of languages. As shown in Table 5, we evaluate the unity and diversity of languages by three similarity metrics, i.e, genetic similarity, word similarity, and sentence similarity based on word embeddings of mT5 model in the datasets. We aim to quantify the unity and diversity of language and use the similarity order to analyze our MOLR models in the next section.

First, (EN, DE) are the most similar language pair in terms of genetic similarity, followed by (EN, IT), (DE, IT), (EN, ES). Second, (ES, TH) are more similar than (EN, TH) in terms of word and sentence similarity. Last but not least, considering the similarity in one aspect is not always meaningful. For example, the comparison of genetic similarity is invalid for "(EN, TH)" and "(ES, TH)", and word similarity of "(EN, DE)" and "(EN, IT)" has very little difference. Another example is that the order of sentence similarity is inconsistent with that of word similarity. Unlike in NLU, the difference in word similarity in DST is small, which might increase the difficulty of distinguishing between sentence embeddings. Hence, it is important to consider all similarity metrics. In this work, we sort the similarity degree of all language pairs in descending order as:

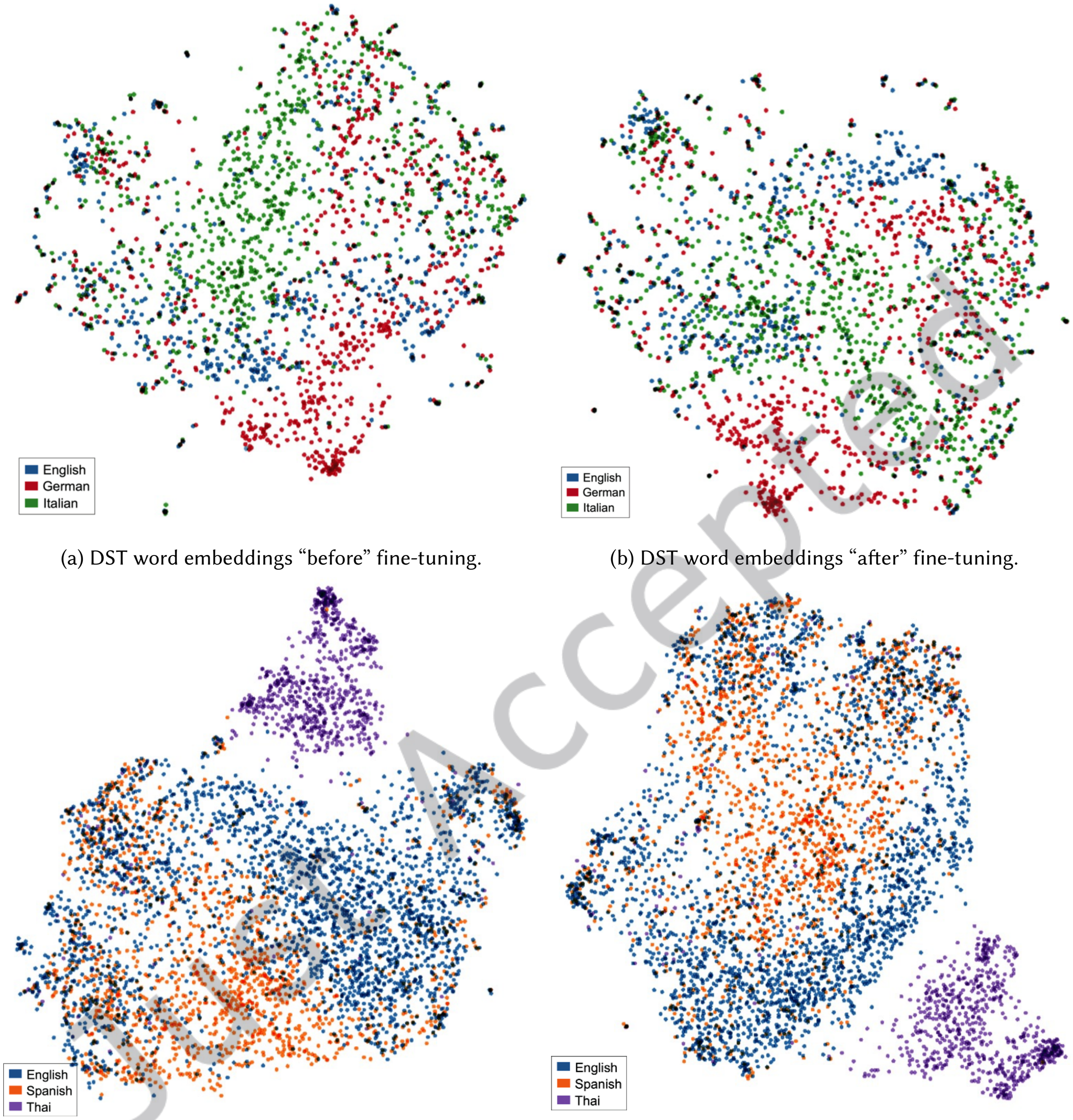
$$\boxed{\phi(EN, DE) > \phi(EN, IT) > \phi(DE, IT) > \phi(EN, ES) > \phi(EN, TH) > \phi(ES, TH)}$$

The point of this order is to fairly compare the similarity between languages; this can be used as language-specific knowledge to analyze how different languages influence MOLR.

5.2.3 Gains of MOLR are language-specific. We compare mT5 and its variants with different language routes on the DST and NLU tasks, as shown in Table 6.

First, gains of MOLR vary when choosing different pivot languages (Section 3.3) at different stages of language routes (Section 4.3). mT5+bMOLR achieves better performance when the similarity between the source language

[12] https://projector.tensorflow.org/



(a) DST word embeddings “before” fine-tuning.

(b) DST word embeddings “after” fine-tuning.

(c) NLU word embeddings “before” fine-tuning.

(d) NLU word embeddings “after” fine-tuning.

