Multilingual Neural Semantic Parsing for Low-Resourced Languages

Menglin Xia
Amazon Research Cambridge
ximengli@amazon.com

Emilio Monti
Amazon Research Cambridge
monti@amazon.com

Abstract

Multilingual semantic parsing is a cost-effective method that allows a single model to understand different languages. However, researchers face a great imbalance of availability of training data, with English being resource rich, and other languages having much less data. To tackle the data limitation problem, we propose using machine translation to bootstrap multilingual training data from the more abundant English data. To compensate for the data quality of machine translated training data, we utilize transfer learning from pretrained multilingual encoders to further improve the model. To evaluate our multilingual models on human-written sentences as opposed to machine translated ones, we introduce a new multilingual semantic parsing dataset in English, Italian and Japanese based on the Facebook Task Oriented Parsing (TOP) dataset. We show that joint multilingual training with pretrained encoders substantially outperforms our baselines on the TOP dataset and outperforms the state-of-the-art model on the public NLMaps dataset. We also establish a new baseline for zero-shot learning on the TOP dataset. We find that a semantic parser trained only on English data achieves a zero-shot performance of 44.9% exact-match accuracy on Italian sentences.

1 Introduction

Semantic parsing is defined as the task of parsing a natural language sentence into a logical form that represents its meaning. The logical form, or sometimes called the meaning representation language (MRL) expression, can be executed against a knowledge base to extract information; therefore, semantic parsing often finds its application in question answering, code generation, information retrieval, etc. Due to its wide range of applications, semantic parsing has drawn a lot of research interest. Among them, neural semantic parsing methods have gained popularity in recent years due to their good results (Dong and Lapata, 2018). Neural semantic parsing often formulates the task as a machine translation problem and uses neural networks to translate the sentences into MRL expressions.

Multilingual neural semantic parsing is a cost-effective method that allows a single model to understand different languages. However, similar to other machine-learning based approaches, neural semantic parsing requires large amounts of training data. To understand texts in different languages, semantic parsing models need training data for each target language. Unfortunately, researchers face a great imbalance of availability of training data for semantic parsing: while we have lots of data in English, the data in non-English languages is often scarce. Although there is a growing number of datasets published for semantic parsing in English, very few datasets are available in other languages. Moreover, manually annotating data for semantic parsing is difficult and time-consuming, as it requires a lot of training and effort for annotators to write MRLs.

Instead of manually annotating semantic parsing data in low-resourced languages, can we bootstrap training data for multilingual semantic parsing from the more abundant English data? In this paper, we aim to tackle the data limitation problem for multilingual semantic parsing with machine translation. We machine translate English sentences into target non-English languages and make use of the alignment information in the English MRL to create MRL annotations in other languages (see Section 3). We then describe our methods to build multilingual semantic parsing models on the machine translated training data (see Section 4). To train the multilingual semantic parser, we mix the training data from all languages together and train a model from scratch (see Section 4.1). We base our neural semantic parser on the sequence-to-sequence model with pointer mecha-

Proceedings of the 10th Conference on Lexical and Computational Semantics, pages 185–194
August 5–6, 2021, Bangkok, Thailand (online)

nism (Sutskever et al., 2014; Vinyals et al., 2015), where both the natural language question and the target MRL are treated as sequences of tokens and the parser learns from the training data a mapping to translate questions into MRLs.

The machine translation-based data generation method allows us to easily extend English data to other languages. However, the quality of the bootstrapped training data is constrained by the accuracy of the machine translation model and other components of the generation method, such as alignment. To mitigate the problem of data quality of the machine translated training data, we make use of transfer learning with pretrained multilingual encoders to further improve the multilingual semantic parsing model (see Section 4.2).

To evaluate the model performance on sentences written by human as opposed to machine translated ones, we introduce a new multilingual semantic parsing dataset based on the Facebook Task Oriented Parsing (TOP) dataset (Gupta et al., 2018). We compare our method against several baselines, including monolingual models and a popular technique in literature that relies on translating the utterances and using an English model to understand them (see Section 4.3). We report the experimental results and our analysis in Section 5. To show that our multilingual semantic parsing models also work with human-generated training data and to compare them against previous work, we report the performance of our models on the public multilingual NLMaps dataset in Section 5.3.

