MTOP: A Comprehensive Multilingual Task-Oriented Semantic Parsing Benchmark

Haoran Li Abhinav Arora Shuohui Chen Anchit Gupta
Sonal Gupta Yashar Mehdad
Facebook
{aimeeli,abhinavarora,shuohui,anchit}@fb.com
{mehdad,sonalgupta}@fb.com

Abstract

Scaling semantic parsing models for task-oriented dialog systems to new languages is often expensive and time-consuming due to the lack of available datasets. Available datasets suffer from several shortcomings: a) they contain few languages b) they contain small amounts of labeled examples per language c) they are based on the simple intent and slot detection paradigm for non-compositional queries. In this paper, we present a new multilingual dataset, called **MTOP**, comprising of 100k annotated utterances in 6 languages across 11 domains. We use this dataset and other publicly available datasets to conduct a comprehensive benchmarking study on using various state-of-the-art multilingual pre-trained models for task-oriented semantic parsing. We achieve an average improvement of +6.3 points on Slot F1 for the two existing multilingual datasets, over best results reported in their experiments. Furthermore, we demonstrate strong zero-shot performance using pre-trained models combined with automatic translation and alignment, and a proposed distant supervision method to reduce the noise in slot label projection.

1 Introduction

With the rising adoption of virtual assistant products, task-oriented dialog systems have been attracting more attention in both academic and industrial communities. One of the first steps in these systems is to extract meaning from the natural language used in conversation to build a semantic representation of the user utterance. Typical systems achieve this by classifying the *intent* of the utterance and tagging the corresponding *slots*. With the goal of handling more complex queries, recent approaches propose hierarchical representations (Gupta et al., 2018) that are expressive enough to capture the task-specific semantics of complex nested queries.

Although, there have been sizable efforts around developing successful semantic parsing models for task-oriented dialog systems in English (Mesnil et al., 2013; Liu and Lane, 2016; Gupta et al., 2018; Rongali et al., 2020), we have only seen limited works for other languages. This is mainly due to the painstaking process of manually annotating and creating large datasets for this task in new languages. In addition to the shortage of such datasets, existing datasets (Upadhyay et al., 2018; Schuster et al., 2019a) are not sufficiently diversified in terms of languages and domains, and do not capture complex nested queries. This makes it difficult to perform more systematic and rigorous experimentation and evaluation for this task across multiple languages.

Building on these considerations and recent advancements on cross-lingual pre-trained models (Devlin et al., 2019; Lample and Conneau, 2019; Conneau et al., 2020), this paper is making an effort to bridge the above mentioned gaps. The main contributions of this paper can be summarized as follows:

- MTOP Dataset: We release an almost-parallel multilingual task-oriented semantic parsing dataset covering **6 languages** and **11 domains**. To the best of our knowledge, this is the first multilingual dataset which contains compositional representations that allow complex nested queries.

- We build strong benchmarks on the released MTOP dataset using state-of-the-art multilingual pre-trained models for both flat and compositional representations. We demonstrate the effectiveness of our approaches by achieving new state-of-the-art result on existing multilingual task-oriented semantic parsing datasets.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2950–2962
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

- We demonstrate strong performance on zero-shot cross-lingual transfer using automatic translation and alignment, combined with a proposed distant supervision approach. We achieve 67.2% exact match accuracy (averaged across 5 languages) without using any target language data compared to best in-language model performance of 77.7%.

2 Related Work

Task-Oriented Semantic Parsing The majority of the work on task-oriented dialog systems has been centered around intent detection and slot filling - for example, the representations used on the ATIS dataset (Mesnil et al., 2013; Liu and Lane, 2016; Zhu and Yu, 2017) and in the Dialog State Tracking Challenge (Williams et al., 2016). This essentially boils down to a text classification and a sequence labeling task, which works great for simple non-compositional queries. For more complex queries with recursive slots, state of the art systems use hierarchical representations, such as the TOP representation (Gupta et al., 2018), that is modeled using Recurrent Neural Network Grammars (Dyer et al., 2016) or as a Sequence to Sequence task (Rongali et al., 2020).

Pre-trained Cross-lingual Representation Over the past few years, pre-trained cross-lingual representations have demonstrated tremendous success in achieving state of the art in various NLP tasks. The majority of the earlier work focuses on cross-lingual emebedding alignment (Mikolov et al., 2013; Ammar et al., 2016; Lample et al., 2018). Schuster et al. (2019b) further extend upon this by aligning contextual word embeddings from the ELMo model (Peters et al., 2018). Later with the success of Transformer (Vaswani et al., 2017) based masked language model pre-training, Devlin et al. (2019) and Lample and Conneau (2019) introduce mBERT and XLM respectively, and Pires et al. (2019) show the effectiveness of these on sequence labeling tasks. Conneau et al. (2020) present XLM-R, a pre-trained multilingual masked language model trained on data in 100 languages, that provides strong gains over XLM and mBERT on classification and sequence labeling tasks.

