Zero-Shot Cross-lingual Semantic Parsing

Tom Sherborne and **Mirella Lapata**
School of Informatics, University of Edinburgh
tom.sherborne@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Recent work in crosslingual semantic parsing has successfully applied machine translation to localize accurate parsing to new languages. However, these advances assume access to high-quality machine translation systems, and tools such as word aligners, for all test languages. We remove these assumptions and study cross-lingual semantic parsing as a zero-shot problem without parallel data for 7 test languages (DE, ZH, FR, ES, PT, HI, TR). We propose a multi-task encoder-decoder model to transfer parsing knowledge to additional languages using only English-Logical form paired data and unlabeled, monolingual utterances in each test language. We train an encoder to generate language-agnostic representations jointly optimized for generating logical forms or utterance reconstruction and against language discriminability. Our system frames zero-shot parsing as a latent-space alignment problem and finds that pre-trained models can be improved to generate logical forms with minimal cross-lingual transfer penalty. Experimental results on Overnight and a new executable version of MultiATIS++ find that our zero-shot approach performs above back-translation baselines and, in some cases, approaches the supervised upper bound.

1 Introduction

Executable semantic parsing translates a natural language *utterance* to a *logical form* for execution in some *knowledge base* to return a *denotation*. The parsing task realizes an utterance as a semantically-identical, but machine-interpretable, expression *grounded* in a denotation. The transduction between natural and formal languages has allowed semantic parsers to become critical infrastructure in the pipeline of human-computer interfaces for question answering from structured data (Berant et al., 2013; Liang, 2016; Kollar et al., 2018).

Sequence-to-sequence approaches have proven capable in producing high quality parsers (Jia and Liang, 2016; Dong and Lapata, 2016) with further modeling advances in multi-stage decoding (Dong and Lapata, 2018; Guo et al., 2019), schema linking (Shaw et al., 2019; Wang et al., 2019) and grammar based decoding (Yin and Neubig, 2017; Lin et al., 2019). In addition to modeling developments, recent work has also expanded to multi-lingual parsing. However, this has primarily required that parallel training data is either available (Jie and Lu, 2014), or requires professional translation to generate (Susanto and Lu, 2017a). This creates an entry barrier to localizing a semantic parser which may not be necessary. Sherborne et al. (2020) and Moradshahi et al. (2020) explore the utility of machine translation, as a cheap alternative to human translation, for training data but found translation artifacts as a performance limiting factor. Zhu et al. (2020) and Li et al. (2021) both examine zero-shot spoken language understanding and observed a significant performance penalty from cross-lingual transfer from English to lower resource languages.

Cross-lingual generative semantic parsing, as opposed to the slot-filling format, has been underexplored in the zero-shot case. This challenging task combines the outstanding difficulty of *structured prediction*, for accurate parsing, with a *latent space alignment* requirement, wherein multiple languages should encode to an overlapping semantic representation. Prior work has identified this penalty from cross-lingual transfer (Artetxe et al., 2020; Zhu et al., 2020; Li et al., 2021) that is insufficiently overcome with pre-trained models alone. While there has been some success in machine-translation based approaches, we argue that inducing a shared multilingual space without parallel data is superior because (a) this nullifies the introduction of translation or word alignment errors and (b) this approach scales to low-resource languages without reliable machine translation.

In this work, we propose a method of **zero-shot executable semantic parsing** using only monolingual data for cross-lingual transfer of parsing knowledge. Our approach uses paired English-logical form data for the parsing task and adapts to additional languages using auxiliary tasks trained on unlabeled monolingual corpora. Our motivation is to accurately parse languages, for which paired training data is unseen, to examine if any translation is required for accurate parsing. The objective of this work is to parse utterances in some language, l, without observing paired training data, (x_l, y), suitable machine translation, word alignment or bilingual dictionaries between l and English. Using a multi-task objective, our system adapts pre-trained language models to generate logical forms from multiple languages with a minimized penalty for cross-lingual transfer from English to German (DE), Chinese (ZH), French (FR), Spanish (ES), Portuguese (PT), Hindi (HI) and Turkish (TR).