Apart from bootstrapping training data, zero-shot learning is also a technique that allows a multilingual model to generalize to low-resourced languages. We study how the multilingual semantic parsers with pretrained encoders can generalize to other languages in a zero-shot scenario (see Section 5.4).

Our main contributions are as follows:

1. We propose a method to automatically generate training data for multilingual semantic parsing from existing English data via machine translation and we use pretrained multilingual encoders to compensate for the data quality. We release a new multilingual semantic parsing dataset in English, Italian and Japanese based on the public TOP dataset, with ~30k machine-translated training and validation data and ~8k manually translated test data for each language. The dataset is available for download at: https://github.com/awslabs/multilingual-top.

2. We show that our multilingual semantic parsing model achieves state-of-the-art performance, outperforming several baselines on the TOP dataset and existing work on the public NLMaps dataset.

3. We establish a new baseline for zero-shot learning on the TOP dataset with semantic parsing model finetuned from pretrained multilingual encoders.

2 Background and Related Work

Semantic parsing has been studied for a few decades. Earlier methods on semantic parsing rely on defining semantic rules to parse the input sentence (Zelle and Mooney, 1996; Zettlemoyer and Collins, 2005). With recent advances in neural networks, there is a trend of formulating semantic parsing as a machine translation problem. In particular, the sequence-to-sequence model (Sutskever et al., 2014) is commonly used in recent works on semantic parsing (Dong and Lapata, 2018; Jia and Liang, 2016; Zhong et al., 2017). Typically, they use a neural network encoder to encode the utterance sentence into a latent vector representation and use a decoder conditioned on the latent representation to predict the MRL as a sequence of symbols.

Due to the research interest in semantic parsing, many public datasets have been made available for English semantic parsing, ranging from small datasets that contain only a few hundred or a few thousand examples, such as GeoQuery (Zelle and Mooney, 1996) and ATIS (Dahl et al., 1994), to larger datasets with tens of thousands of question-logical form pairs, such as WikiSQL (Zhong et al., 2017) and Overnight (Wang et al., 2015).

Multilingual semantic parsing, however, has only begun to draw research attention in more recent years. Therefore, very few datasets have been published for semantic parsing in non-English languages. So far, almost all of the multilingual semantic parsing datasets are manually translated from their English versions. Due to the cost of manual translation, they are limited to small datasets. For example, Jones et al. (2012) translated the GeoQuery dataset into German, Greek, and Thai. Susanto and Lu (2017) translated the ATIS dataset

into Indonesian and Chinese. Haas and Riezler (2016) created the NLMaps dataset which contains around 2,400 queries to a geographic database in English. The authors translated the queries into German but kept the MRL annotation the same as that for English. Apart from the semantic parsing datasets for question answering, there are some multilingual datasets with other logical form representations, such as multilingual GraphQuestions with graphs as the meaning representation (Reddy et al., 2017), Parallel Meaning Bank with DRT (Discourse Representation Theory) representation (Abzianidze et al., 2017), and multilingual AMR test set with Abstract Meaning Representation (Damonte and Cohen, 2018). The logical form representation in these datasets are very different from the MRLs used for question answering and thus cannot be easily harnessed by many semantic parsers.

Among the limited literature on multilingual semantic parsing, several different methods have been proposed. The first attempts on multilingual semantic parsing (Haas and Riezler, 2016; Damonte and Cohen, 2018) use statistical/neural machine translation methods to translate non-English questions into English and rely on using an English semantic parser to parse all the utterances. Annotation projection is an alternative technique to deal with the lack of multilingual data. It maps the annotation from one language to another using word alignment. It has been applied to many NLP applications, including POS tagging (Yarowsky et al., 2001), role-labeling (Akbik et al., 2015), semantic CCG parsing (Evang and Bos, 2016), and AMR parsing (Damonte and Cohen, 2018). In addition, Susanto and Lu (2017) approached multilingual semantic parsing with a multi-task learning technique. They used separate encoders to encode sentences in different languages and used a shared decoder to predict the MRL. Duong et al. (2017) used cross-lingual word embeddings in a sequence-to-sequence model. They observed that using cross-lingual word embeddings improves the results on both English and German over their baseline models on the NLMaps dataset. They also compared training a model with a single encoder on multilingual data against training with separate encoders for each language and found that keeping separate encoders actually harms semantic parsing accuracy. Based on their observation, we will use a single encoder for multiple languages in our experiments.