The models discussed above are encoder-only models. More recently, multilingual seq-to-seq pre-training has become popular. Liu et al. (2020a) introduce mBART, a seq-to-seq denoising auto-encoder pre-trained on monolingual corpora in many languages, which extends BART (Lewis et al., 2020b) to a multilingual setting. More recently, Lewis et al. (2020a) introduced a seq-to-seq model pre-trained on a multilingual multi-document paraphrasing objective, which self-supervises the reconstruction of target text by retrieving a set of related texts and conditions on them to maximize the likelihood of generating the original. Tran et al. (2020) is another contemporary work that mines parallel data using encoder representations and jointly trains a seq-to-seq model on this parallel data.

Cross-Lingual Task-Oriented Semantic Parsing Due to the ubiquity of digital assistants, the task of cross-lingual and multilingual task-oriented dialog has garnered a lot of attention recenty, and few multilingual benchmark datasets have been released for the same. To the best of our knowledge, all of them only contain simple non-compositional utterances, suitable for the intent and slots detection tasks. Upadhyay et al. (2018) release a benchmark dataset in Turkish and Hindi (600 training examples), obtained by translating utterances from the ATIS corpus (Price, 1990) and using Amazon Mechanical Turk to generate phrase level slot annotation on translations. Schuster et al. (2019a) release a bigger multilingual dataset for task-oriented dialog in English, Spanish and Thai across 3 domains. They also propose various modeling techniques such as using XLU embeddings (see Ruder et al. (2017) for literature review) for cross-lingual transfer, translate-train and ELMo (Peters et al., 2018) for target language training. BERT-style multilingual pre-trained models have also been applied to task-oriented semantic parsing. Castellucci et al. (2019) use multilingual BERT for joint intent classification and slot filling, but they don’t evaluate on existing multilingual benchmarks. Instead, they introduce a new Italian dataset obtained via automatic machine translation of SNIPS (Coucke et al., 2018), which is of lower quality. For zero shot transfer, Liu et al. (2020b) study the idea of selecting some parallel word pairs to generate code-switching sentences for learning the inter-lingual semantics across languages and compare the performance using various cross-lingual pre-trained models including mBERT and XLM.

3 Data

Existing multilingual task-oriented dialog datasets, such as Upadhyay et al. (2018); Schuster et al.

Domain	Number of utterances (training/validation/testing)						Intent	Slot
	English	German	French	Spanish	Hindi	Thai	types	types
Alarm	2,006	1,783	1,581	1,706	1,374	1,510	6	5
Calling	3,129	2,872	2,797	2,057	2,515	2,490	19	14
Event	1,249	1,081	1,050	1,115	911	988	12	12
Messaging	1,682	1,053	1,239	1,335	1,163	1,082	7	15
Music	1,929	1,648	1,499	1,312	1,508	1,418	27	12
News	1,682	1,393	905	1,052	1,126	930	3	6
People	1,768	1,449	1,392	763	1,408	1,168	17	16
Recipes	1,845	1,586	1,002	762	1,378	929	3	18
Reminder	1,929	2,439	2,321	2,202	1,781	1,833	19	17
Timer	1,488	1,358	1,013	1,165	1,152	1,047	9	5
Weather	2,372	2,126	1,785	1,990	1,815	1,800	4	4
Total	22,288	18,788	16,584	15,459	16,131	15,195	117	78

Table 1: Summary statistics of the MTOP dataset. The Data is roughly divided into 70:10:20 percent splits for train, eval and test.

(2019a), rely on expensive manual work for preparing guidelines and annotations for other languages; which is probably why they only contain very few languages and few labeled data examples for other languages. Furthermore, annotations will be more complicated and expensive if they were to include compositional queries, where slots can have nested intents. To this end we create an almost parallel multilingual task-oriented semantic parsing corpora which contains **100k examples** in total for **6 languages** (both high and low resource): *English*, *Spanish*, *French*, *German*, *Hindi* and *Thai*. Our dataset contains a mix of both simple and compositional nested queries across **11 domains**, **117 intents** and **78 slots**. Table. 1 shows a summary statistics of our MTOP dataset.

We release the dataset at `https://fb.me/mtop_dataset`.

