Multi-level Alignment Pretraining for Multi-lingual Semantic Parsing

Bo Shao[1,4], **Yeyun Gong**[2], **Weizhen Qi**[2,3], **Nan Duan**[2], **Xiaola Lin**[1]
[1]Sun Yat-sen University
[2]Microsoft Research Asia
[3]University of Science and Technology of China
[4]Microsoft STCA Bing
shaobo2@mail2.sysu.edu.cn, {yegong, nanduan}@microsoft.com,
weizhen@mail.ustc.edu.cn, linxl@mail.sysu.edu.cn

Abstract

In this paper, we present a multi-level alignment pretraining method in a unified architecture for multi-lingual semantic parsing. In this architecture, we use an adversarial training method to align the space of different languages and use sentence level and word level parallel corpus as supervision information to align the semantic of different languages. Finally, we jointly train the multi-level alignment and semantic parsing tasks. We conduct experiments on a publicly available multi-lingual semantic parsing dataset ATIS and a newly constructed dataset. Experimental results show that our model outperforms state-of-the-art methods on both datasets.

1 Introduction

[1] The goal of semantic parsing is to convert a natural language sentence to an executable logical form, which has been studied in the past few years and used on various applications, such as question answering (Kwiatkowski et al., 2011), task-oriented dialog systems (Yih et al., 2015) and interpreting instructions (Artzi and Zettlemoyer, 2013).

Due to the importance of semantic parsing, various approaches have been proposed for this task, such as (Kwiatkowski et al., 2011; Jia and Liang, 2016; Dong and Lapata, 2018; Chen et al., 2018). However, most existing methods only handle monolingual semantic parsing, while in real world applications such as Chatbot and search engine, we generally need to handle multi-lingual semantic parsing. Table 1 shows an example of the multi-lingual semantic parsing task, and the task aims to convert the question from different languages into the corresponding lambda calculus. For multi-lingual semantic parsing, previous works such as Jie and Lu (2014) and Susanto and Lu (2017) study it from different perspectives. Jie and Lu (2014) train the model for each language respectively and use ensemble method to combine the models on a multi-lingual semantic parsing dataset. Susanto and Lu (2017) propose a hybrid combination method to model multi-source input. Both of them need enough multi-lingual semantic parsing data for training. However, it is very hard to collect enough multi-lingual semantic parsing data.

EN	Who is the director of Inception?
ZH	谁是电影Inception的导演
LF	$\lambda x.film\_film\_director(Inception, x)$

Table 1: An example of our multi-lingual semantic parsing dataset, including a lambda calculus (LF) with the English (EN) and Chinese (ZH) question.

Recently, various pretraining methods have been successfully used to solve the labeled data insufficient problem in different tasks. In these methods, unsupervised data (Peters et al., 2017; Alec Radford, 2018; Devlin et al., 2018) or richly supervised data (McCann et al., 2017; Lample and Conneau, 2019) from other tasks are used to pretrain their models and achieve significant performance improvement in different tasks.

[1]This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 28th International Conference on Computational Linguistics, pages 3246–3256
Barcelona, Spain (Online), December 8-13, 2020

In this paper, we propose a multi-level alignment pretraining method to align the space level, word level and sentence level semantic representations for different languages. We design an adversarial training method to align the space level representation using unsupervised data. And to align the semantic level representation of parallel corpus in different languages, we use machine translation corpus and bilingual tokens to learn a shared cross-lingual encoder for our semantic parsing model. To better evaluate our method, we construct an open domain multi-lingual semantic parsing dataset, since most existing multi-lingual semantic parsing datasets (Hemphill et al., 1990; Zettlemoyer and Collins, 2012) are for specific domain and relatively small in scale.

The main contributions of this paper are:

- We design a multi-level alignment pretraining method to pretrain the multi-lingual semantic parsing model.
- We construct a new multi-lingual semantic parsing dataset on open domain, we will release this dataset to help the research of multi-lingual semantic parsing tasks.
- We conduct an experiment on ATIS and our dataset. Experimental results show that our model achieves new state-of-the-art results on both datasets.