Fig. 4. Visualization of words in the DST and NLU datasets “before” and “after” fine-tuning. We conduct dimension reduction using the UMAP algorithm [67] and plot all scatter in 2D coordinates using the Tensorflow embedding projector.[12]

and pivot language is larger in most settings. Specifically, it obtains 1.35% higher joint goal accuracy for Italian DST using English rather than German as the pivot language. Besides, it gets 1.45% improvements in slot F1

Table 5. Similarity between languages. We report genetic similarity, as well as word and sentence similarity based on the DST and NLU datasets.

	DST: Similarity				NLU: Similarity		
Language pair	**Genetic**	**Word**	**Sentence**	**Language pair**	**Genetic**	**Word**	**Sentence**
(EN, DE)	0.1667	0.6725	0.8813	(EN, ES)	0.0833	0.7448	0.8777
(EN, IT)	0.1250	0.6711	0.9036	(EN, TH)	0.0000	0.4787	0.5706
(DE, IT)	0.0909	0.6486	0.9066	(ES, TH)	0.0000	0.4056	0.5512

Table 6. The performance of the proposed mT5-based models with different language routes on the DST and NLU tasks. The bold numbers are the best results in terms of different evaluation metrics for target languages.

	DST: Joint Goal / Request Accuracy (%)					
Model	**English (EN)**		**German (DE)**		**Italian (IT)**	
mT5	89.53/97.02		79.06/95.92		87.58/95.44	
	EN,DE→EN	EN,IT→EN	DE,EN→DE	DE,IT→DE	IT,EN→IT	IT,DE→IT
mT5+bMOLR	91.42/97.32	91.11/**97.57**	**81.62**/96.65	**81.62**/96.23	**88.25**/**96.53**	86.98/96.41
	1.DE,IT→DE 2.EN,DE→EN	1.IT,DE→IT 2.EN,IT→EN	1.EN,IT→EN 2.DE,EN→DE	1.IT,EN→IT 2.DE,IT→DE	1.EN,DE→EN 2.IT,EN→IT	1.DE,EN→DE 2.IT,DE→IT
mT5+mMOLR	**91.84**/97.02	91.42/97.14	81.56/**97.02**	81.38/96.23	87.77/96.41	86.00/96.35
	NLU: Intent Accuracy/ Slot F1 (%)					
	English (EN)		**Spanish (ES)**		**Thai (TH)**	
mT5	99.35/96.40		98.68/88.45		97.52/89.48	
	EN,ES→EN	EN,TH→EN	ES,EN→ES	ES,TH→ES	TH,EN→TH	TH,ES→TH
mT5+bMOLR	99.29/96.49	99.29/96.34	**99.08**/89.59	98.78/88.94	97.28/94.81	97.64/93.39
	1.ES,TH→ES 2.EN,ES→EN	1.TH,ES→TH 2.EN,TH→EN	1.EN,TH→EN 2.ES,EN→ES	1.TH,EN→TH 2.ES,TH→ES	1.EN,ES→EN 2.TH,EN→TH	1.ES,EN→ES 2.TH,ES→TH
mT5+mMOLR	**99.40/96.50**	99.30/96.39	98.91/89.53	98.88/90.21	97.70/94.61	97.52/94.59

for Thai NLU using English rather than Spanish as the pivot language. For mT5+mMOLR, the first stage is pre-training, and the second stage is the main procedure. It achieves better performance when the similarity between the source language and the pivot language in the second stage is larger in most settings. For example, it improves 1.77% of joint goal accuracy for Italian DST using English rather than German as the pivot language in the second stage. Language transfer is easier if the source and pivot languages are more similar, and it can avoid introducing too many language gaps in the early stage of a language route.

Second, compared with mT5+bMOLR, the performance of mT5+mMOLR varies with different additional languages in the second stage. For example, mT5+mMOLR increases 0.42% in joint goal accuracy by adding Italian into the additional route of "DE,IT→DE". However, mT5+mMOLR decreases 0.98% in joint goal accuracy by adding English into the additional route of "DE,EN→DE". The second stage is essential, and it is helpful if the source language and pivot language are similar in the second stage.

Third, the performance of mT5+mMOLR is dependent on the volume of additional languages and the difficulty of tasks. For most settings, the changes are small compared with mT5+bMOLR. Particularly, mT5+mMOLR

Table 7. Comparison of the effect of different combination policies on mT5+bMOLR model. The bold numbers are the best results in terms of different evaluation metrics for target languages.

	DST: Joint Goal Accuracy / Request Accuracy (%)					
Model	**English (EN)**		**German (DE)**		**Italian (IT)**	
mT5	89.53/97.02		79.06/95.92		87.58/95.44	
	EN,DE→EN	EN,IT→EN	DE,EN→DE	DE,IT→DE	IT,EN→IT	IT,DE→IT
w/ bDA	89.59/96.71	91.05/96.90	79.98/96.35	81.25/96.53	85.82/**96.71**	87.89/96.17
w/ route-addressing	89.11/97.14	89.17/95.92	79.12/95.44	77.97/95.74	84.97/95.86	80.85/95.01
w/ parameter-sharing	**91.42**/97.32	91.11/**97.57**	**81.62**/**96.65**	**81.62**/96.23	**88.25**/96.53	86.98/96.41
	NLU: Intent Accuracy / Slot F1 (%)					
	English (EN)		**Spanish (ES)**		**Thai (TH)**	
mT5	99.35/96.40		98.68/88.45		97.52/89.48	
	EN,ES→EN	EN,TH→EN	ES,EN→ES	ES,TH→ES	TH,EN→TH	TH,ES→TH
w/ bDA	**99.34**/**96.53**	99.32/96.47	98.98/89.92	98.72/88.57	**97.87**/94.45	97.52/92.06
w/ route-addressing	99.25/95.93	99.22/95.86	98.72/**89.89**	98.45/87.19	97.52/93.67	97.16/91.33
w/ parameter-sharing	99.29/96.49	99.29/96.34	**99.08**/89.59	98.78/88.94	97.28/**94.81**	97.64/**93.39**

increases 1.27% of slot F1 by adding English into the additional route of "TH,EN→TH" for Spanish NLU. Similarly, mT5+mMOLR increases 1.20% of slot F1 by adding English into the additional route of "ES,EN→ES" for Thai NLU. One reason is that English is a high-resource language compared with Spanish and Thai in NLU, which is able to provide sufficient extra information for improvement. Another reason is that the slot filling task is more difficult than intent identification, and the potential for improvement of the former is larger than the latter.