3.1 Dataset Creation

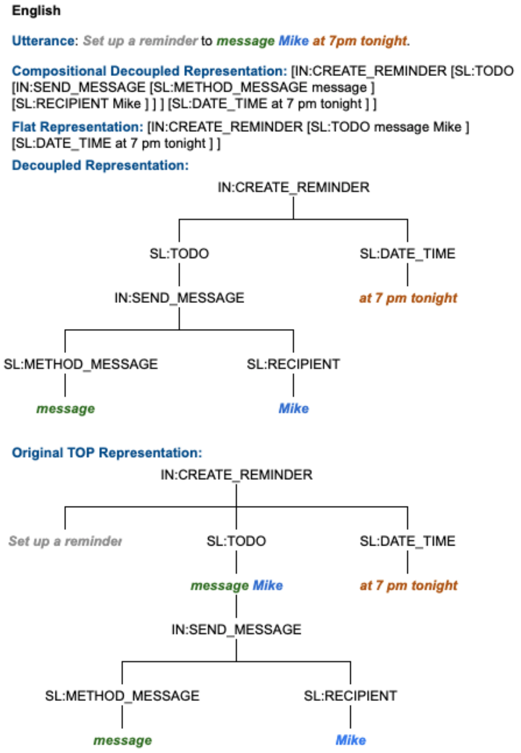
Our approach for creating this dataset consists of two main steps: i) generating synthetic utterances and annotating in English, ii) translation, label transfer, post-processing, post editing and filtering for other 5 languages. Generating the English utterances and their annotations, for the 11 domains, follows the exact process as described in (Gupta et al., 2018). We ask crowdsourced workers to generate natural language sentences that they would ask a system which could assist in queries corresponding to our chosen domains. These queries are labeled by two annotators. A third annotator is used only to adjudicate any disagreements. Once an annotated English dataset is available, we build the multilingual dataset through the following steps:

Translation: We first extract slot text spans from English annotation and present the utterances along with slot text spans to professional translators for translation to the target language. We prepare detailed guidelines, where we ask the translators to ensure that the translation for each slot span is exactly in the same way as it occurs in the translated utterance. For example, when translating the slot span *mom* in utterance *call my mom*, we ask the translators to use the same target language word for *mom*, that they used in the translation for *call my mom*.

Post-processing: After we obtain the translation of utterances and corresponding slot text spans, we use the tree structure of English and fill in the translated slot text spans to construct the annotation in the target languages. Our representation, described in §3.2.1, enables us to reconstruct the annotations.

Post-editing and Quality Control: We further run two rounds of quality control over translated utterances and slots, and revise the data accordingly. In the first round, we ask translators to review and post-edit the errors in translations and slot alignments. In the second round, the constructed target language data is presented to different annotators for a lightweight annotation quality review. 83% of the data was marked as good quality data and passed our quality standards, which can be interpreted as the inter-annotator agreement rate on the translated data. Based on this feedback, we remove low quality annotations from the dataset.

To create this dataset, for each target language we had three translators: two were responsible for translation and the third one for review and edits.



English

Utterance: *Set up a reminder* to *message Mike at 7pm tonight*.

Compositional Decoupled Representation: [IN:CREATE_REMINDER [SL:TODO [IN:SEND_MESSAGE [SL:METHOD_MESSAGE message] [SL:RECIPIENT Mike]]] [SL:DATE_TIME at 7 pm tonight]]

Flat Representation: [IN:CREATE_REMINDER [SL:TODO message Mike] [SL:DATE_TIME at 7 pm tonight]]

Decoupled Representation:

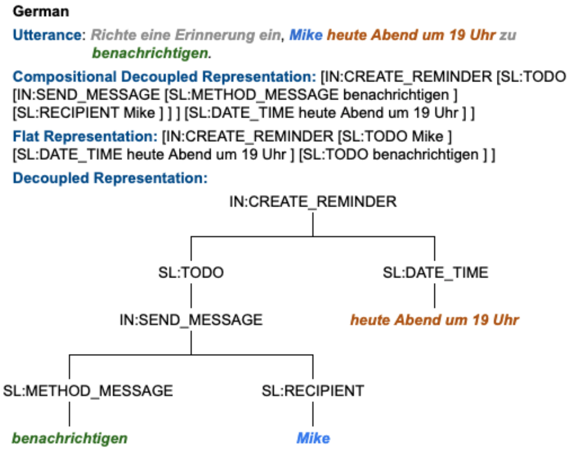
Original TOP Representation:

Figure 1: An English example from the data, showing its *flat* representation and *compositional decoupled* representation and a comparison between the decoupled and the original TOP representations in tree format.

All the translators were professional translators, with native or close to native speaker skills. The overall time spent was 15 to 25 days for each language. Even though we run rigorous quality control, a dataset built by translation is bound to have few errors, such as using words or phrases that are not commonly used in spoken language.

3.2 Data Format

In this dataset, we release two kinds of representations, which we refer to as *flat* representations and *compositional decoupled* representations, that are illustrated in Figure 1 for an English utterance. Most existing annotations for task-oriented dialog systems follow the intent classification and slot tagging paradigm, which is what we refer to as the flat representation. Since our data contains compositional utterances with nested slots with intents within them, flat representations are constructed by only using the top level slots. We include the flat representation so that the data and the discussed modeling techniques are comparable to other task-oriented dialog benchmarks. To ensure the reproducibility of our results, we also release the tokenized version of utterances obtained via our in-house multilingual tokenizer.