2 Related Work

Cross-lingual Modeling This area has recently gained increasing interest across a range of natural language understanding settings, with benchmarks such as XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020) providing classification and generation benchmarks across a range of languages (Zhao et al., 2020). There has been significant interest in cross-lingual approaches to dependency parsing (Tiedemann et al., 2014; Schuster et al., 2019), sentence simplification (Mallinson et al., 2020) and spoken-language understanding (SLU; He et al., 2013; Upadhyay et al., 2018). Within this, pre-training has shown to be widely beneficial for cross-lingual transfer using models such as multilingual BERT (Devlin et al., 2019) or XLM-Roberta (Conneau et al., 2020a).

Broadly, pre-trained models trained on massive corpora purportedly learn an overlapping cross-lingual latent space (Conneau et al., 2020b) but have also been identified as under-trained for some tasks (Li et al., 2021). A subset of cross-lingual modeling has focused on engineering alignment of multi-lingual word representations (Conneau et al., 2017; Artetxe and Schwenk, 2018; Hu et al., 2021) for tasks such as dependency parsing (Schuster et al., 2019; Xu and Koehn, 2021) and word alignment (Jalili Sabet et al., 2020; Dou and Neubig, 2021).

Semantic Parsing For parsing, there has been recent investigation into multilingual parsing using multiple language ensembles using hybrid generative trees (Lu, 2014; Susanto and Lu, 2017b) and LSTM-based sequence-to-sequence models (Susanto and Lu, 2017a). This work has largely affirmed the benefit of "high-resource" multi-language ensemble training (Jie and Lu, 2014). Code-switching in multilingual parsing has also been explored through mixed-language training datasets (Duong et al., 2017; Einolghozati et al., 2021). For adapting a parser to a new language, machine translation has been explored as mostly reasonable proxy for in-language data (Sherborne et al., 2020; Moradshahi et al., 2020). However, machine translation, in either direction, can introduce limiting artifacts (Artetxe et al., 2020) and generalisation is limited due to a "parallax" error between gold test utterances and "translationese" training data (Riley et al., 2020).

Zero-shot parsing has primarily focused on 'cross-domain' challenges to improve parsing across varying query structure and lexicons (Herzig and Berant, 2018; Givoli and Reichart, 2019) or different databases (Zhong et al., 2020; Suhr et al., 2020). The Spider dataset (Yu et al., 2018) formalises this challenge with unseen tables at test time for zero-shot generalisation. Combining zero-shot parsing with cross-lingual modeling has been examined for the UCCA formalism (Hershcovich et al., 2019) and for task-oriented parsing (see below). Generally, we find that cross-lingual executable parsing has been under-explored in the zero-shot case.

Executable parsing contrasts to more abstract meaning expressions (AMR, λ-calculus) or hybrid logical forms, such as decoupled TOP (Li et al., 2021), which cannot be executed to retrieve denotations. Datasets such as ATIS (Dahl et al., 1994) have both executable parsing and spoken language understanding format. We focus on only the former, as a generation task, and note that results for the latter classification task are not comparable.

Dialogue Modeling Usage of MT has also been extended to a task-oriented setting using the MTOP dataset (Li et al., 2021), employing a pointer-generator network (See et al., 2017) to generate dialogue-act style logical forms. This has been combined with adjacent SLU tasks, on MTOP and MultiATIS++ (Xu et al., 2020), to combine cross-lingual task-oriented parsing and SLU classification. Generally, these approaches additionally require word alignment to project annotations between languages. Prior zero-shot cross-lingual work in SLU (Li et al., 2021; Zhu et al., 2020; Krishnan et al., 2021) similarly identifies a penalty for cross-lingual transfer and finds that pre-trained models and machine translation can only partially mitigate this error.

Compared with similar exploration of cross-lingual parsing such as Xu et al. (2020) and Li et al. (2021), the zero-shot case is our primary focus. Our study assumes a case of no paired data in the test and our proposed approach is more similar to Mallinson et al. (2020) and Xu and Koehn (2021) in that we objectify the convergence of pre-trained representations for a downstream task. Our approach is also similar to work in zero-resource (Firat et al., 2016) or unsupervised machine translation with monolingual corpora (Lample et al., 2018). Contrastingly, our approach is not pairwise between languages owing to our single multilingual latent representation.

3 Problem Formulation

The primary challenge for cross-lingual parsing is learning parameters, that can parse an utterance, x, from any test language. Typically, a parser trained on language l, or multiple languages $\{l_1, l_2, \ldots, l_N\}$, is only capable for these languages and performs poorly outside this set. For a new language, conventional approaches requires parallel datasets and models (Jie and Lu, 2014; Haas and Riezler, 2016; Duong et al., 2017).