Question:
`Any festivals this weekend`

Hierarchical intent-slot representation:
```
[IN:GET_EVENT Any
[SL:CATEGORY_EVENT festivals ]
[SL:DATE_TIME this weekend ] ]
```

Adapted MRL representation:
```
[IN:GET_EVENT
[SL:CATEGORY_EVENT festivals ]
[SL:DATE_TIME this weekend ] ]
```

Table 1: An example of the English TOP dataset

3 Multilingual Semantic Parsing Data

To tackle the data scarcity problem for multilingual semantic parsing, we aim to utilize machine translation to automatically generate training data from the more abundant English data for other languages. In this section, we introduce the English semantic parsing dataset we are using and describe our strategy to bootstrap training data for multilingual semantic parsing.

3.1 English Semantic Parsing Data

We use the Facebook Task Oriented Parsing (TOP) dataset (Gupta et al., 2018) as our source English semantic parsing data. The TOP dataset contains around 44k navigation and event questions created by crowd-sourced workers. The questions are annotated to semantic frames comprising of hierarchical intents and slots. We adapted the original intent-slot representation to a representation that is more similar to other question answering MRLs. More specifically, we dropped the text mentions in the intent label and kept only the entity text in the slot label. The resulting MRL is still a valid meaning representation because the text in the intent label does not affect the execution of the query on a knowledge base. Table 1 shows an example of the original TOP data and its corresponding MRL representation in the adapted task.

We also remove the utterances where the root intent is `IN:UNSUPPORTED`, as it is a noisy catch-all class for out-of-domain utterances. The final dataset contains 28,414 training, 4,032 validation, and 8,241 test data points.

3.2 Bootstrapping Multilingual Semantic Parsing Data

Creating multilingual semantic parsing data from the English data is not a trivial task, because the

Question (English):
Any festivals|x_0 this|x_1 weekend|x_1

MRL:
[IN:GET_EVENT
[SL:CATEGORY_EVENT x_0]
[SL:DATE_TIME x_1]]

Table 2: Replacing text in the MRL with placeholder tokens and marking the positions of placeholder tokens in the question (on the same example as in Table 1).

MRL annotation is highly intertwined with the input question. Directly translating the text in the MRL into another language is likely to generate an incorrect MRL, as it may not match the translation of the input question. In order to obtain valid multilingual equivalents of both the natural language question and its meaning representation, rather than translating the MRL directly, we apply a similar method to annotation projection. We make use of the text alignment information between the question and the MRL to ensure that the translated MRL matches with the translated question. This is done in three steps:

Step 1: First, we reformat the question-MRL pair in English by replacing the text tokens in the MRL with placeholder tokens $x_0, x_1, ...$ that correspond to text tokens in the question. We also mark the positions of placeholder tokens in the question. Table 2 gives an example.

Step 2: We then use the Amazon Machine Translation Service[1] to translate the natural language question into the target language. Next, we use the fast align algorithm (Dyer et al., 2013) to align the text between the translation and the original English sentence so as to identify the positions of the placeholder tokens in the translation. Figure 1 illustrates the alignment of texts between the source English sentence and its Italian translation and the identified placeholder tokens in the translation.

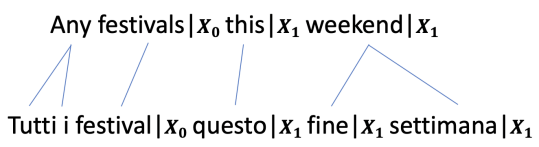

Figure 1: Using fast align algorithm to identify corresponding placeholder tokens in the translation.

Step 3: Finally, to obtain a valid MRL in the target language, we substitute the placeholder tokens in the MRL back with their corresponding text tokens in the translation. In this way, a valid pair of question and its MRL annotation in the target language is created (see Table 3).

Question (Italian):
Tutti i festival questo fine settimana

MRL:
[IN:GET_EVENT
[SL:CATEGORY_EVENT festival]
[SL:DATE_TIME questo fine settimana]]

Table 3: English semantic parsing data translated into Italian

Following this method, we generate training data for Italian and Japanese semantic parsing from the English TOP dataset. We machine translated the training and validation splits of the TOP dataset into the two target languages.