German

Utterance: *Richte eine Erinnerung ein, Mike heute Abend um 19 Uhr zu benachrichtigen*.

Compositional Decoupled Representation: [IN:CREATE_REMINDER [SL:TODO [IN:SEND_MESSAGE [SL:METHOD_MESSAGE benachrichtigen] [SL:RECIPIENT Mike]]] [SL:DATE_TIME heute Abend um 19 Uhr]]

Flat Representation: [IN:CREATE_REMINDER [SL:TODO Mike] [SL:DATE_TIME heute Abend um 19 Uhr] [SL:TODO benachrichtigen]]

Decoupled Representation:

Figure 2: German utterance constructed from the English example of Figure 1. Even though the slot text order changed, we can still easily build a decoupled representation with the same structure.

3.2.1 Compositional Decoupled Representation

Gupta et al. (2018) demonstrate the inability of flat representations to parse complex compositional requests and propose a hierarchical annotation scheme (TOP representation) for semantic parsing, that allows the representation of such nested queries. We further use a representation, called the decoupled representation, that removes all the text from the TOP representation that does not appear in a leaf slot, assuming this text does not contribute to the semantics of the query. Figure 1 highlights the difference between this decoupled representation and the original TOP representation. The decoupled representation makes the semantic representation more flexible and allows long-distance dependencies within the representation. It also makes translation-based data creation approach feasible for different languages despite syntactic differences, as the representation is *decoupled* from the word order of the utterance. For example, in the German translation of the English example as shown in Figure 2, translations of *message* and *Mike* were separated by other words between them. However, it is straight forward to construct a decoupled representation as the representation is not bound by a word-order constraint.

4 Model Architecture

4.1 Joint intent and slot tagging for flat representation

For flat representation, where there is a single top-level intent, the traditional way is to model it as an intent classification and a slot tagging problem. Our baseline model is a bidirectional LSTM intent slot model as described in Liu and Lane (2016); Zhang and Wang (2016) with pre-trained XLU embeddings. Since existing pre-trained XLU embeddings (e.g., MUSE (Lample et al., 2018)) don't provide embedding for Hindi and Thai, we train our own using multiCCA following Ammar et al. (2016). Compared to previous state-of-the-art work on existing multilingual task-oriented parsing datasets (Liu et al., 2020b; Castellucci et al., 2019) which use Multilingual BERT, we use XLM-R (Conneau et al., 2020) since it's shown to outperform Multilingual BERT in cross-lingual performance on a variety of tasks. Specifically we use XLM-R Large in all our experiments. We use the same model architecture as in Chen et al. (2019) and replace BERT encoder with XLM-R encoder.

4.2 Seq-to-seq for hierarchical representation

Even though there are few existing works on cross lingual transfer learning for parsing flat representations, to the best of our knowledge, we are not aware of any other work that studies cross-lingual transfer for parsing complex queries in task-oriented dialog. In this section, we outline our modeling approaches for the compositional decoupled representation discussed in §3.2.1.

Seq-to-seq with Pointer-generator Network Our model adopts an architecture similar to Rongali et al. (2020), where source is the utterance and target is the compositional decoupled representation described in §3.2.1. Given a source utterance, let $[\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_n]$ be the encoder hidden states and $[\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_m]$ be the corresponding decoder hidden states. At decoding time step t, the model can either generate an element from the ontology with generation distribution $\boldsymbol{p}_t^{\mathrm{g}}$, or copy a token from the source sequence with copy distribution $\boldsymbol{p}_t^{\mathrm{c}}$. Generation distribution is computed as:

$$\boldsymbol{p}_t^{\mathrm{g}} = \mathrm{softmax}\left(\mathrm{Linear}_{\mathrm{g}}[\boldsymbol{d}_t]\right)$$

Copy distribution is computed as:

$$\boldsymbol{p}_t^{\mathrm{c}}, \boldsymbol{\omega}_t = \mathrm{MHA}\left(\boldsymbol{e}_1, ..., \boldsymbol{e}_n; \mathrm{Linear}_{\mathrm{c}}[\boldsymbol{d}_t]\right)$$

where MHA stands for Multi-Head Attention (Vaswani et al., 2017) and $\boldsymbol{\omega}_t$ is the attended vector used to compute the weight of copying $\boldsymbol{p}_t^{\mathrm{w}}$:

$$p_t^{\mathrm{w}} = \mathrm{sigmoid}\left(\mathrm{Linear}_{\alpha}\left[\boldsymbol{d}_t; \boldsymbol{\omega}_t\right]\right)$$