5.3 Analysis of key components of mT5+bMOLR (RQ4.3)

5.3.1 Combination policies are essential. We compare mT5 with its variants, i.e., mT5+bDA which is mT5 with bilingual data training, route-addressing, and parameter-sharing combination policies, as shown in Table 7.

First, mT5 with parameter-sharing outperforms mT5 in all settings. Specifically, it improves 2.56%/0.73%, 1.83%/0.61%, and 0.19%/0.96% for German, English, and Italian DST in terms of joint goal accuracy and request accuracy, respectively. Meanwhile, it improves 0.18%/5.13%, 0.23%/1.76%, and 0.05%/0.10% for Thai, Spanish, and English NLU, respectively. This proves the overall effectiveness of MOLR models.

Second, mT5 with parameter-sharing outperforms or is on par with bDA (i.e., bilingual data augmentation) in all settings. In DST, "EN, DE→EN", "EN, DE→DE", "IT,EN→DE" changes +1.83%/+0.61%, +1.64%/+0.3%, +2.43%/-0.17% in terms of joint goal accuracy and request accuracy, given a pivot language similar to source language. However, mT5+bDA cannot always benefit from multilingual data, e.g., "IT,EN→IT" decreases 1.76% of joint goal accuracy compared with mT5. In NLU, "TH,ES→TH" and "ES, TH→ES" mutually increase as much as 2.53% and 1.64% in terms of slot F1, while the changes are quite small (<0.40%) for the other settings. This reveals that the gains are also from MOLR and global optimization, along with multilingual data.

Third, policy parameter-sharing outperforms route-addressing in general. In DST, parameter-sharing beats route-addressing by 1.94%–6.13% and 0.18%–1.64% in terms of joint goal accuracy and request accuracy. In NLU, parameter-sharing outperforms route-addressing by 0.04%–2.40% and 0.48%–2.06% in terms of joint goal accuracy and request accuracy, excluding the intent accuracy for "TH,EN→TH" (-0.24%) and slot F1 for "ES,EN→ES" (-0.30%). Thus, we use parameter-sharing as the combination policy for our best-performing models.

Table 8. Model complexity by agent models with the different number of layers.

	DST: Joint Goal Accuracy / Request Accuracy (%)					
	English (EN)		**German (DE)**		**Italian (IT)**	
#Layers	EN,DE→EN	EN,IT→EN	DE,EN→DE	DE,IT→DE	IT,EN→IT	IT,DE→IT
8	**91.42**/**97.32**	**91.11**/**97.57**	**81.62**/**96.65**	81.62/96.23	**88.25**/**96.53**	86.98/**96.41**
6	90.75/97.20	90.44/97.14	80.83/95.01	**82.47**/95.98	85.51/96.10	**87.16**/**96.41**
4	88.92/97.20	88.98/97.02	78.76/96.04	79.79/**96.53**	85.09/96.35	83.99/95.98
2	86.49/97.14	86.79/97.02	76.38/95.56	78.03/94.64	81.98/95.19	81.56/95.19
	NLU: Intent Accuracy / Slot F1 (%)					
	English (EN)		**Spanish (ES)**		**Thai (TH)**	
	EN,ES→EN	EN,TH→EN	ES,EN→ES	ES,TH→ES	TH,EN→TH	TH,ES→TH
8	99.29/**96.49**	99.29/96.34	**99.08**/89.59	**98.78**/**88.94**	97.28/**94.81**	97.64/**93.39**
6	**99.38**/96.37	99.27/96.33	98.88/**89.87**	98.68/88.33	**97.58**/93.27	**97.66**/91.58
4	99.28/96.07	99.27/96.01	98.95/87.28	98.55/83.75	97.40/91.61	97.22/84.87
2	99.30/93.53	**99.33**/93.68	98.62/83.59	98.39/66.25	97.64/84.82	96.93/69.16

5.3.2 Impact of the number of layers varies with tasks and languages. We study the influence of the different number of layers for each expert model in Table 8.

First, the best settings of layers for expert agents (Section 3.2) vary for different tasks. In DST, joint goal accuracy notably decreases 3.59%–6.27%, and request accuracy only decreases by 0.17%–1.59% with reducing the number of layers from 8 to 2. In NLU, slot F1 dramatically reduces by 2.96%–24.23%, in contrast, intent accuracy reduces or even increases slightly, e.g., "TH,EN→TH" improves 0.36%. The difficulty of different tasks varies, and the number of layers has less influence on simpler tasks.

Second, the influence of the number of layers is language-specific. In DST, we reduce the number of layers from 8 to 2. The mixture of German and Italian (i.e., "DE,IT→DE" and "IT,DE→IT") does not always drop like the rest of the settings. Since the amount of multilingual data is comparable, it is likely caused by the language specification, i.e., (DE, IT) are less similar than (EN, DE) and (EN, IT), and the mixture of German and Italian can preserve more diversity.

Third, the number of layers is sensitive to high-resource pivot languages. Reducing the number of layers from 8 to 4, the changes in "EN,ES→EN " and "EN,TH→EN" are less than 0.1% and 0.5% in terms of intent accuracy and slot F1. The pivot languages (i.e., Spanish and Thai) have much fewer data samples compared with the high-resource language, i.e., English.