In order to evaluate the performance of our multilingual models on human-written sentences rather than machine-translated ones, we hire professional translators to manually translate the test set into Italian and Japanese. Table 4 shows the data distribution of the multilingual TOP dataset. It should be noted that as the fast align algorithm may fail to align the tokens between the translation and the source text, especially when the source and target languages are dissimilar, we may lose some data points in the automatic multilingual data generation process. Overall, the vast majority of the training data can be bootstrapped successfully following our method (97.9% data for Italian and 89.9% for Japanese).

Language	Train	Dev	Test
English	28414	4032	8241
Italian	27830	3955	8241
Japanese	25544	3629	8241

Table 4: The distribution of the multilingual TOP dataset

4 Multilingual Semantic Parsing Models

4.1 Model Architecture

Following the work of the state-of-the-art semantic parsers in English (Dong and Lapata, 2018; Rongali et al., 2020), we base our multilingual semantic parsing model on the sequence-to-sequence

[1] https://aws.amazon.com/translate/

method. We train a model that is similar in architecture to the Transformer encoder-decoder model described in Vaswani et al. (Vaswani et al., 2017). More specifically, we use a multilayer bidirectional Transformer encoder to encode the input question and a Transformer decoder to predict the MRL as a sequence of tokens. An encoder-decoder attention layer in the decoder learns to attend to the input tokens. We also implement an attention-based pointer mechanism (Vinyals et al., 2015) to learn to copy text tokens from the input question. We concatenate the attention scores from the attention layer with the output vocabulary distribution from the final layer of the decoder. We then feed the concatenated vector to a Softmax layer to obtain a final probability distribution of possible actions. At each time step, the decoder either generates a symbol from the output vocabulary or outputs a pointer to one of the input tokens based on the scores from the final probability distribution. We use beam-search at inference time to select the prediction that maximizes the probability of the entire sequence. To train our baseline multilingual semantic parsing model, we mix the data from all languages together and train a single model from scratch to parse all questions. We apply Byte-Pair Encoding (BPE) tokenization (Sennrich et al., 2016) to preprocess the data. BPE tokenization learns to break rare words into subword units. It is frequently used in machine translation and has contributed to better translation quality in many shared tasks (Denkowski and Neubig, 2017). For multilingual tasks, we believe that subword representation helps to encode shared information between similar languages, and therefore facilitates multilingual semantic parsing.

4.2 Multilingual Semantic Parsing with Pretrained Encoders

Transfer learning is a technique that aims to transfer information from a model trained on a source task to improve performance of the model on a target task. For neural network models, transfer learning typically consists of two stages: a pretraining stage and a finetuning stage. In the pretraining stage, the model is trained on the source task. In the finetuning stage, the knowledge of the trained model is transferred to the target task and adapted on that task. Existing literature has shown that transferring knowledge from pretrained models can improve the downstream performance on many NLP tasks (Devlin et al., 2019).

As all our non-English semantic parsing training data are automatically generated from machine translation, it may not be as natural as real human-written sentences. We believe that transferring knowledge from a model that is pretrained on a huge amount of authentic multilingual text will allow our multilingual semantic parser to learn a better representation for the input utterance and to generalize better on real human-written sentences. To do that, we first initialize the encoder parameters with pretrained encoder parameters. We compare two state-of-the art multilingual encoders for initializing the multilingual semantic parser: the multilingual BERT (mBERT) model (Vaswani et al., 2017) and the XLM-R model (Conneau and Lample, 2019). Both models cover all the languages required in our semantic parsing tasks. The mBERT model is based on the multi-layer Transformer architecture. It is trained using the masked language objective and the next sentence prediction objective (Devlin et al., 2019) on Wikipedia texts for the top 100 languages with the largest Wikipedia dumps. In our experiment, we use the public multilingual cased BERT-Base model[2] (12-layer, 768-hidden, 12 heads) to initialize our semantic parsing encoder. The XLM-R model is a Transformer model trained using multilingual masked language model objectives (Conneau and Lample, 2019). It is trained for 100 languages on the CommonCrawl corpus, which is several orders of magnitude larger than the Wikipedia dump, especially for low-resourced languages. We use the public XLM-R Base model[3] (12-layer, 768-hidden, 12 heads) in our experiment.