The final probability distribution is computed as a mixture of the generation and copy distributions:

$$\boldsymbol{p}_t = p_t^{\mathrm{w}} \cdot \boldsymbol{p}_t^{\mathrm{g}} + (1 - p_t^{\mathrm{w}}) \cdot \boldsymbol{p}_t^{\mathrm{c}}.$$

As a baseline, we use a standard LSTM encoder-decoder architecture with XLU embeddings. We also experiment with various transformer-based state of the art multilingual pre-trained models to improve upon the baseline. We use both pre-trained encoder-only models as well as pre-trained seq-to-seq encoder and decoder models. Here we outline the different models that we experimented with:

- **XLM-R** encoder, pre-trained with masked language model objective in 100 languages. For decoder, we use randomly initialized transformer decoder as in Vaswani et al. (2017).
- **mBART** (Liu et al., 2020a) is pre-trained seq-to-seq model using denoising autoencoder objective on monolingual corpora in 25 languages.
- **mBART on MT**: Machine translation is another common task for pre-training multilingual models. We follow Tang et al. (2020) to further fine-tune mBART on English to 25 languages translation task.
- **CRISS** (Tran et al., 2020) is pre-trained on parallel data in an unsupervised fashion. It iteratively mines parallel data using its own encoder outputs and trains a seq-to-seq model on the parallel data. CRISS has been shown to perform well on sentence retrieval and translation tasks.
- **MARGE** (Lewis et al., 2020a) is learned with an unsupervised multi-lingual multi-document paraphrasing objective. It retrieves a set of related texts in many languages and conditions on them to maximize the likelihood of generating the original text. MARGE has shown to outperform other models on a variety of multilingual benchmarks including document translation and summarization.

5 Experiments

We conduct thorough experiments on the new dataset we describe in in §3. To further demonstrate the effectiveness of our proposed approaches,

We provide reproducibility details and all hyperparameters in Appendix A we also run additional experiments on the existing multilingual task-oriented semantic parsing datasets including *Multilingual ATIS* (Upadhyay et al., 2018) and *Multilingual TOP* (Schuster et al., 2019a). Note that both these data sets only include flat representation, while our data set contains hierarchical representations.

5.1 Experimental Settings

For all benchmarks, we have three different evaluation settings:

- IN-LANGUAGE MODELS: We only use target language training data.
- MULTILINGUAL MODELS: We use training data in all available languages and train a single model for multiple languages.
- ZERO-SHOT TARGET LANGUAGE MODELS: We only use English data during training.

Next in each subsection we talk about details of approaches we use in these experiments.

5.1.1 Translate and Align

With zero or few target language annotated examples, *translate-train* is a common approach to augment target language training data. For semantic parsing tasks, besides translation we need alignment to project slot annotations to target language. This process is similar to how we collect our dataset, but using machine translation and alignment methods. For translation, we use our in-house machine translation system. We also tried other publicly available translation APIs and didn't find significant difference in final task performance. For alignment, we experimented with both, using attention weights from translation as in Schuster et al. (2019a) and fastalign (Dyer et al., 2013) and found data generated through fastalign leads to better task performance. Thus we only report results that use fastalign.

5.1.2 Multilingual Training

With the advancement of multilingual pre-trained models, a single model trained on multiple languages has shown to outperform in-language models (Conneau et al., 2020; Hu et al., 2020). As a result, we also experiment with multilingual training on our benchmark, including training jointly on all in-language data and training on English plus translated and aligned data in all other languages for the zero-shot setting. Instead of concatenating data in all languages together as in Conneau et al. (2020), we adopt a multitask training approach where for each batch we sample from one language based on a given sampling ratio so that languages with fewer training data can be upsampled. We found this setting to perform better than mixed-language batches in our experiments.

5.1.3 Distant Supervision in Zero-Shot Setting for Flat Representations

Alignment models are not perfect, especially for low resource languages. To combat the noise and biases introduced in slot label projection, we experiment with another distant supervision approach in the zero-shot setting for learning flat representation models. We first concatenate the English utterance and its corresponding translation (using machine translation) in target language as input and then replace the English slot text with MASK token at random (30% of the time, chosen empirically as a hyper-parameter). With the masked source utterance and the translated utterance as the concatenated input, we train a model to predict the overall intent and slot labels on the original English source. In this way, the MASK token can also attend to its translation counterpart to predict its label and the translated slot text could be distantly supervised by English labeled data.

6 Results and Discussions

6.1 Results on MTOP

Flat Representation Results Table. 2 shows the result on our MTOP dataset for all languages, using the flat representation. For both in-language and multilingual settings, XLM-R based models significantly outperform the BiLSTM models using XLU. We also observe that multilingual models outperform in-language models. Interestingly, for Hindi and Thai (both non-European languages), the improvements from multilingual training are considerably higher for XLM-R as compared to XLU BiLSTM. This observation highlights the remarkable cross-lingual transferability of the pre-trained XLM-R representations where fine-tuning on syntactically different languages also improves target language performance.