5.4 Feasibility of MOLR framework with decoder-only LLM backbones (RQ4.4)

5.4.1 Outcome. We evaluate the performance of extended MOLR models using various LLMs as backbones on the DST task, which poses greater challenges and opportunities for improvement than the NLU task. We compare the results of several models with both encoder-decoder and decoder-only LLMs as backbones in Table 9. First, LLaMa-2-Chat+mMOLR significantly outperforms all models in terms of all metrics, using English, German, and Italian as target languages. Specifically, joint goal accuracy increases by 1.48% in English, 2.52% in German, and 2.33% in Italian, compared with mT5+mMOLR. The primary factor is the substantial increase in the number of learnable parameters in the latest decoder-only LLMs, surpassing those of encoder-decoder LLMs by a significant margin. This verifies the feasibility of extending MOLR framework with various decoder-only LLM backbones. Second, the MOLR framework exerts a more significant impact on decoder-only LLMs. The MOLR

Table 9. Comparison with models based on decoder-only LLMs for supervised learning of the dialogue state tracking (DST) task, using English, German, and Italian as target languages. In the cells with results, the numbers before and after "/" denote joint the goal accuracy and request accuracy, respectively. Boldface indicates leading results.

	Joint Goal Accuracy / Request Accuracy (%)		
DST Models	**English**	**German**	**Italian**
Encoder-decoder LLMs			
mT5	89.53/97.02	79.06/95.92	87.58/95.44
mT5+mMOLR	91.84/97.02	81.56/97.02	87.77/96.41
Decoder-only LLMs			
OpenChat3.5	14.52/95.75	15.74/91.49	15.55/93.32
Bloom	87.97/96.66	75.58/95.81	80.50/95.20
LLaMa-2-Chat	90.40/97.21	79.83/96.05	86.63/96.42
LLaMa-2-Chat+mMOLR	**93.32/97.93**	**84.08/97.27**	**90.10/97.08**

Table 10. Comparison with models based on decoder-only LLMs for supervised learning of the natural language generation (NLG) task, using English, German, and Italian as target languages. In the cells with results, the numbers before and after "/" denote BLEU-4 and ROUGE-L, respectively. Boldface indicates leading results.

	BLEU-4 / ROUGE-L (%)		
NLG Models	**English**	**German**	**Italian**
Encoder-decoder LLMs			
mT5	**33.97**/32.66	32.96/33.03	32.80/32.17
mT5+mMOLR	33.91/**32.94**	**35.53/33.81**	**34.14/33.09**
Decoder-only LLMs			
OpenChat3.5	5.01/10.54	4.11/5.35	4.54/8.23
Bloom	29.14/27.24	28.65/27.14	28.11/26.79
LLaMa-2-Chat	31.50/29.21	30.95/29.22	30.68/28.62
LLaMa-2-Chat+mMOLR	33.24/31.39	32.59/30.46	32.61/30.54

framework improves the mT5 performance by 2.31% in English, 2.5% in German, 0.19% in Italian, in terms of joint goal accuracy, while it boosts LLaMa-2-Chat performance by 2.92% in English, 4.25 in German, 3.47% in Italian. Third, decoder-only LLMs are still facing difficulties in instruction following. For example, OpenChat3.5 drops dramatically in all languages in terms of joint goal accuracy. We carefully checked the generated output and found the reason to be that OpenChat3.5 generates lots of redundant and repeat inform-slot-value triplets.

We evaluate the performance of extended MOLR models using various LLMs as backbones on the NLG task, as shown in Table 10. First, we observed that the utilization of MOLR led to enhancements in the performance of both the encoder-decoder LLM mT5 and the decoder-only LLM LLaMa-2-Chat. For instance, with MOLR, mT5 achieved a notable increase of 2.57% in BLEU-4 for German and 1.34% for Italian, while experiencing a slight decrease of 0.06% in English. Similarly, LLaMa-2-Chat exhibited improvements, with BLEU-4 increasing by 1.74% in English, 1.64% in German, and 1.93% in Italian. Second, the combination of mT5 with mMOLR consistently attains the highest scores, closely trailed by LLaMa-2-Chat with mMOLR. This trend can be attributed to mT5's encoder-decoder architecture, while LLaMa-2-Chat operates within a decoder-only framework and heavily relies on the quality of the provided prompt as input. Notably, the prompt for the DST task is comparatively simpler and less well-designed compared to that for the NLG task.

5.4.2 Discussion. In the past few of months, decoder-only LLMs have more and more become prevailing. We are cognizant of this and are actively working towards extending MOLR by integrating open-sourced decoder-only LLMs. This initiative facilitates seamless adoption of the most recent LLMs. However, tracking belief states with LLMs is still challenging even for some commercial models such as ChatGPT. Several recent studies show that the generated outputs are sensitive to a well-designed prompt with an exhaustive list of schema [51, 65]. Moreover, it unavoidably has extrinsic hallucination beyond the given schema and knowledge [1]. Finally, it is worth noting that tailoring LLMs for downstream tasks necessitates ample task-specific data on a large scale. This underscores the potential of MOLR, offering a pathway for multilingual data augmentation.

6 CONCLUSION AND FUTURE WORK

In this work, we have studied multilingual task-oriented dialogue systems in a collaborative task-oriented dialogue system framework, where expert agents work on monolingual and cross-lingual dialogues, and the chair agent accounts for a mixture-of-experts approach for globally optimizing multilingual dialogues. We have proposed a mixture-of-languages routing framework, which aims to fully make use of multilingual data, capture language relationships, and globally optimize multilingual performance simultaneously. We have conducted experiments on two benchmark multilingual task-oriented dialogue system datasets to verify the effectiveness of the proposed mixture-of-languages routing based on a pre-trained mT5 model.

Our main finding is that mixture-of-languages routing can be greatly influenced by data availability, language characteristics, as well as collaboration policies. To be precise, training mixture-of-languages routing with sufficient multilingual data can significantly improve performance over training with little data in a low-resource language. Moreover, mixture-of-languages routing with increasing amounts of data in different languages can perform very differently, so the gains of mixture-of-languages routing are language-specific. Different combination policies enable global optimization, and their performance varies a lot; this demonstrates the versatility and effectiveness of the collaborative paradigm. Together, these findings and insights provide an affirmative answer to the leading research question for this work: multiple languages can indeed be used in a collaborative way to improve the performance of task-oriented dialogue systems in every single language.

As to broader implications of multilingual task-oriented dialogue systems, researchers in this domain should consider as many languages as possible. They should also enable their models to select valuable data for enhancement. Language characteristics (e.g., unity and diversity) should never been underestimated.