2 Model

In this section, we will first briefly introduce the basic sequence-to-sequence (S2S) model as our baseline model. Then, we introduce the architecture of our Multi-level Alignment pretraining for multi-lingual Semantic Parsing (MASP) model.

2.1 S2S Model for Semantic Parsing

The S2S model has been successfully used in recent semantic parsing task (Dong and Lapata, 2016). The input of the model is a natural language question $q = [x_1, x_2 ... x_{|q|}]$ and output is a logical form sequence $l = [y_1, y_2, ... y_{|l|}]$. The tokens of the question q are fed one-by-one into the encoder, producing a sequence of encoder hidden states $h = [h_1, h_2, ... h_{|q|}]$. In the decoding process, at each time step t, the decoder computes the attention distribution to obtain a context vector c^t as follows:

$$\begin{aligned} e_i^t &= u^T f(W_e[h_i; s^t] + b_e); \\ a_i^t &= \frac{e_i^t}{\sum_{j=1}^{|q|} e_j^t}; c^t = \sum_{i=1}^{|q|} a_i^t h_i \end{aligned} \tag{1}$$

where f is a non-linear function, and we use $tanh$ here. u, W_e and b_e are parameters. s^t is the decoder hidden state at step t.

The context vector c^t is used to compute the generation distribution P_v of the target vocabulary with the hidden state s^t:

$$P_v = softmax(W_p(W[s^t; c^t] + b) + b_p) \tag{2}$$

where W_p, W, b, b_p are parameters.

In particular, to tackle out-of-vocabulary words, we incorporate the same copy mechanism as in (See et al., 2017) in our decoder. Attention score a_i is used as probability distribution of the copy mechanism over the source words. The copy distribution P_c is defined as follows:

$$P_c(y_t) = \sum_{i:x_i=y_t} a_i^t \tag{3}$$

To combine the copy distribution with the generation distribution, we use a gate g_c to choose whether to copy from q or generate from the target vocabulary:

$$g_c = \sigma(W^*[c^t; s^t; z^{t-1}] + b^*) \tag{4}$$

where vectors W^*, b^* are parameters. z^{t-1} is the word embedding of the previous word. We get final distribution score on each step t:

$$P_f(y_t) = (1 - g_c)P_v(y_t) + g_c P_c(y_t) \tag{5}$$

where $P_f(y_t)$ is considered as the final vocabulary distribution for step t.

We compute the overall loss of all steps as:

$$\mathcal{L}_{s2s} = \frac{\sum_{i=0}^{|l|} -logP_f(y_i)}{|l|} \tag{6}$$

2.2 MASP

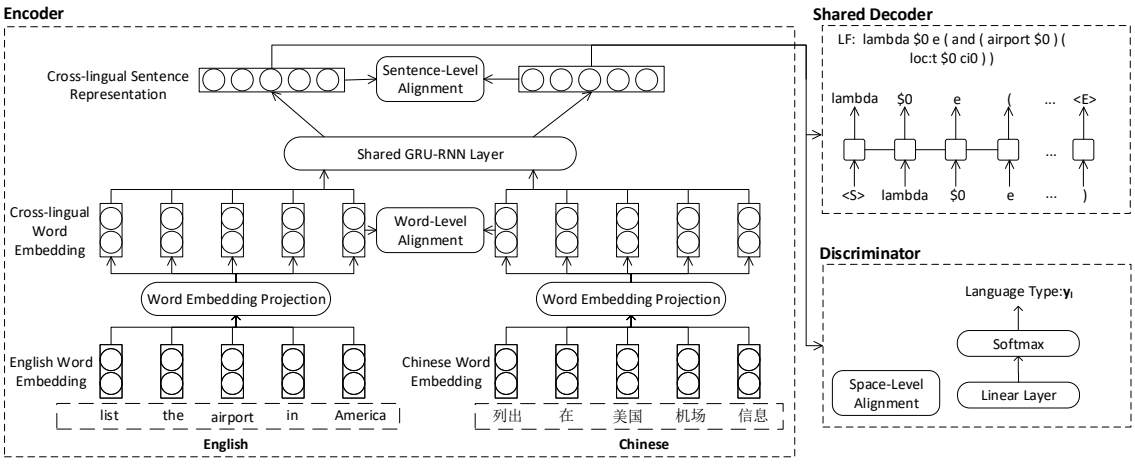

Figure 1: Overview of Multi-level Alignment semantic parsing Model.