After initializing the semantic parsing model with pretrained encoder parameters, we finetune the models on the mixed multilingual semantic parsing data. To effectively adapt the pretrained encoder to our data, we implement gradual unfreezing (Howard and Ruder, 2018) in the finetuning steps. Instead of tuning all encoder layers from the beginning, which may cause the model to forget what it learnt in pretraining, we slowly unfreeze the encoder layer weights to be tuned, from not changing the weights at all in the beginning until we finetune all the layers.

[2] https://github.com/google-research/bert/blob/master/multilingual.md

[3] https://github.com/pytorch/fairseq/tree/master/examples/xlmr

4.3 Baselines

We compare our multilingual semantic parsing models against two groups of baselines: monolingual models trained for each target language and a common method in previous research that also makes use of machine translation.

4.3.1 Monolingual Baselines

We investigate how our multilingual semantic parsing models compare to monolingual models trained on each language separately. In accordance with the multilingual models, we build two types of monolingual baselines: monolingual models without pretraining and monolingual models finetuned from pretrained encoders. We use the same model architecture as the multilingual models for the monolingual baselines. For the monolingual pretrained encoders, we use the public English RoBERTa (Liu et al., 2019) model to initialize the English model, because semantic parsers finetuned from the English RoBERTa model have achieved state-of-the-art result on the original TOP dataset (Rongali et al., 2020). As there is no public RoBERTa model available for Italian and Japanese, we use Italian and Japanese BERT models trained on Wikipedia data instead.

In addition to using monolingual pretrained encoders, we also investigate a baseline with multilingual pretrained encoders (mBERT and XLM-R) finetuned on monolingual data for each target language.

4.3.2 Multilingual Semantic Parsing through Machine Translation

An alternative to multilingual semantic parsing is to translate all non-English languages into English and use an English semantic parsing model to understand the translated utterances (Haas and Riezler, 2016). We compare our multilingual semantic parsing models against this method. We train an semantic parser on the English training data by finetuning from the RoBERTa model. We then use the Amazon Machine Translation Service to translate the Italian and Japanese sentences in the TOP test set into English. The translated texts are fed into the English semantic parser to get their MRL predictions. We use the MRL annotation of the English test set as the gold-standard for evaluation.

5 Experiments and Results

5.1 Experiment Setup

We measure the performance of the semantic parsing models by exact match accuracy. By its definition, an MRL prediction is considered accurate only if the entire predicted sequence is exactly the same as the gold-standard MRL. The models are trained on AWS P3 instances with Tesla V100 GPU. We use the Adam optimizer in training and introduce early stopping if the loss doesn't improve on the validation set. We tune the hyperparameters for each model by random search on the validation set and report the results on the test set.

5.2 Results and Analysis on the TOP Dataset

Table 5 shows the performance of the multilingual models on the TOP dataset and Table 6 shows the results of the baselines models. Comparing the multilingual models against the monolingual baselines, we find that training semantic parsing models on multilingual data jointly outperforms models trained on monolingual data only, even without using a pretrained encoder. The joint training is not only helpful for non-English languages, where the training data were machine translated, but it is also helpful for English, with or without a multilingual pretrained encoder.

In addition, we observe that transfer learning from pretrained encoders can improve the multilingual model performance further. Among the multilingual models, finetuning from pretrained XLM-R model achieves the best performance, which yields a parsing accuracy of 85.1% for English, 62.4% for Italian, and 36.3% for Japanese. It substantially outperforms the monolingual baselines as well as the method that relies on machine translating utterances into English and using the English semantic parser to understand the utterances. The results prove that bootstrapping training data from English using machine translation is an effective method for constructing training data for multilingual semantic parsing.