One limitation of this work is that mixture-of-languages routing only works when multilingual data can bring both commonalities and peculiarities of languages. In the extreme case where two languages do not have any commonalities, mixture-of-languages routing can hardly learn to transfer across languages for cross-lingual adaptation. And vice versa, if two languages do not have any peculiarities, mixture-of-languages routing can hardly gain from language transfer for a multilingual model.

As to future work, we believe that multilingual task-oriented dialogue systems can be advanced in many directions. First, we plan to use different pre-trained language models (e.g., GPT2, mBART, etc.) and compare them with the mT5 model. Second, we plan to explore different collaboration policies and see how they influence the overall performance. Third, we plan to experiment on new datasets with full dialogue tasks and more languages and step forward to practical applications of multilingual task-oriented dialogue systems.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for their valuable feedback.

This research was partially funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO), and the FINDHR (Fairness and Intersectional Non-Discrimination in Human Recommendation) and the VOXReality (Voice-driven Interaction in XR Spaces) projects that received funding from the European Union's Horizon Europe research and innovation program under grant agreements No 101070212 and No 101070521, and the SURF cooperative grant funded by NWO under grant agreement No EINF-7705.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[2] Lisa Beinborn and Rochelle Choenni. 2020. Semantic drift in multilingual representations. *Computational Linguistics* 46, 3 (2020), 571–603.

[3] Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate production using character-based machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. 883–891.

[4] Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics* 45, 2 (2019), 381–389.

[5] Susanne Burger, Karl Weilhammer, Florian Schiel, and Hans G Tillmann. 2000. Verbmobil data collection and annotation. In *Verbmobil: Foundations of Speech-to-speech Translation*. Springer, 537–549.

[6] Hugo C.C. Carneiro, Felipe M.G. França, and Priscila M.V. Lima. 2015. Multilingual part-of-speech tagging with weightless neural networks. *Neural Networks* 66 (2015), 11–21.

[7] Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *Comput. Surveys* 54, 2 (2021), 1–37.

[8] Guan-Lin Chao and Ian Lane. 2019. BERT-DST: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. *Proceedings of Interspeech* (2019).

[9] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2 (2017), 25–35.

[10] Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. XL-NBT: A cross-lingual neural belief tracking framework. In *EMNLP*. 414–424.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*. 8440–8451.

[12] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems* 32 (2019).

[13] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* (2017).

[14] David Crystal. 2008. Two thousand million? *English Today* 24, 1 (2008), 3–6.

[15] Richard Csaky and Gabor Recski. 2020. The gutenberg dialogue dataset. *arXiv preprint arXiv:2004.12752* (2020).

[16] Michael Cysouw. 2013. Predicting language-learning difficulty. In *Approaches to Measuring Linguistic Differences*. De Gruyter.

[17] Raj Dabre, Aizhan Imankulova, Masahiro Kaneko, and Abhisek Chakrabarty. 2021. Simultaneous multi-pivot neural machine translation. *arXiv preprint arXiv:2104.07410* (2021).

[18] Ewa Dabrowska. 2015. What exactly is universal grammar, and has anyone seen it? *Frontiers in Psychology* 6 (2015), 852.

[19] Michael Daniel. 2011. Linguistic typology and the study of language. In *The Oxford Handbook of Linguistic Typology*. Oxford University Press.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*. 4171–4186.

[21] Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing multiwoz to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1639–1657.

[22] Robert M.W. Dixon. 2010. *I am a linguist: with a foreword by Peter Matthews*. Brill.

[23] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*. PMLR, 1126–1135.

[24] W. Tecumseh Fitch. 2011. Unity and diversity in human language. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366, 1563 (2011), 376–388.

[25] Alexandre François. 2015. Trees, waves and linkages: Models of language diversification. In *The Routledge Handbook of Historical Linguistics*. Routledge, 161–189.

[26] Pascale Fung and Tanja Schultz. 2008. Multilingual spoken language processing. *IEEE Signal Processing Magazine* 25, 3 (2008), 89–97.

[27] Milan Gritta and Ignacio Iacobacci. 2021. XeroAlign: Zero-shot cross-lingual transformer alignment. In *ACL Findings*. 371–381.

[28] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints* (2023).

[29] Martin Haspelmath. 2004. How hopeless is genealogical linguistics, and how advanced is areal linguistics? *Studies in Language* 28, 1 (2004), 209–223.

[30] Claudia Hauff, Julia Kiseleva, Mark Sanderson, Hamed Zamani, and Yongfeng Zhang. 2021. Conversational search and recommendation: Introduction to the special issue. *ACM Transactions on Information Systems* 39, 4 (2021), 1–6.

[31] Hiyouga. 2023. LLaMA Factory. https://github.com/hiyouga/LLaMA-Factory.

[32] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems* 33 (2020), 20179–20191.

[33] Eduard Hovy, Nancy Ide, Robert Frederking, Joseph Mariani, and Antonio Zampolli. 2001. *Multilingual Information Management: Current Levels and Future Abilities.* Istituti Editoriali e Poligrafici Internazionali, Pisa.

[34] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank adaptation of large language models. In *ICLR*. OpenReview.net. https://openreview.net/forum?id=nZeVKeeFYf9

[35] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems* 38, 3 (2020), 1–32.

[36] Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. Multi2WOZ: A robust multilingual dataset and conversational pretraining for task-oriented dialog. *arXiv preprint arXiv:2205.10400* (2022).

[37] Pratik Jayarao and Aman Srivastava. 2018. Intent detection for code-mix utterances in task oriented dialogue systems. In *ICEECCOT*. IEEE, 583–587.

[38] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).

[39] Armand Joulin, Piotr Bojanowski, Tomáš Mikolov, Hervé Jégou, and Édouard Grave. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *EMNLP*. 2979–2984.

[40] Prabhu Kaliamoorthi, Aditya Siddhant, Edward Li, and Melvin Johnson. 2021. Distilling Large Language Models into Tiny and Effective Students using pQRNN. *arXiv preprint arXiv:2101.08890* (2021).