In this section, we will introduce our Multi-level Alignment pretraining for multi-lingual Semantic Parsing (MASP) model. Our model uses pretraining method to incorporate rich unsupervised and supervised corpus to align sentences in different languages into shared multi-lingual space. And then we apply our model to multi-lingual semantic parsing task.

2.2.1 Multi-level Alignment

In this section, we will introduce our alignment strategies. Our model integrates three alignment strategies in space, word and sentence level, to learn shared semantic information during training. The space level alignment only uses monolingual corpus, word-level alignment needs bilingual dictionary, and sentence level alignment learns shared semantic information from parallel corpus. The input of the multi-lingual model is pair-wised, including two sentences, $Q_E = \{x_1, x_2, ..., x_n\}$ in English, $Q_C = \{c_1, c_2, ..., c_m\}$ in Chinese, where n and m is the length of Q_E and Q_C.

Space-level Alignment In this section, we design an adversarial learning method to maximize the confusion between two language representations, which has been successfully used in domain adaptation (Tzeng et al., 2017). The distributions of their representations are quite different, which will harm the performance of shared semantic parsing model. To align the distribution space of the two languages, we use the adversarial learning method to maximize the confusion between the two languages, which aligns the distribution of the sentence representations. And questions from two languages can be considered as two special domains.

The discriminator D is aimed to distinguish whether the input representation is from English or Chinese. In our model, the discriminator D is a binary classifier with a standard softmax layer. The input of D is all hidden states, i.e h^e of Q_E, from the shared RNN encoder. Furthermore, we give an extra label $y_l \in \{0, 1\}$ for discriminator D to indicate which language the input of discriminator belongs to.

The discriminator D sums up all hidden states as input features, and predicts which language the encoded sentence belongs to. For the English question Q_E, the final distribution in discriminator can be formulated as,

$$P_{ad} = softmax(M^a(\sum_{i=1}^{n} h_i^e) + b^a) \tag{7}$$

where M^a, b^a are trainable parameters, P_{ad} is the probability distribution of labels that indicate the language type. The final distribution Q_C is the same as Q_E. Then we compute the cross entropy loss $\mathcal{L}_{ad}$ of the discriminator D:

$$\mathcal{L}_{ad} = -y_l log(P_{ad}(y_l)) \tag{8}$$

For our multi-lingual model, we maximize the reversal classification loss to optimize the parameters, which aims to confuse the discriminator, and the reversal loss $\mathcal{L}_g$ is formulated as follows,

$$\mathcal{L}_g = -y_l log(P_{ad}(1 - y_l)) \tag{9}$$

This strategy can align the sentence representation space of different languages to help our model learn shared semantic information.

Word-Level Alignment Space-level alignment strategy can align the distribution space of the two languages. However, the shared semantic information is not aligned. In this section, we will introduce our word level alignment strategy to map monolingual word embedding into shared cross-lingual semantic space with the dictionary of bilingual lexicons. The model is first initialized with a pretrained word embedding matrix, trained by word2vec based methods (Mikolov et al., 2013; Bojanowski et al., 2017) in the two different languages. Here we define the two word embedding matrices, $X_E = R^{|X_E|*d}$ in English and $X_C = R^{|X_C|*d}$ in Chinese, d is the dimension of word embedding. The word embedding matrix in each language is pretrained respectively, embeddings of words that have the same meanings are unaligned, which will increase difficulty to encode sentences in our model. Thus, in order to map X_E and X_C into a shared semantic space, we define two linear transformation matrices W^E and W^C using as the multi-lingual projection. The matrices apply a linear transformation on X_E and X_C to align their embedding in each dimension. Then, we optimize our model by monolingual corpus with an extra bilingual lexicon dictionary B.