In our work, zero-shot parsing refers to parsing utterances in languages *without paired data during training*. For some language, l, there exists no pairing of x_l to a logical form, y, except for English. During training, the model only generates logical forms conditioned on English input data. Other languages are incorporated into the model using auxiliary objectives detailed in Section 4. Zero-shot parsing happens at test time, when a logical form is predicted conditioned upon an input question from any test language.

Our approach improves the zero-shot case using only monolingual objectives to parse additional languages beyond English[1] without semantic parsing training data. We explore a hypothesis that a multilingual latent space can be learned through auxiliary tasks in tandem with logical form generation. We desire to learn a *language-agnostic* representation space to minimize the penalty of cross-lingual transfer and improve parsing of languages without training data. To generate this latent space, we posit that only unpaired monolingual data in the target language, and some pre-trained encoder, are required. We remove the assumption that machine translation is suitable and study the inverse case wherein only paired data in English and monolingual data in target languages are available. This frames the cross-lingual parsing task as one of *latent representation alignment* only, to explore a possible upper bound of parsing accuracy without errors from external dependencies.

The benefit of our zero-shot method is that our approach requires only external corpora in the test language. Using only English paired data and monolingual corpora, we can generate logical forms above back-translation baselines and compete with fully supervised in-language training. For example, consider the German test of the Overnight dataset (Wang et al., 2015; Sherborne et al., 2020) which lacks German paired training data. To competitively parse this test set, our approach minimally requires only the original English training data and collection of unlabeled German utterances.

4 Method

We adopt a multi-task sequence-to-sequence model (Luong et al., 2016) combining logical form generation with two auxiliary objectives. The first is a monolingual reconstruction loss, similar to *domain-adaptive pre-training* (Gururangan et al., 2020), and the second is a language identification task. We describe each model component below:

Generating Logical Forms Predicting logical forms is the primary output objective for the system. For an utterance $x = (x_1, x_2, \ldots, x_T)$, we desire to predict a logical form $y = (y_1, y_2, \ldots, y_M)$ representing the same meaning in a machine-executable language. We model this transduction task using a sequence-to-sequence neural network (Sutskever et al., 2014) based upon the Transformer architecture (Vaswani et al., 2017).

The sequence x is encoded to a latent representation $z = (z_1, z_2, \ldots, z_T)$ through Equation (1) using a stacked self-attention Transformer encoder, E, with weights θ_E. The conditional probability of the output sequence y is expressed as Equation (2) as each token y_i is autoregressively generated based upon z and prior outputs, $y_{<i}$. The distribu-

[1] English acts as the “source” language in our work as the source language for all explored datasets. We express all other languages as “target” languages.

tion is modeled in Equation (3) using Transformer decoder, D_{LF} for logical forms with associated weights $\theta_{D_{LF}}$.

$$z = E\left(x \mid \theta_E\right) \tag{1}$$

$$p\left(y \mid x\right) = \prod_{i=0}^{M} p\left(y_i \mid y_{<i}, x\right) \tag{2}$$

$$p\left(y_i | y_{<i}, x\right) = \operatorname{softmax}\left(D_{\mathrm{LF}}\left(y_{<i} \mid z, \theta_{D_{\mathrm{LF}}}\right)\right) \tag{3}$$

For semantic parsing dataset $\mathcal{S}_{\mathrm{LF}} = \{x^n, y^n\}_{n=0}^{N}$, we generate an output prediction, $\hat{y}$, through the encoder and logical form decoder, $\{E,\ D_{LF}\}$. Equation (4) describes our minimization objective computed using cross-entropy loss between y and $\hat{y}$.

$$\mathcal{L}_{\mathrm{LF}} = -\sum_{(x,\ y) \in \mathcal{S}_{\mathrm{LF}}} \log p\left(y \mid x\right) \tag{4}$$