[41] Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, Matthew Henderson, and Koichiro Yoshino. 2016. The fifth dialog state tracking challenge. In *SLT Workshop*. IEEE, 511–517.

[42] Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*. 12–18.

[43] Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *MRL Workshop*. 211–223.

[44] Adarsh Kumar, Peter Ku, Anuj Goyal, Angeliki Metallinou, and Dilek Hakkani-Tur. 2020. Ma-dst: Multi-attention-based scalable dialog state tracking. In *AAAI*, Vol. 34. 8107–8114.

[45] Tuan Manh Lai, Quan Hung Tran, Trung Bui, and Daisuke Kihara. 2020. A simple but effective bert model for dialog state tracking on resource-limited systems. In *ICASSP*. IEEE, 8034–8038.

[46] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv.2211.05100* (2022).

[47] Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. SUMBT: Slot-utterance matching for universal and scalable belief tracking. In *ACL*. 5478–5483.

[48] Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *EACL*. 2950–2962.

[49] Juntao Li, Chang Liu, Chongyang Tao, Zhangming Chan, Dongyan Zhao, Min Zhang, and Rui Yan. 2021. Dialogue history matters! Personalized response selection in multi-turn retrieval-based chatbots. *ACM Transactions on Information Systems* 39, 4 (2021), 1–25.

[50] Tomasz Limisiewicz and David Mareček. 2020. Syntax Representation in Word Embeddings and Neural Networks–A Survey. *arXiv preprint arXiv:2010.01063* (2020).

[51] Zhaojiang Lin, Bing Liu, Andrea Madotto, Seungwhan Moon, Zhenpeng Zhou, Paul A Crook, Zhiguang Wang, Zhou Yu, Eunjoon Cho, Rajen Subba, et al. 2021. Zero-shot dialogue state tracking via cross-task transfer. In *EMNLP*. 7890–7900.

[52] Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021. XPersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 3rd Workshop on Natural Language Processing for*

Conversational AI. 102–112.

[53] Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. *arXiv preprint arXiv:2106.02787* (2021).

[54] Yanxiang Ling, Fei Cai, Jun Liu, Honghui Chen, and Maarten de Rijke. 2023. Generating relevant and informative questions for open-domain conversations. *ACM Transactions on Information Systems* 41, 1 (2023), Article 2.

[55] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* 8 (2020), 726–742.

[56] Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot Cross-lingual dialogue systems with transferable latent variables. In *EMNLP-IJCNLP*. 1297–1303.

[57] Zihan Liu, Genta I Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021. On the importance of word order information in cross-lingual sequence labeling. In *AAAI*, Vol. 35. 13461–13469.

[58] Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *AAAI*, Vol. 34. 8433–8440.

[59] Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *ICLR*.

[60] Samuel Louvan and Bernardo Magnini. 2020. Simple is better! Lightweight data augmentation for low resource slot filling and intent classification. In *PACLIC*. 167–177.

[61] Jianjun Ma, Jiahuan Pei, and Degen Huang. 2016. Identification of English functional noun phrases using CRFs combining the semantic information. *Journal of Chinese Information Processing* 30, 6 (2016), 59–66.

[62] Jianjun Ma, Jiahuan Pei, Degen Huang, and Dingxin Song. 2018. Syntactic parsing of clause constituents for statistical machine translation. *International Journal of Computational Science and Engineering* 17, 1 (2018), 126–132.

[63] Longxuan Ma, Mingda Li, Wei-Nan Zhang, Jiapeng Li, and Ting Liu. 2021. Unstructured Text Enhanced Open-Domain Dialogue System: A Systematic Survey. *ACM Transactions on Information Systems* 40, 1 (2021), 1–44.

[64] Brian MacWhinney. 2005. A unified model of language acquisition. In *Handbook of Bilingualism: Psycholinguistic Approaches*, Judith F. Kroll and Annette M.B. de Groot (Eds.). Vol. 4967. Oxford University Press, 50–70.

[65] Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118* (2021).

[66] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *Advances in Neural Information Processing Systems* 30 (2017).

[67] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software* 3, 29 (2018), 861.

[68] David P Medeiros. 2018. ULTRA: Universal grammar as a universal parser. *Frontiers in Psychology* 9 (2018), 155.

[69] Nikola Mrkšić, Diarmuid O Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *ACL*. 1777–1788.

[70] Nikola Mrkšić and Ivan Vulić. 2018. Fully statistical neural belief tracking. In *ACL*. 108–113.

[71] Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-Lingual constraints. *Transactions of the Association for Computational Linguistics* 5 (2017), 309–324.

[72] André Müller, Søren Wichmann, Viveka Velupillai, Cecil H. Brown, Pamela Brown, Sebastian Sauppe, Eric W. Holman, Dik Bakker, Johann-Mattis List, Dmitri Egorov, Oleg Belyaev, Robert Mailhammer, Matthias Urban, Helen Geyer, and Anthony Grant. 2010. ASJP world language tree of lexical similarity: Version 3 (July 2010). https://asjp.clld.org/static/WorldLanguageTree-003.pdf.

[73] Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence Research* 36 (2009), 341–385.

[74] Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Adiga, and Erik Cambria. 2021. Recent advances in deep learning based dialogue systems: A systematic survey. *arXiv preprint arXiv:2105.04387* (2021).

[75] Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *CICLing*. Springer, 3–16.

[76] Elnaz Nouri and Ehsan Hosseini-Asl. 2018. Toward scalable neural dialogue state tracking model. *arXiv preprint arXiv:1812.00899* (2018).

[77] Javad Nouri and Roman Yangarber. 2016. From alignment of etymological data to phylogenetic inference via population genetics. In *CogACLL Workshop*. 27–37.

[78] Nathaniel Oco, Leif Romeritch Syliongka, Rachel Edita Roxas, and Joel Ilao. 2013. Dice's coefficient on trigram profiles as metric for language similarity. In *O-COCOSDA/CASLRE*. IEEE, 1–4.

[79] Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *COLING Technical Papers*. 1297–1308.

[80] Subhadarshi Panda, Caglar Tirkaz, Tobias Falke, and Patrick Lehnen. 2021. Multilingual paraphrase generation for bootstrapping new features in task-oriented dialog systems. In *Workshop on NLP for Conversational AI*. 30–39.