For zero-shot cross-lingual transfer, we restrict ourselves to an XLM-R baseline to explore improvements using translate and align, and the distant supervision techniques as described in 5.1.1 and 5.1.3 respectively. Our results demonstrate that

Model	en	es	fr	de	hi	th	Avg(5 langs)
			(Exact Match Accuracy)				
In-language models (only use target language training data)							
XLU biLSTM	78.2	70.8	68.9	65.1	62.6	68	67.1
XLM-R	85.3	81.6	79.4	76.9	76.8	73.8	77.7
Multilingual models (use training data from multiple languages)							
XLU biLSTM	78.2	73.8	71.5	65.8	63.1	68.7	68.6
XLM-R	86.3	83.6	81.8	79.2	78.9	76.7	80
Zero-shot target language models (only use English training data)							
XLM-R on EN	N/A	69.1	65.4	64	55	43.8	59.5
XLM-R with mask in §5.1.3	N/A	68	69.5	**69.2**	**63.3**	35.3	61.1
XLM-R on EN + translate align §5.1.1	N/A	74.5	**72.6**	64.7	58.3	**56.5**	65.3
XLM-R with mask + translate align	N/A	**74.6**	72.2	65.7	62.5	53.2	**65.6**

Table 2: Results on flat representation for 6 languages. We report exact match accuracy in this table. More metrics including intent accuracy and slot F1 is in Table 5 in Appendix. Notice that average is calculated across 5 languages except English to be comparable to zero-shot results. Best result for zero-shot is in bold. Taking best zero shot setting for each language, average exact match accuracy is 67.2. Note that for zero-shot setting, we only use EN train and eval data without any target language data.

distant supervision is able to considerably improve over the baselines for French, German and Hindi, while there is a small drop for Spanish. In the same setting, performance for Thai significantly degrades compared to the baseline. We suspect this is due to imperfect Thai tokenization that leads to learning noisy implicit alignments through distant supervision. The translate and align approach consistently improves over the baseline for all languages. It also performs better than distant supervision for all languages except German and Hindi. Our hypothesis is that the compounding nature of German inhibits the learning of hard alignment from fastalign. In summary, the XLM-R trained on all the 6 languages significantly outperforms all other models for this task.

In Appendix B, we further report intent accuracy and slot F1 metrics for the flat representation, as these are commonly used metrics in previous benchmarks for intent-slot prediction (Price, 1990; Schuster et al., 2019a).

Compositional Decoupled Representation Table. 3 shows the results on our MTOP dataset using compositional decoupled representation. In all settings, using multilingual pre-trained models significantly outperform the baseline. Surprisingly, mBART doesn't demonstrate strong performance compared to other models with fine-tuning on our task, even though fine-tuning BART on English achieves the best performance on English data. We hypothesize that mBART was under-trained for many languages and did not learn good cross-lingual alignments. In order to prove our hypothesis, we further fine-tune mBART on English to 25 languages translation task. The obtained mBART fine-tuned on translation significantly outperform the original mBART. The performance of CRISS and MARGE are at par with each other and among our best performing models across 5 languages, except Thai. XLM-R with random decoder performs the best on Thai. We believe this is because neither CRISS nor MARGE are pre-trained on Thai, while XLM-R pre-training includes Thai.

Similar to previous observations, multilingual training improves over the monolingual results. With multilingual training, XLM-R and CRISS are the best performing models for every language. Since XLM-R uses a randomly initialized decoder, it makes intuitive sense that such a decoder is better trained with multilingual training and thus obtains higher gains from more training data. Interestingly, mBART performance also improves a lot, which is another evidence that it was originally under-trained, as discussed in the previous paragraph. In the zero-shot setting, using the models fine-tuned on English does not perform well. In fact Thai zero shot using CRISS gives a 0 exact match accuracy, as the model was not pre-trained on any Thai data. Both XLM-R and CRISS show significant improve-

Model	**en**	**es**	**fr**	**de**	**hi**	**th**	**Avg(5 langs)**
			(Exact Match Accuracy)				
In-language models (only use target language training data)							
XLU biLSTM	77.8	66.5	65.6	61.5	61.5	62.8	63.6
XLM-R encoder + random decoder	83.9	76.9	74.7	71.2	70.2	**71.2**	72.8
mBART	81.8	75.8	68.1	69.1	67.6	61.2	68.4
mBART on MT	**84.3**	77.2	74.4	70.1	69.2	66.9	71.6
CRISS	84.2	**78**	**75.5**	**72.2**	**73**	68.8	**73.5**
MARGE	84	77.7	75.4	71.5	70.8	70.8	73.2
Multilingual models (use training data from multiple languages)							
XLM-R encoder + random decoder	83.6	**79.8**	**78**	74	74	**73.4**	**75.8**
mBART	83	78.9	76	72.9	72.8	68.8	73.9
CRISS	**84.1**	79.1	77.7	**74.4**	**74.7**	71.3	75.4
Zero-shot target language models (only use English training data)							
XLM-R on EN	N/A	50.3	43.9	42.3	30.9	26.7	38.8
XLM-R on EN + translate align	N/A	71.9	70.3	62.4	63	**60**	**65.5**
CRISS on EN	N/A	48.6	46.6	36.1	31.2	0	32.5
CRISS on EN + translate align	N/A	**73.3**	**71.7**	**62.8**	**63.2**	53	64.8