Reconstructing Utterances The secondary objective encourages multi-lingual semantic representations in encoding space, z, in Equation (1), using an additional decoder to recover a noisy input. This *co-adaptation* strategy steers a latent space suitable for both decoders, improving the parsing accuracy of languages without training logical forms. An utterance, x, is input to the encoder, E, and a reconstruction decoder, D_{NL}, then tries to recover x from the latent representation. We follow the *denoising* objective from Lewis et al. (2020) and replace x with noised input $\tilde{x} = \mathrm{N}\left(x\right)$ for some noising function N. The output probability of reconstruction, or denoising, is described in Equation (6) with each token predicted through Equation (7) using decoder, D_{NL}, with associated weights $\theta_{D_{\mathrm{NL}}}$.

$$z = E\left(\tilde{x} \mid \theta_E\right) \tag{5}$$

$$p\left(x \mid \tilde{x}\right) = \prod_{i=0}^{T} p\left(x_i \mid x_{<i}, \tilde{x}\right) \tag{6}$$

$$p\left(x_i | x_{<i}, \tilde{x}\right) = \operatorname{softmax}\left(D_{\mathrm{NL}}\left(x_{<i} \mid z, \theta_{D_{\mathrm{NL}}}\right)\right) \tag{7}$$

The reconstruction objective is trained using both the utterances from $\mathcal{S}_{\mathrm{LF}}$ and monolingual data, $\mathcal{S}_{\mathrm{NL}} = \{\{x^n\}_{n=0}^{N}\}_{l=0}^{L}$, in L languages. The sub-model, $\{E,\ D_{NL}\}$, predicts the reconstruction of x from $\tilde{x}$ with an optimization objective described in Equation (8). Our approach only considers monolingual reconstruction and we leave an additional translation objective, using bitext, to future work.

$$\mathcal{L}_{\mathrm{NL}} = -\sum_{x} \log p\left(x \mid \tilde{x}\right) \tag{8}$$

Language Prediction We augment the model with a third objective designed to encourage language-agnostic representations by reducing the discriminability of the source language, l, from z. Equation (9) defines a **L**anguage **P**rediction (LP) network to predict l from z using a feedforward classifier over L training languages. Here, $W_{di} \in \mathbb{R}^{|z| \times |z|}$ and $W_{do} \in \mathbb{R}^{L \times |z|}$ are layer weights, $b_{di} \in \mathbb{R}^{|z|}$ and $b_{do} \in \mathbb{R}^{L}$ are layer biases and G is the Gaussian Error Linear Unit (Hendrycks and Gimpel, 2020). Equation (10) describes the conditional model for the output distribution, as a language label is predicted using the time-average of the input encoding z of length T. Equation (11) describes the objective function for the LP network, however, we add a *gradient reversal layer* in the backward pass prior to the LP network to encourage the encoder to produce language invariant representations (Ganin et al., 2016). The LP network is optimized to discriminate the source language from z, but the encoder is now optimized *against* this objective[2]. Our intuition here is that discouraging language discriminability in z encourages latent representation similarity across languages, and therefore reduce the penalty for cross-lingual transfer.

$$\mathrm{LP}\left(x\right) = W_{do}\, G\left(W_{di} x + b_{di}\right) + b_{do} \tag{9}$$

$$p\left(l \mid x\right) = \operatorname{softmax}\left(\mathrm{LP}\left(\frac{1}{T} \sum_{t} z_t\right)\right) \tag{10}$$

$$\mathcal{L}_{\mathrm{LP}} = -\sum_{x} \log p\left(l \mid x\right) \tag{11}$$

Combined Model The combined model uses a single encoder, E, and the three objective decoders $\{D_{\mathrm{LF}}, D_{\mathrm{NL}}, \mathrm{LP}\}$ to generate outputs. During training, an English query is encoded and input to all three output systems to express output loss as $\mathcal{L}_{\mathrm{LF}} + \mathcal{L}_{\mathrm{NL}} + \mathcal{L}_{\mathrm{LP}}$. For additional languages without (x, y) pairs, the utterance is encoded and then input only to the auxiliary objectives for a combined loss as $\mathcal{L}_{\mathrm{NL}} + \mathcal{L}_{\mathrm{LP}}$. During inference, an utterance is encoded and *always* input to D_{LF} to predict a logical form, $\hat{y}$, regardless of source language, l.