[81] Hyunji Hayley Park, Katherine J Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics* 9 (2021), 261–276.

[82] Nicholas D. Pattengale, Eric J. Gottlieb, and Bernard M.E. Moret. 2007. Efficiently computing the Robinson-Foulds metric. *Journal of Computational Biology* 14, 6 (2007), 724–735.

[83] Michael Paul, Andrew Finch, and Eiichrio Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Transactions on Asian Language Information Processing (TALIP)* 12, 4 (2013), 1–17.

[84] Jiahuan Pei, Pengjie Ren, and Maarten de Rijke. 2019. A modular task-oriented dialogue system using a neural mixture-of-experts. In *SIGIR Workshop on Conversational Interaction Systems*.

[85] Jiahuan Pei, Pengjie Ren, and Maarten de Rijke. 2021. A cooperative memory network for personalized task-oriented dialogue systems with incomplete user profiles. In *The Web Conference*. 1552–1561.

[86] Jiahuan Pei, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2020. Retrospective and prospective mixture-of-generators for task-oriented dialogue response generation. In *ECAI*. 2148–2155.

[87] Carol Peters, Martin Braschler, and Paul Clough. 2012. *Multilingual Information Retrieval*. Springer-Verlag, Berlin, Heidelberg.

[88] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2021. CoSDA-ML: multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *IJCAI*. 3853–3860.

[89] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[90] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21 (2020), 1–67.

[91] Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Anna Korhonen, and Ivan Vulic. 2021. Crossing the conversational chasm: A primer on multilingual task-oriented dialogue systems. *arXiv preprint arXiv:2104.08570* (2021).

[92] Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo Ponti, and Ivan Vulić. 2022. Natural language processing for multilingual task-oriented dialogue. In *ACL Tutorial Abstracts*. 44–50.

[93] Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. 2022. Data augmentation and learned layer aggregation for improved multilingual language understanding in dialogue. In *ACL Findings*. 2017–2033.

[94] Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *EMNLP*. 2780–2786.

[95] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten de Rijke. 2021. Conversations with search engines: SERP-based conversational response generation. *ACM Transactions on Information Systems* 39, 4 (2021), 1–29.

[96] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research* 65 (2019), 569–631.

[97] Sergio Scalise, Elisabetta Magni, and Antonietta Bisetto. 2009. *Universals of Language Today*. Springer.

[98] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *NAACL-HLT*. 3795–3805.

[99] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems. *Dialogue & Discourse* 9, 1 (2018), 1–49.

[100] Maurizio Serva and Filippo Petroni. 2008. Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)* 81, 6 (2008), 68005.

[101] Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *AAAI*, Vol. 34. 8854–8861.

[102] Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui. 2019. Cross-lingual word embeddings. *Synthesis Lectures on Human Language Technologies* 12, 2 (2019), 1–132.

[103] Georgios P Spithourakis, Ivan Vulić, Michał Lis, Iñigo Casanueva, and Paweł Budzianowski. 2022. Evi: Multilingual spoken dialogue tasks and dataset for knowledge-based enrolment, verification, and identification. *arXiv preprint arXiv:2204.13496* (2022).

[104] Chaoju Tang and Vincent J. van Heuven. 2007. Mutual intelligibility and similarity of Chinese dialects: Predicting judgments from objective measures. *Linguistics in the Netherlands* 24, 1 (2007), 223–234.

[105] Sandra A. Thompson, Robert E. Longacre, Shin Ja J. Hwang, and Timothy Shopen. 2007. *Language typology and syntactic description*. Cambridge University Press.

[106] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[107] Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (Almost) zero-shot cross-lingual spoken language understanding. In *ICASSP*. IEEE, 6034–6038.

[108] Phi Nguyen Van, Tung Cao Hoang, Dung Nguyen Manh, Quan Nguyen Minh, and Long Tran Quoc. 2022. ViWOZ: A multi-domain task-oriented dialogue systems dataset for low-resource language. *arXiv preprint arXiv:2203.07742* (2022).

[109] Piet van Sterkenburg (Ed.). 2008. *Unity and Diversity of Languages*. John Benjamins Publishing.

[110] Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *ACL*. 56–68.

[111] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235* (2023).

[112] Jiapeng Wang and Yihong Dong. 2020. Measurement of text similarity: a survey. *Information* 11, 9 (2020), 421.

[113] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *EACL*. 438–449.

[114] Lindsay J. Whaley. 1996. *Introduction to Typology: The Unity and Diversity of Language*. SAGE publications.

[115] Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *ACL*. 808–819.

[116] Lu Xiang, Junnan Zhu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2021. Robust cross-lingual task-oriented dialogue. *Transactions on Asian and Low-Resource Language Information Processing* 20, 6 (2021), 1–24.

[117] Puyang Xu and Qi Hu. 2018. An end-to-end approach for handling unknown slot values in dialogue state tracking. In *ACL*. 1448–1457.

[118] Ruijian Xu, Chongyang Tao, Jiazhan Feng, Wei Wu, Rui Yan, and Dongyan Zhao. 2021. Response ranking with multi-types of deep interactive representations in retrieval-based dialogues. *ACM Transactions on Information Systems* 39, 4 (2021), 1–28.

[119] Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *EMNLP*. 5052–5063.

[120] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *NAACL-HLT*.

[121] Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2022. ReMeDi: Resources for multi-domain, multi-service, medical dialogues. In *SIGIR*. 3013–3024.

[122] Rui Yan, Weiheng Liao, Dongyan Zhao, and Ji-Rong Wen. 2021. Multi-response awareness for retrieval-based conversations: Respond with diversity via dynamic representation learning. *ACM Transactions on Information Systems* 39, 4 (2021), 1–29.

[123] Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dialog state tracking with reinforced data augmentation. In *AAAI*, Vol. 34. 9474–9481.

[124] Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences* 63, 10 (2020), 2011–2027.