Formally, we compute the multi-lingual representations and add the word alignment loss as,

$$\mathcal{L}_w = \sum_{i=0}^{n}\sum_{j=0}^{m} 1 - cos(W^E X_E(x_i), W^C X_C(c_j)) \quad [(x_i, c_j) \in B] \tag{10}$$

where $cos(.,.)$ is the function that computes the cosine distance of two vectors. Through Eq.10, we align the word embedding in the pretraining process to help the performance of the encoder, which share the parameters between the two languages.

Sentence-Level Alignment Word level alignment is used to align monolingual embedding, but it can not cover more complex semantic information between different languages, since semantic equivalent sentences in the two languages are different in structures. To align the two representations for semantic parsing, end-to-end training a semantic parser requires a large amount of multi-lingual semantic parsing data, which is costly to annotate. However, there are sufficient machine translation data that contains the semantic alignment information across different languages. Thus, we pretrain our model on these data to align the representation of semantic equivelant sentences .

In our model, for each sentence pair Q_E and Q_C from the two languages, the shared BiLSTM encoder compute its contextual representations $[h_1^e, h_2^e, ..., h_n^e]$ of Q_E and $[h_1^c, h_2^c, ..., h_m^c]$ of Q_C respectively. We use the final hidden h_n^e and h_m^c state as the sentence representation. Sentence pairs in the equivalent semantics should have similiar sentence representations that will be used in the decoder to generate the same logical form. We also construct some negative sentence pair with different meaning by randomly

sampling sentences in both languages. It can be considered as an auxiliary task that predict whether the sentences' pair from different languages has the same meaning. We also randomly pair questions in the two language as the negative samples, and we use extra label $y_s \in \{0, 1\}$ to indicate whether the sentence pair is semantic equivalent.

Then we can compute the sentence alignment loss as,

$$\mathcal{L}_s = y_s(1 - cos(h_n^e, h_m^c)) + (1 - y_s)(1 + cos(h_n^e, h_m^c)) \tag{11}$$

where $\mathcal{L}_s$ is the loss of sentence-level alignment.

2.2.2 Training Process

The full training process contains two steps, firstly, pretraining our model with the three alignment strategies, then jointly training the model on multi-lingual semantic parsing datasets with pretraining corpus.

Pretraining We pretrain our multi-lingual model with machine translation parallel corpus(for the experiment without sentence alignment, we use monolingual corpus instead) and bilingual dictionary. The pretraining model contains the multi-lingual encoder and the discriminator D. We feed each pair of sentences in the two languages in our model with the bilingual lexicon vocabulary B. The overall loss of multi-lingual model contains word-level alignment loss, sentence-level alignment loss sentence and language confusion loss,

$$\mathcal{L}_{pre} = \mathcal{L}_w + \mathcal{L}_s + \mathcal{L}_g \tag{12}$$

Simultaneously, we alternately optimize the discriminator D with the loss $\mathcal{L}_{ad}$ until both losses converges. Then the pretrained model will be saved for multi-lingual semantic parsing tasks.

Joint Training We initialize the parameters with the pretrained multi-lingual model, and then finetune on the corresponding semantic parsing dataset. The input sample in a multi-lingual semantic parsing dataset, contains questions in different languages and corresponding logical forms. And we train our model with these samples by Eq. 6. In order to keep the alignment property, we use a joint training method with the pretraining corpus and optimize the model with the loss $\mathcal{L}_{ft}$, which contains the generation loss $\mathcal{L}_{s2s}$ and alignment loss $\mathcal{L}_{pre}$:

$$\mathcal{L}_{ft} = \mathcal{L}_{s2s} + \alpha\mathcal{L}_{pre} \tag{13}$$

where α is used to control the weight of the alignment loss.