[2] This approach is similar to adversarial methods (Goodfellow et al., 2014), however, our approach is not generative.

$$\frac{\partial \mathcal{L}}{\partial \theta_E} = \alpha_{LF}\frac{\partial \mathcal{L}_{\mathcal{L}F}}{\partial \theta_E} + \alpha_{NL}\frac{\partial \mathcal{L}_{NL}}{\partial \theta_E} - \lambda\alpha_{LP}\frac{\partial \mathcal{L}_{LP}}{\partial \theta_E} \tag{12}$$

$$\lambda = \frac{2}{1+e^{-10p}} + 1 \tag{13}$$

During the backward pass, each output head backpropagates the gradient signal from the respective objective function. For the encoder, these signals are combined as Equation (12) where $\alpha_{\{\text{LF, NL, LP}\}}$ are loss weightings and λ is the reversed gradient scheduling parameter from (Ganin et al., 2016). The value of λ increases with training progress p according to Equation (13) to limit the impact of noisy predictions during early training.

Our intuition is that the parser will adapt and recognize an encoding from an unfamiliar language through jointly training to generate logical forms and our auxiliary objectives using monolingual data. This sequence-to-sequence approach is highly flexible and may be useful for zero-shot approaches to additional generation tasks.

5 Data & Resources

5.1 Semantic Parsing

We analyze our model using two datasets to examine a broader language ensemble and a three-language multi-domain parsing benchmark. The former is a new version of the **ATIS** dataset of travel information (Dahl et al., 1994). Starting with the English utterances and simplified SQL queries from Iyer et al. (2017), using the same dataset split as Kwiatkowski et al. (2011), we align these to the parallel utterances from MultiATIS++ dataset for spoken language understanding (Xu et al., 2020). Therefore, we add executable SQL queries to 4473 training, 493 development, and 448 test utterances in Chinese (ZH), German (DE), French (FR), Spanish (ES), and Portuguese (PT) that were previously constructed for the slot-filling format of ATIS. Additionally, we align the test set to the Hindi (HI) and Turkish (TR) utterances from Upadhyay et al. (2018). We now can predict SQL for the ATIS dataset from eight natural languages. This represents a significant improvement over 2 or 3 language approaches (Sherborne et al., 2020; Duong et al., 2017). Note that MultiATIS++ Japanese set is excluded as the utterance alignment between this language and others was not recoverable.

We also examine **Overnight** (Wang et al., 2015), an eight-domain dataset covering *Basketball*, *Blocks*, *Calendar*, *Housing*, *Publications*, *Recipes*, *Restaurants*, and *Social Network* domains. This dataset comprises 13,682 English utterances paired with $\lambda-$DCS logical forms, executable in SEMPRE (Berant et al., 2013), split into 8,754/2,188/2,740 for training/validation/test respectively. The Overnight training data is only available in English and we use the Chinese and German test set translations from Sherborne et al. (2020) for multilingual evaluation. Each domain employs some distinctive linguistic structures, such as spatial relationships in *Blocks* or temporal relationships in *Calendar*, and our results on this dataset permit a detailed study on how cross-lingual transfer applies to various linguistic phenomena.

Reconstruction Data For the reconstruction task, we use questions from the MKQA corpus (Longpre et al., 2020), a translation of 10,000 samples from NaturalQuestions (Kwiatkowski et al., 2019) into 26 languages. We posit that MKQA is suitable for our auxiliary objective as (a) the utterances, as questions, closely model our test set while being highly variable in domain, (b) the corpus includes all training languages in our experiments, and (c) the balanced data between languages limits overexposure effects to any one test language to the detriment of others. For evaluating ATIS, we use 60,000 utterances from MKQA in languages with a training set (EN, DE, ZH, FR, ES, PT) for comparison. For Overnight, we use only 30,000 utterances in EN, DE, and ZH.

5.2 Pre-trained Models

The model described in Section 4 relies on some encoder model, E, to generate latent representations amenable to both semantic parsing and our additional objectives. We found our system performs poorly without using some pre-trained model within the encoder to provide prior knowledge from large external corpora. Prior work observed improvements using multilingual BERT (Devlin et al., 2019) and XLM-Roberta (Conneau et al., 2020a). However, we find that these models underperformed at cross-lingual transfer into parsing and we instead report results using the encoder of the **mBART50** pre-trained sequence-to-sequence model (Tang et al., 2020). This model extends **mBART** (Liu et al., 2020), trained on a monolingual de-noising objective over 25 languages, to 50 languages using multilingual bitext. We note that the pre-training data is not balanced across languages, with English as the highest resource and Portuguese as the lowest.