On the other hand, constrained by the method we created our training data, the semantic parsing accuracy is heavily dependent on the machine translation quality. The better the machine translation model is, the more similar the automatically generated multilingual training data can be to real data. We measured the BLEU scores of the machine translation models on a random sample of English-Italian and English-Japanese sentences and found

Languages	multilingual (no pretraining)	mBERT	XLM-R
English	79.1%	84.6%	**85.1%**
Italian	57.4%	61.4%	**62.4%**
Japanese	31.9%	34.2%	**36.3%**
Mixed	56.1%	60.1%	**61.2%**

Table 5: Results on the multilingual TOP dataset

Languages	monolingual baselines				machine translated to English
	no pretraining	monolingual BERTs	mBERT	XLM-R	
English	78.3%	85.3%	83.3%	83.8%	85.3% (English model)
Italian	55.9%	55.1%	59.8%	60.2%	35.3%
Japanese	28.0%	32.1%	33.0%	32.5%	15.1%

Table 6: Results from baseline models on the multilingual TOP dataset

that the BLEU scores are 57.5 for Italian and 27.2 for Japanese, which shows that the English-Italian machine translation model is substantially more accurate than the English-Japanese one. Therefore, we observe a big difference between the semantic parsing accuracy for Italian and for Japanese.

During error analysis, we find that a large group of errors in Italian semantic parsing is due to the inclusion or exclusion of articles copied in the MRL, which has minimal influence over the understanding. Table 7 gives an example. As a heuristic solution, we filter out articles from both the expectation and the prediction and the exact match accuracy rises from 62.4% to 75.4% by our best performing model. Similarly, a large group of errors in Japanese is due to the inclusion or omission of postpositions and grammatical particles in the MRL when they are copied from the input question. If we filter out the postpositions and grammatical particles from the gold-standard and the predicted MRLs, the exact match accuracy is raised from 36.3% to 52.3%.

Question (Italian):
`dove posso vedere i fuochi d'artificio questa sera`

Gold-standard MRL:
`[IN:GET_EVENT [SL:CATEGORY_EVENT i fuochi d'artificio ] [SL:DATE_TIME questa sera ]]`

Predicted MRL:
`[IN:GET_EVENT [SL:CATEGORY_EVENT fuochi artificio ] [SL:DATE_TIME questa sera ]]`

Table 7: An example of missing article "i" in Italian semantic parsing

5.3 Experiment on the NLMaps Dataset

Apart from experimenting on the machine translated training data, we also want to see how our multilingual models perform with training data created by human and how our models compare to existing work. Therefore, we report the results of our models on the multilingual NLMaps dataset. The multilingual NLMaps dataset (Haas and Riezler, 2016) is one of the largest multilingual semantic parsing dataset published in previous literature. It contains around 2,400 English utterances and their manual translation into German. The queries are paired with a MRL representation that can be executed on a geographic database. Because NLMaps doesn't have a validation set, we randomly split 10% of the training data as the validation set and trained our models on the remaining 90% of the data. The resulting dataset contains 1,350 training utterance-MRL pairs, 150 validation pairs and 880 test pairs for both English and German.

Table 8 shows the results. Our best performing model on the NLMaps dataset is the multilingual semantic parser finetuned from the mBERT model, which yields an accuracy of 79.7% for English and 79.5% for German. The best result reported on the multilingual NLMaps dataset in literature was by Duong et al. (2017). However, their model was trained on the full training dataset for 10k iterations without splitting a separate validation set. Therefore, we retrain our best performing model under the same condition and present the result in Table

Languages	monolingual baselines				multilingual (no pretraining)	mBERT	XLM-R
	no pretraining	monolingual BERTs	mBERT	XLM-R			
English	73.5%	74.1%	75.7%	63.3%	72.1%	**79.7%**	74.3%
German	68.0%	70.3%	71.6%	59.5%	66.9%	**79.5%**	73.9%
Mixed	-	-	-	-	69.5%	**79.6%**	74.1%

Table 8: Results on the multilingual NLMaps dataset

Languages	Our best (mBERT)	Duong et al. (2017)
English	**85.9%**	85.7%
German	**85.5%**	82.3%

Table 9: Comparing the mBERT-based model with SOTA model on the NLMaps dataset (trained on the full training data for 10k iterations)

9. The result shows that our multilingual model outperforms the state-of-the-art model in German by 3.2% while keeping the same level of accuracy in English (with a slight improvement of 0.2%).