	ATIS		**MLSP**	
Method	EN	ZH	EN	ZH
Translate Test	20.31	33.04	21.98	19.47
SL-SINGLE (Susanto and Lu, 2017)	81.85	73.66	-	-
SL-SHARED (Susanto and Lu, 2017)	81.77	73.96	-	-
SL-SEPARATE (Susanto and Lu, 2017)	81.40	75.89	-	-
SS-SINGLE	82.37	74.55	68.32	62.74
SS-SHARED	82.14	75.45	70.44	67.88
MASP	**85.04**	**79.02**	**76.39**	**73.83**
w/o SPACE	83.48	78.13	75.45	72.70
w/o WORD	83.26	78.57	73.32	72.45
w/o SENT	84.15	77.68	74.08	70.57

Table 2: Accuracy on ATIS and MLSP datasets. "EN" represents the accuracy of English, and "ZH" represents the result of Chinese. In our methods, "w/o" means to ablate each alignment strategy respectively.

3 Dataset

3.1 Dataset Construction

In this section, we introduce a new **m**ulti-**l**ingual **s**emantic **p**arsing (MLSP) dataset based on Satori [2] It contains a set of nodes and edges that are represented by triple $\{s, p, o\}$. Each triple denotes two nodes,

[2]https://microsoft.sharepoint.com/teams/SatoriHelpAndSupport/SitePages/Home.aspx

a subject entity s, an object entity o and the directed edge p between them as a predicate. We collect our dataset by crowd sourcing, which involves two steps:

1) First, we collect the connected triples in Freebase randomly. Second, we annotate a simple question for each triple as seed questions. Third, we automatically generate complex questions for the connected triples with the simple questions of selected triples using a template, following the procedure from ComplexWebQuestion (CWQ) (Talmor and Berant, 2018). Fourth, we ask native speakers to paraphrase the questions generated from the template. Fifth, three other annotators verify the quality of the paraphrased results, and annotate three additional labels to indicate whether the paraphrased questions are the semantic equivalents of the automatically generated questions. We obtained a two-vote consensus of 97% and dropped the 3% additional samples.

2) To generate Chinese questions for our dataset, we first use Microsoft's translator[3] to translate English questions into Chinese. Then we ask annotators to translate the English questions into Chinese given the machine translated questions as a reference. For the questions which are difficult to translate, we label them as "None" and drop them from our experiment. After this step, about 92% of the questions are retained.

3.2 Dataset Analysis

In total, MLSP contains 15,991 samples. Each sample in our dataset has an English question, a Chinese question and a corresponding lambda calculus, which contains primary functions such as *Argmax, Argmin,Argmore, Argless, Max, Min* defined to denote basic functions. We also calculate the number of question patterns and logical from patterns, whose entity name are replaced with a placeholder, and our dataset contains 7,482 qustion patterns and 3,429 logical form patterns. Compared with existing datasets GEO and ATIS, which contain 880 and 5,410 samples respectively, MLSP is a large scale dataset in open domain. We will release this dataset to advance research in multi-lingual semantic parsing.

To evaluate the quality of the dateset, we randomly select 1% annotated samples to double check, and we find that 95% of these samples are correct. We will publish this dataset with more detailed instructions.

4 Experiment

We conduct our experiments on two datasets, ATIS and MLSP.

4.1 Datasets

ATIS contains 5410 queries from a flight booking system (Hemphill et al., 1990). The data samples have been split into 4348 training instances, 491 validation instances, and 448 test instances. Each pair contains a question and the lambda-calculus expression with the identified values for the variables of date, time, city, aircraft code, airport, airline, and number. The corpus was translated into Chinese with segmentation from (Susanto and Lu, 2017).

For our MLSP dataset introduced in Section 3, we randomly split the data into 0.8/0.1/0.1 as train/dev/test sets in our model.