6 Experiments

Setting The implementation of our model, described in Section 4, largely follows parameter settings from Liu et al. (2020) for a Transformer encoder-decoder model. The encoder, E, decoders, $\{D_{\text{LF}}, D_{\text{NL}}\}$, and embedding matrices all use a dimension size of 1024 with the self-attention projection of 4096 and 16 heads per layer. Both decoders are 6-layer stacks. Weights were initialized by sampling from normal distribution $\mathcal{N}\left(0, 0.02\right)$. When using mBART50 to initialize the encoder, we use all 12 pre-trained layers frozen and append one additional layer. The domain prediction network is a two-layer feed-forward network projecting from z to 1024 hidden units then to $|L|$ for L languages. For the reconstruction noising function, we use token masking to randomly replace u tokens in x with "<mask>". u is sampled from $U\left(0, v\right)$ and we optimize $v = 3$ as the best setting.

The system was trained using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1×10^{-4}, and a weight decay factor of 0.1. We use a "Noam" schedule for the learning rate (Vaswani et al., 2017) with a warmup of 5000 steps. Loss weighting values for $\alpha_{\{\text{LF, NL, LP}\}}$ were set to $\{1, 0.33, 0.1\}$ respectively. Batches during training were size 50 and homogeneously sampled from either $\mathcal{S}_{\text{LF}}$ or $\mathcal{S}_{\text{NL}}$, with an epoch consuming one pass over both. Models were trained for a maximum of 100 epochs with early stopping. Model selection and hyperparameters were tuned on the $\mathcal{S}_{\text{LF}}$ validation set in English. Test predictions were generated using beam search with 5 hypotheses. Evaluation is reported as *denotation accuracy*, computed by comparing the retrieved denotation from the prediction, $\hat{y}$, inside the knowledge base to that from executing the gold-standard logical form.

Each model is trained on 1 NVIDIA RTX3090 GPU. All models were implemented using AllenNLP (Gardner et al., 2018) and PyTorch (Paszke et al., 2017), using pre-trained models from HuggingFace (Wolf et al., 2019).

Baseline We compare to a back-translation baseline for both datasets. Here, the test set in all languages is translated to English using Google Translate (Wu et al., 2016) and input to reference sequence-to-sequence model trained on only English queries. We consider improving upon this "minimum effort" approach as a lower-bound for justifying our approach. Additionally, we compare to an upper-bound of professional training data translation for MultiATIS++ in DE, ZH, FR, ES, and PT. This represents the "maximum effort" strategy that our system approaches.

7 Results

We implement the model described in Section 4, with hyper-parameters described in Section 6, and report our results for ATIS in Table 1 and Overnight in Table 2. Additional results for Overnight are reported in Tables 3-5 in Appendix A. For both datasets, we present ablated results for a parser without auxiliary objectives (D_{LF} only), using only parsing and reconstruction ($D_{\text{LF}}+D_{\text{NL}}$) and finally our full model using two auxiliary objectives with parsing ($D_{\text{LF}} + D_{\text{NL}} + \text{LP}$).

Comparing to baselines, we find that generally, the approach without auxiliary objectives performs worse than back-translation. This is unsurprising, as this initial approach relies on only information captured during pre-training to be capable at parsing non-English. This result is similar to Li et al. (2021) in observing a large penalty from cross-lingual transfer. However, our approach is differentiated in that we now *improve* upon this with our multi-task model. Comparing our auxiliary objectives, we broadly find more benefit to monolingual reconstruction than language prediction. Four ATIS test languages improve by $> 10\%$ using the reconstruction objective, whereas the largest improvement from adding language prediction is $+8\%$.

ATIS We identify improvements in accuracy across all languages by incorporating our reconstruction objective and then further benefit with the language prediction objective. The strongest improvements are for Chinese ($+20.7\%$) and Portuguese ($+18.0\%$) from Model (1) to Model (3). This is a sensible result for Portuguese, given that this language has comparatively very little data (49,446 sentences) during pre-training in Tang et al. (2020). Chinese is strongly represented during this same pre-training, with 10,082,367 sentences, however, Model (1) performs extremely poorly here. Our observed improvement for Chinese could be a consequence of this language being less orthographically similar to other pre-training languages. This could result in poorer cross-lingual information sharing during pre-training and therefore leave