On the NLMaps dataset, we find that training the model on mixed multilingual data does not outperform a monolingual model if the models are trained without using a pretrained encoder. However, joint multilingual training is still helpful when a pretrained encoder is used. For example, when we finetune the mBERT model on English and German data separately, the resulting models yield an accuracy of 75.7% for English and 71.6% for German, which are markedly lower than the results from the mBERT model finetuned on mixed multilingual data. In addition, we find that using pretrained multilingual mBERT model outperforms pretrained monolingual BERT models.

5.4 Experiment on Zero-shot Learning

Encoder weights unfreezing rate		Italian	Japanese
unfreeze all	mBERT	24.9%	4.6%
	XLM-R	36.1%	1.7%
unfreeze 10%	mBERT	28.6%	**7.3%**
	XLM-R	**44.9%**	4.9%
freeze all	mBERT	16.0%	2.5%
	XLM-R	15.6%	3.9%

Table 10: Zero-shot learning on the TOP dataset

Zero-shot learning is a problem setup in which a model is tested on tasks that are not observed at training time. It studies the model's ability to generalize to unseen tasks. For multilingual models, we are interested in the zero-shot performance of a model when it is trained on one language and tested on other languages. To explore the zero-shot ability of our multilingual semantic parsers on the TOP dataset, we train a model with pretrained multilingual encoder on the English training data and apply the model to Italian and Japanese test data directly without further finetuning. We experiment with different ratios for unfreezing the pretrained encoder weights when tuning the models on the English data. Table 10 shows the results. We find that setting a small unfreezing rate to the pretrained encoder leads to a higher zero-shot accuracy.

Question (Italian):
`Concerti di Beyonce questo fine settimana`

Predicted MRL:
`[IN:GET_EVENT [SL:CATEGORY_EVENT Concerti ] [SL:NAME_EVENT Beyonce ] [SL:DATE_TIME questo fine settimana ] ]`

Table 11: An example of correct zero-shot prediction

Multilingual models trained only on English data can achieve 44.9% zero-shot accuracy when parsing Italian sentences, even though it has not seen any Italian semantic parsing data in training. Table 11 shows an example. However, their zero-shot performance on Japanese sentences is very poor. This is not surprising as English and Italian are more similar and they share a lot more BPE subword units than English and Japanese.

6 Conclusion

In this paper, we describe our method to build multilingual semantic parsing models when the multilingual data is limited. We introduce a new multilingual semantic parsing dataset in English, Italian and Japanese based on the public TOP dataset, with training and validation data automatically generated from English and 8k test data manually translated. The multilingual TOP test set is so far the largest dataset for multilingual semantic parsing, which will be useful for future research. By leveraging joint multilingual training and transfer learn-

ing from pretrained encoders, our semantic parsing models outperform several baselines on the TOP dataset and the state-of-the-art on the NLMaps dataset. We show that semantic parsing models with pretrained multilingual encoders can generalize from English to Italian with 44.9% zero-shot accuracy. However, we find that there is a gap between Italian and Japanese semantic parsing with our method. In future work, we plan to improve our models with both language-invariant and language-specific encodings and apply our method to more languages.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition Banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the atis task: The atis-3 corpus. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual abstract meaning representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27, Vancouver. Association for Computational Linguistics.

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.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–742, Melbourne, Australia. Association for Computational Linguistics.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2017. Multilingual semantic parsing and code-switching. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 379–389, Vancouver, Canada. Association for Computational Linguistics.

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

Kilian Evang and Johan Bos. 2016. Cross-lingual learning of an open-domain semantic parser. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 579–588, Osaka, Japan. The COLING 2016 Organizing Committee.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. *CoRR*.

Carolin Haas and Stefan Riezler. 2016. A corpus and semantic parser for multilingual natural language querying of OpenStreetMap. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 740–750, San Diego, California. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

Bevan Jones, Mark Johnson, and Sharon Goldwater. 2012. Semantic parsing with Bayesian tree transducers. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 488–496, Jeju Island, Korea. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. *CoRR*, abs/1702.03196.

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*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Raymond Hendy Susanto and Wei Lu. 2017. Neural architectures for multilingual semantic parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–44, Vancouver, Canada. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2692–2700. Curran Associates, Inc.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1332–1342, Beijing, China. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, page 1050–1055. AAAI Press.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, page 658–666, Arlington, Virginia, USA. AUAI Press.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*.