In pretraining, we use the English-Chinese translation corpus, News Commentary v12 of WMT 2017 (Bojar et al., 2017). The English corpus is tokenized by NLTK (Bird and Loper, 2004) and the Chinese corpus is tokenized by Jieba segmenter[4]. In space-level and word-level alignment, we use the unsupervised corpus of Wikipedia[5]. We also randomly sample the same number of sentence pairs as the MT dataset used as the negative samples in sentence level alignment experiment. In experiment of word level alignment, we also construct a simple bilingual lexicon dictionary by translating the words contained in the English version into Chinese. We randomly collect 1k word pairs as bilingual lexicons. If the word pair appears in the sentence pair, we will mark their positions with labels for word-level alignment experiment. For ATIS, the pre-processing is the same as (Dong and Lapata, 2016), which

[3]https://www.bing.com/translator

[4]https://github.com/fxsjy/jieba

[5]https://dumps.wikimedia.org/

replace entities with their type name. To evaluate our method in situations when there is not an annotated multi-lingual semantic parsing dataset, we translate the English semantic parsing corpus by the open translation service of Microsoft. This is expected to be a common scenario in practice.

4.2 Settings

We set the vocabulary size to 50k for both languages in our model. We use Glove (Pennington et al., 2014) 6B and Fasttext pretrained Chinese (Bojanowski et al., 2017) as English and Chinese pretrained word embedding. For words in vocabulary which do not have pretrained embeddings, we assign them uniform randomized values. The size of the word embedding is set to 300. During training, we update all word embeddings. We use accuracy on the development set to implement early stopping. Parameters are randomly initialized from a uniform distribution (-0.01, 0.01). For regularization, we use dropout and set the dropout rate to 0.5. Dimensions of hidden vectors in encoder and decoder are 300. α in joint training is set to be 0.1. Adagrad (Duchi et al., 2011) is used in training with an initial accumulator value of 0.1.

4.3 Results

Table 2 shows the results of our model and the state-fo-the-art methods on multi-lingual ATIS, and MLSP datasets, we report accuracy of exact match to evaluate our model. "SL-SINGLE" represents applying SEQ2TREE (Dong and Lapata, 2016) to each language respectively, "SL-SHARED/SEPARATE" denotes training the model with shared/separate encoder in (Susanto and Lu, 2017). "SS-SINGLE" represents training seq2seq model (described in 2.1) for each language respectively. "SS-SHARED" denotes using both English and Chinese data to train the model. "MASP" is the proposed model in this paper. Specially, we report the baseline "Translated Test" which represents we translate questions of one language in the test dataset and evaluate on baseline model trained with data in the other language.

From the results, we observe that our model achieves a new state-of-the-art results on all dataset. Comparing "SS-SHARED" with "SS-SINGLE", we see that merging the data in different languages does not achieve promising improvement, this is because the Chinese and English are different in word and sentence level. Compared with "SS-SHARED", the proposed model "MASP" achieves significant improvement in both languages which illustrates the effectiveness of multi-level alignment method.

We conduct an ablation study on the variants of "MASP" and investigate the effect of our alignment strategy. The last four lines in Table 2 show the results by ablating each aligment strategy respectively. "w/o SENT" represents the model without sentence level alignment, "w/o SPACE" and "w/o WORD" denotes without space level alignment and word level alignment respectively. From the results, we observe that "MASP" outperforms "w/o SPACE", "w/o WORD" and "w/o SENT" on all the results, which illustrates that removing each alignment harms the performance of our model.

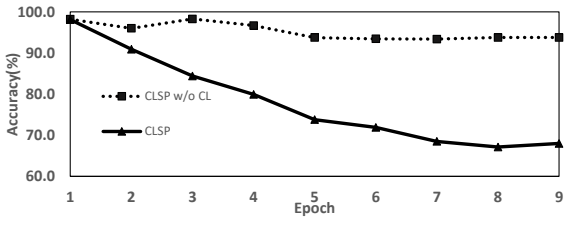

Figure 2: Performance of discriminator in pretraining.

4.4 Alignment Analysis

We analyze the space level, word level and sentence level alignments in this section.