Table 3: Results on compositional decoupled representation for 6 languages. Metric is exact match accuracy. Average is calculated across 5 languages except English. Best result for each setting is in bold. For reference, exact match accuracy for BART model in-language training for en is 84.6.

Model	**Multilingual ATIS**		**Multilingual TOP**	
	hi	**tr**	**es**	**th**
In-language models (only use target language training data)				
Original paper	-/-/74.6	-/-/75.5	74.8/96.6/83.0	84.8/96.6/90.6
XLM-R	53.6/80.6/84.4	52.6/90.0/80.4	84.3/98.9/90.2	90.6/97.4/95
Multilingual models (use training data from multiple languages)				
original paper (bilingual)	-/-/80.6	-/-/78.9	76.0/97.5/83.4	86.1/96.9/91.5
XLM-R ALL	62.3/85.9/87.8	65.7/92.7/86.5	83.9/99.1/90	91.2/97.7/95.4
Zero-shot target language models (only use English training data)				
Original paper	N/A	N/A	55/85.4/72.9	**45.6/95.9/55.4**
MBERT MLT	N/A	N/A	-/87.9/73.9	-/73.46/27.1
XLM-R on EN	40.3/80.2/76.2	15.7/78/51.8	**79.9/97.7/84.2**	35/90.4/46
XLM-R with mask	49.4/85.3/84.2	19.7/79.7/60.6	76.9/98.1/85	23.5/95.9/30.2
XLM-R EN + translate align	53.2/85.3/84.2	**49.7/91.3/80.2**	66.5/98.2/75.8	43.4/97.3/52.8
XLM-R mask + translate align	**55.3/85.8/84.7**	46.4/89.7/79.5	73.2/98/83	41.2/96.9/52.8

Table 4: Results on Multilingual ATIS and Multilingual TOP, metrics are exact match accuracy / intent accuracy / slot F1 respectively. For zero-shot, first line is from original dataset paper. Best result for zero-shot is in bold.

ments when they utilized the machine translated and aligned data.

6.2 Results on Existing Benchmarks

Table. 4 shows results on two previously released multilingual datasets: Multilingual ATIS and Multilingual TOP. Similar to our findings in 6.1, XLM-R based models significantly outperform the best results reported by the original papers and sets a new state-of-the-art on these benchmarks. Also, multilingual models trained on all available languages further improve the result.

For Multilingual ATIS, in the zero-shot setting, our distant supervised masking strategy shows considerable gains compared to direct transfer using English. Using translate and aligned data also helps in improving the results significantly. When multi-task trained together with masked data, it achieves the best zero-shot performance on Hindi. For both languages (Hindi and Turkish) this comes very close to the performance using target language training data.

For multilingual TOP, direct transfer proves to be effective for Spanish, direct transfer from English overall yield better result than what's reported in Mixed-Language Training (MLT) with MBERT (Liu et al., 2020b). While masking and translating generated data degrade its performance. Based on our error analysis, we find that tokenization mismatch, derived from translation data, causes such performance drop due to errors in slot text boundaries. For Thai, all our translation-based techniques perform worse than translate-train results from original paper. We attribute this primarily to the tokenization difference between our translated data and original test data. Unlike Spanish, Thai is much more sensitive to tokenization as it rarely uses whitespace.

7 Conclusion

In this paper, we release a new multilingual task-oriented semantic parsing dataset called **MTOP** that covers 6 languages, including both flat and compositional representations. We develop strong and comprehensive benchmarks for both representations using state-of-the-art multilingual pre-trained models in both zero-shot and with target language settings. We hope this dataset along with proposed methods benefit the research community in scaling task-oriented dialog systems to more languages effectively and efficiently.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multi-lingual intent detection and slot filling in a joint bert-based model. *arXiv preprint arXiv:1907.02884*.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648. The Association for Computational Linguistics.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *HLT-NAACL*, pages 199–209. The Association for Computational Linguistics.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Pavel Izmailov, Dmitry Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI 2018)*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization.