[125] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[126] Zijian Zhao, Su Zhu, and Kai Yu. 2019. Data augmentation with atomic templates for spoken language understanding. In *EMNLP-IJCNLP*. 3637–3643.

[127] Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *ACL*. 1458–1467.

[128] Han Zhou, Ignacio Iacobacci, and Pasquale Minervini. 2022. XQA-DST: Multi-domain and multi-lingual dialogue state tracking. *arXiv preprint arXiv:2204.05895* (2022).

[129] Lei Zuo, Kun Qian, Bowen Yang, and Zhou Yu. 2021. AllWOZ: Towards multilingual task-oriented dialog systems for all. *arXiv preprint arXiv:2112.08333* (2021).

A EXAMPLES OF PROMPT

In Table 11 we show examples of prompts used in the decoder-only framework in DST and NLG task, respectively.

Table 11. Examples of prompts used in the extension of the decoder-only MOLR framework.

<table>
<tr><th>Task</th><th>Prompt</th></tr>
<tr><td>DST</td><td>You are a helpful AI assistant tasked with generating key-value pairs from a dialogue context based on schema.
Task: Slot Extraction aims to extract all slots and corresponding values mentioned in the given dialogue context.
If the value of a slot is mentioned, then the substring is formatted as “inform [slot] [value]”.
If the value of a slot is not mentioned, then the substring is formatted as “request slot [slot]”.
The output is a concatenation of all substrings of all slots.
Schema:
food: the cuisine of the restaurant you are looking for, such as “british”.
area: the location or area of the restaurant, including “centre”, “north”, “west”, “south”, “east”.
price range: price budget for the restaurant, including “cheap”, “moderate”, and “expensive”.
request: the attribute of a restaurant you are looking for, including “address”, “area”, “food”, “phone”, “price range”, “postcode”, “name”.
Example:
{“input”: “<|user|>i want to find a moderately priced restaurant in the west part of town .
what is the address and the postcode ?”,
“output”: “request slot postcode, request slot address, inform price range moderate, inform area west”}</td></tr>
<tr><td>NLG</td><td>You are a helpful AI assistant tasked with generating a response given the current user query and dialogue history.</td></tr>
</table>

B SUMMARY OF ZERO-SHOT CROSS-LINGUAL BENCHMARKS

We summarize all the zero-shot crosslingual results on the DST (Table 12) and NLU (Table 13) datasets, as well as our implementation of mT5 models. We find that mT5 and its variants achieve the state-of-the-art for zero-shot crosslingual adaptation. This justifies our choice mT5 as our base model in the main results.

Table 12. Comparison of dialogue state tracking (DST) models for zero-shot learning from English (EN) to German (DE) and Italian (IT).

		Joint/Request (%)	
Models	**Settings**	**EN → DE**	**EN → IT**
NBT			
XL-NBT [10] (from [58])	Teacher-student NBT+bilingual data augmentation	30.80/68.32	41.23/81.23
MUSE			
MUSE [58]	Word alignment using MUSE [13]	21.57/74.22	20.66/79.09
MUSE+AMLT [58]	MUSE+attention-based bilingual code-switching	36.51/82.99	39.35/84.23
XLM			
XLM [58]	XLM [12] context encoder	16.34/75.73	–
XLM+AMLT [58]	XLM+attention-based bilingual code-switching	33.12/82.96	–
XLM+CLCSA [88]	XLM+multilingual code-switching	48.70/88.30	–
XQA-DST [128]	XLM-R [11]+value span extraction	64.88/–	68.63/–
BERT			
mBERT [58]	mBERT [20] context encoder	14.95/75.31	12.88/76.12
mBERT+AMLT [58]	mBERT+attention-based bilingual code-switching	34.36/86.97	33.35/84.96
mBERT+CLCSA [88]	mBERT+multilingual code-switching	63.20/94.00	61.30/**94.20**
T5 (Ours)			
mT5	Multilingual T5 (small) with 300M parameters	28.42/92.27	32.14/87.22
mT5+AMLT	mT5+bilingual code-switching	40.96/93.37	47.90/87.65
mT5+CLCSA	mT5+multilingual code-switching	**67.86**/**95.80**	**71.15**/88.07

Table 13. Comparison of natural language understanding (NLU) models for zero-shot learning from English (EN) to German (DE) and Italian (IT).

		Intent/Slot F1 (%)	
Models	**Settings**	**EN → ES**	**EN → TH**
RNN			
biRNN [56]	An implantation of bidirectional RNN [107]	46.64/15.41	35.64/12.11
CoVe [98]	biLSTM-CRF based translation model to English	37.13/ 5.35	54.24/ 8.84
mCoVe [98]	Multilingual CoVe [66]	53.34/22.50	66.35/32.52
mCoVe+Auto [98]	mCoVe w/autoencoder objective	53.89/19.25	70.70/35.62
biLSTM [56]	biLSTM w/noise, refinement, delexicalization	90.20/65.79	73.43/32.24
MUSE			
RCSLS [39]	MUSE+relaxed cross-domain similarity local scaling	37.67/22.23	35.12/ 8.72
RCSLS+AMLT [58]	RCSLS+attention-based bilingual code-switching	87.05/57.75	81.44/30.42
Transformers			
Transformer [57]	Transformer w/frozen word embeddings	89.71/67.10	74.68/31.20
Transformer+ORT [57]	Order-reduced transformer	91.46/71.36	75.02/34.61
mBERT			
mBERT [58]	mBERT [20] context encoder	74.15/54.28	26.54/11.34
mBERT+AMLT [58]	mBERT+attention-based bilingual code-switching	87.88/73.89	73.46/27.12
mBERT+CLCSA [88]	mBERT+multilingual code-switching	92.80/75.20	74.80/28.10
XLM-R [27]	XLM-R encoder with 270M parameters	90.70/70.10	71.90/53.10
T5 (Ours)			
mT5	Multilingual T5 with 300M parameters	92.17/71.26	81.38/52.13
mT5+AMLT	mT5+bilingual code-switching	92.77/71.92	91.61/**57.46**
mT5+CLCSA	mT5+multilingual code-switching	**94.71/75.77**	**93.20**/47.02