4.4.1 Space-level Alignment

We evaluate the performance of discriminator in the space-level alignment. We use questions in machine translation dataset with their language label as input, and feed the encoded representation into the discriminator. Figure 2 shows the discriminator results during pretraining. "MASP w/o CL" represents the

model without confusion loss in space alignment. The results show that the discriminator achieves high accuracy of discriminating the representations in "MASP w/o CL", while it is hard to discriminate the representations in "MASP" after 5 epochs. The results illustrate our space alignment method successfully align representation distribution of different languages and confuses the discriminator, which helps our model to handle multi-lingual questions.

4.4.2 Word-level Alignment

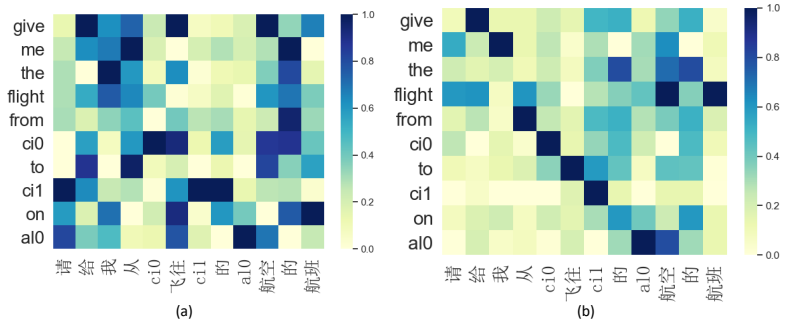

Figure 3: Word Alignment result. (a) represents the result without word alignment, and (b) is the alignment result

To evaluate word alignment result, we compute the cosine distance between the tokens of the questions in the two languages. Figure 3(a) shows the results without word level alignment and Figure 3(b) shows the results using word level alignment. From the results, we observe that most words from different languages with the same meaning have been aligned by our word alignment method. For example, "fight" and "航班" are semantic equivalent in English and Chinese, and their embedding are closed in cosine similarity but it is not show the same trend in the model without our alignment strategy. It shows that our word alignment layer can successfully transform the word embedding into a shared multi-lingual space, thus help to improve the model performance.

4.4.3 Sentence-Level Alignment

We also evaluate our sentence alignment results by a classification task as auxiliary. We use the pre-trained model to encode the sentence pair in different languages, and use the cosine distance of sentence representations to predict whether the two sentences have the same meaning. We assume that if the cosine distance is greater than 0, the two sentences are semantic equivalent. We evaluate our model on multi-lingual ATIS datasets. We use their question pair data as positive samples and randomly select the same number of Chinese questions and English questions, pair as negative samples to compute the classification accuracy. Then we find that the accuracy is up to 97% in both datasets by our sentence-level alignment method. However, without sentence-level alignment the accuracy is 61.3%. This experiment shows our sentence-level alignment method successfully aligns sentence representations in semantic.

4.5 Results on Translated Corpus

	ATIS		MLSP	
Method	**EN**	**ZHMT**	**EN**	**ZHMT**
SS-SINGLE	82.37	25.00	68.32	23.67
SS-SHARED	79.24	36.61	69.63	25.74
MASP	**83.48**	**43.08**	**73.64**	**34.56**
w/o SPACE	81.03	42.19	72.89	33.63
w/o WORD	81.92	41.74	71.51	31.50
w/o SENT	82.14	39.73	72.07	28.62

Table 3: Accuracy on ATIS and MLSP with machine translated Chinese Questions

Usually in real world scenarios, we only have monolingual semantic parsing dataset instead of multi-lingual dataset. In this section, we use Microsoft Translator to generate Chinese semantic parsing corpus

from the English corpus and use these data to evaluate our model. Table 3 shows the results of our model and baseline methods. From the results in Table 3 and Table 2, we find that the performance using translated corpus on both English and Chinese are lower than using annotated data. However, with these translated corpus, our methods can improve the target language performance, which shows its robustness. The results demonstrate that both the multi-lingual data on ATIS and MLSP effectively improve the semantic parsing performance. And also we see that our model achieves state-of-the-art results on all the results, which shows the effectiveness and robustness of our method. This experiment illustrates that our methods can be used in real world scenarios with the help of existing machine translator.