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

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *ICLR (Poster)*. OpenReview.net.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. *arXiv preprint arXiv:2006.15020*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 685–689. ISCA.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020b. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8433–8440. AAAI Press.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*, pages 3771–3775.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *ACL (1)*, pages 4996–5001. Association for Computational Linguistics.

P. J. Price. 1990. Evaluation of spoken language systems: the atis domain. In *HLT*. Morgan Kaufmann.

Subendhu Rongali, Luca Soldaini, Emilio Monti, and Wael Hamza. 2020. Don't parse, generate! a sequence to sequence architecture for task-oriented semantic parsing. *arXiv preprint arXiv:2001.11458*.

Sebastian Ruder, Ivan Vulic, and Anders Sogaard. 2017. A survey of cross-lingual word embedding models. Cite arxiv:1706.04902.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019a. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019b. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. *arXiv preprint arXiv:2006.09526*.

Shyam Upadhyay, Manaal Faruqui, Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *ICASSP*, pages 6034–6038. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2993–2999. IJCAI/AAAI Press.

Su Zhu and Kai Yu. 2017. Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding. In *ICASSP*, pages 5675–5679. IEEE.

A Training Details

Settings for MTOP results in Table. 2 For fine-tuning XLM-R, we use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.99, \epsilon = 1e-6$ and batch size of 16. We fine-tune for 20 epochs and search over learning rates $\in \{1, 2, 3\}e-5$ on dev set. All XLM-R models were run on single 32GB V100 Nvidia GPU.

For the XLU models in Table. 2, we use 300 dim XLU embeddings and feed them to a 2-layer 200 dim BiLSTM. The intent classification head contains an attention pooling layer as described in Lin et al. (2017) with with attention dim 128 followed by a 200 dim linear projection before the softmax. The slot tagging head also contains a 200 dim linear layer followed by a CRF decoder. We use the we use the Adam optimizer with the same settings as above and a batch size of 32 for 40 epochs. The learning rate and BiLSTM dropouts are picked via a param sweep over the dev set.

Settings for MTOP results in Table. 3 For training seq-2-seq models, we use stochastic weight averaging (Izmailov et al., 2018) with Lamb optimizer (You et al., 2019) and exponential learning rate decay for all models. For fine-tuning pre-trained models: we use batch size of 16 for all models except Marge, we use batch size 4 for Marge since we were not able to fit larger batch size into 32GB memory; We finetune for 50 epochs and again search over learning rates on dev set.

For copy pointer We use 1 layer multihead attention(MHA) with 4 attention heads to get copy distribution. For seq-2-seq model with XLM-R encoder, the decoder is a randomly initialized 3-layer transformer, with hidden size 1024 and 8 attention heads. XLM-R encoder (24 layers) is larger than mBART/CRISS/MARGE encoder (12 layers) so we were not able to fit a larger decoder into GPU memory.

For the XLU models specifically we use a 2-layer BiLSTM encoder with a hidden dimension of 256. For the decoder, we use a 2-layer LSTM with 256 dimension and a single attention head. Similar to the flat models, learning rate and LSTM dropouts are picked via a param sweep over the dev set.

Settings for other benchmark results in Table. 4 We use the same setting as described for Table. 2 except for multilingual ATIS which doesn't have dev set, we just use the checkpoint after a fixed number of epochs.

B More Results

We report additional metrics for our experiments in this section. Table. 5 contains the intent accuracy and slot F1 metrics of models for flat representation.

Model	**en**	**es**	**fr**	**de**	**hi**	**th**
			(Intent Accuracy / Slot F1)			
In-language models (only use target language training data)						
XLU biLSTM	94.0/88.6	90.1/83.0	89.6/81.8	88.8/81.4	85.9/79.6	91.2/80.4
XLM-R	96.7/92.8	95.2/89.9	94.8/88.3	95.7/88.0	94.4/87.5	93.4/85.4
Multilingual models (use training data from multiple languages)						
XLU biLSTM	94.6/88.4	91.3/84.6	91.3/83.0	90.3/81.2	87.6/78.9	91.9/80.5
XLM-R	97.1/93.2	96.6/90.8	96.3/89.4	96.7/88.8	95.4/88.4	95.1/86.3
Zero-shot target language models (only use English training data)						
XLM-R on EN	N/A	93.5/81.7	90.7/81.6	91.2/78.7	88.4/71.8	88.0/63.3
XLM-R with mask in §5.1.3	N/A	94.7/81.0	93.9/82.0	**94.0/81.8**	**94.1/77.3**	92.0/56.4
XLM-R on EN + translate align §5.1.1	N/A	96.2/84.6	**95.4/82.7**	96.1/78.9	94.7/72.7	**92.7/70.0**
XLM-R with mask + translate align	N/A	**96.3/84.8**	95.1/82.5	94.8/80.0	94.2/76.5	92.1/65.6

Table 5: Intent Accuracy / Slot F1 for models in Table 2.