5 Related Work

Semantic parsing, as an important task in natural language understanding, has attracted significant attention in the research and industry. Recently, various semantic parsing models have been proposed such as (Kwiatkowski et al., 2011; Xiao et al., 2017; Yin and Neubig, 2017; Fan et al., 2017; Dong and Lapata, 2016; Chen et al., 2018; Dong and Lapata, 2018). Kwiatkowski et al. (2011) propose a combinatory categorical grammar induction technique for semantic parsing. Xiao et al. (2017; Yin and Neubig (2017) use grammar and syntax information to improve semantic parsing models. (Fan et al., 2017) apply a transfer learning method to semantic parsing. Dong and Lapata (2016) propose a tree-based decoder to model structure of logical forms. Chen et al. (2018) translate the decode process as a sequence of actions with a sequence-to-sequence model. Recently, Dong and Lapata (2018) propose a two-stage model to decode the logical form with the help of sketches, which contain structure and predicates in logical forms.

In multi-lingual semantic parsing, Jie and Lu (2014) use majority voting ensemble method to combine outputs from parsers for certain languages to apply on multi-lingual semantic parsing. Zhang et al. (2018) use a sequence-to-sequence model to map the questions in the source language into decompositional semantic representations in the target language. In Susanto and Lu (2017)'s work, they propose a combination method to combine questions in different language simultaneously for multi-source input and achieve promising improvement on ATIS (Hemphill et al., 1990). They also explore different architectures for single-source input without their combination mechanism. Zou and Lu (2018) propose a method to learn a cross lingual representation and use it in their semantic parsing model (Zettlemoyer and Collins, 2012).

6 Conclusion

In this paper, we propose a multi-lingual semantic parsing model, which is first pretrained using a multi-level alignment mechanism, and then we jointly train the multi-lingual semantic parsing and multi-level alignment tasks. Most existing multi-lingual semantic parsing datasets are based on specific domain, to better evaluate our method on open domain, we annotate a relative large scale multi-lingual semantic parsing dataset on open domain. Experimental results on ATIS and our dataset show the effectiveness and robustness of our model.

References

Tim Salimans andIlya Sutskever Alec Radford, Karthik Narasimhan. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association of Computational Linguistics*, 1:49–62.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Bo Chen, Le Sun, and Xianpei Han. 2018. Sequence-to-action: End-to-end semantic graph generation for semantic parsing. In *Proceedings of ACL*, pages 766–777. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of ACL*, pages 33–43. Association for Computational Linguistics.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. *arXiv preprint arXiv:1805.04793*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Xing Fan, Emilio Monti, Lambert Mathias, and Markus Dreyer. 2017. Transfer learning for neural semantic parsing. *arXiv preprint arXiv:1706.04326*.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of ACL*, pages 12–22. Association for Computational Linguistics.

Zhanming Jie and Wei Lu. 2014. Multilingual semantic parsing: Parsing multiple languages into semantic representations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1291–1301.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2011. Lexical generalization in ccg grammar induction for semantic parsing. In *Proceedings of EMNLP*, pages 1512–1523.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

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)*, volume 2, pages 38–44.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176.

Chunyang Xiao, Marc Dymetman, and Claire Gardent. 2017. Sequence-based structured prediction for semantic parsing, November 28. US Patent 9,830,315.

Scott Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base.

Pengcheng Yin and Graham Neubig. 2017. A syntactic neural model for general-purpose code generation. *arXiv preprint arXiv:1704.01696*.

Luke S Zettlemoyer and Michael Collins. 2012. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*.

Sheng Zhang, Xutai Ma, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2018. Cross-lingual decompositional semantic parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1664–1675.

Yanyan Zou and Wei Lu. 2018. Learning cross-lingual distributed logical representations for semantic parsing. *arXiv preprint arXiv:1806.05461*.