Model	EN	DE	ZH	FR	ES	PT	HI	TR
Baseline								
Monolingual training	77.2	66.6	64.9	67.8	64.1	66.1	-	-
Back-translation	-	56.9	51.4	58.2	57.9	57.3	52.6	52.7
Our model								
(1) D_{LF} only	77.2	50.2	38.5	61.3	46.5	42.5	40.4	37.3
(2) $D_{\mathrm{LF}} + D_{\mathrm{NL}}$	**77.7**	61.1	51.2	62.7	**58.2**	57.5	49.5	44.7
(3) $D_{\mathrm{LF}} + D_{\mathrm{NL}} + \mathrm{LP}$	76.3	**67.1**	**59.2**	**66.9**	**58.2**	**60.5**	**54.1**	**47.1**

Table 1: Denotation accuracy for ATIS (Dahl et al., 1994) in English (EN), German (DE), Chinese (ZH), French (FR), Spanish (ES), Portuguese (PT), Hindi (HI) and Turkish (TR) using data from Upadhyay et al. (2018); Xu et al. (2020). We report results for multiple systems: (1) using only English semantic parsing data, (2) Multitask combined parsing and reconstruction and (3) Multitask parsing, reconstruction and language prediction. We compare to a back-translation baseline and monolingual upper-bound results where available.

Model	EN	DE	ZH
Baseline			
Back-translation	-	60.1	48.1
Our model			
(1) D_{LF} only	80.5	58.4	48.0
(2) $D_{\mathrm{LF}} + D_{\mathrm{NL}}$	81.3	62.7	49.5
(3) $D_{\mathrm{LF}} + D_{\mathrm{NL}} + \mathrm{LP}$	**81.4**	**64.3**	**52.7**

Table 2: Average denotation accuracy across all domains for Overnight (Wang et al., 2015). for English (EN), German (DE) and Chinese (ZH). We report results for multiple systems: (1) using only English semantic parsing data, (2) Multi-task combined parsing and reconstruction and (3) Multitask parsing, reconstruction and language prediction. We compare to a back-translation baseline for both ZH and DE.

more to be gained in our downstream task. We find less improvement across our models in languages closer to English, finding only a $+5.6\%$ improvement for French and $+11.7\%$ for Spanish. However, only for these languages do we approach the supervised upper-bound with -0.9% error for French and a $+0.5\%$ gain for German. Given we used no semantic parsing data in these languages, this represents significant competition to methods requiring dataset translation for success.

We also note that our system improves parsing on Hindi and Turkish, even though these are absent from the auxiliary training objectives. The model is not trained to reconstruct or identify either of these languages, however, we find that our system improves parsing regardless. By adapting our latent representation to multiple generation objectives and encouraging language-agnostic encodings – we find the model can generate better latent representations for two typologically diverse languages without guidance. This result suggests that our approach has a wider benefit to cross-lingual transfer beyond the languages we explicitly desire to transfer towards. We find our system improves above our baseline for Hindi but not Turkish. This could be another consequence of unbalanced pre-training, given the 1,327,206 Hindi sentences compared to 204,200 for Turkish.

Overnight Table 2 similarly identifies improvement for both German and Chinese using monolingual reconstruction and language prediction objectives. We observe smaller improvements between Model (1) and Model (3) compared to ATIS, $+5.9\%$ for German and $+4.7\%$ for Chinese, which may be a consequence of increased question complexity across Overnight domains. Results for individual domains are given in Appendix A for brevity, however, we did not observe any strong trends across domains in either language. Comparatively, our best system is more accurate for German than Chinese by 9.6%, which may be another consequence of orthographic dissimilarity between languages during pre-training and our approach.

Finally, we also note the capability of our model for English appears mostly unaffected by our additional objectives. For both ATIS and Overnight, we observe no catastrophic forgetting (McCloskey and Cohen, 1989) for the source language and, in some cases, a marginal performance improvement from our multi-task objectives. While semantic parsing for English is well studied and not the focus of our work, this result suggests there is minimal required compromise in maintaining parsing accuracy for English and transferring this capability to other languages.

8 Conclusion

We present a new approach to zero-shot cross-lingual semantic parsing for accurate parsing of languages without paired training data. We define and evaluate a multi-task model combining logical form generation with auxiliary objectives that require only unlabeled, monolingual corpora. This approach minimizes the error from cross-lingual transfer and improves parsing accuracy across all test languages. We demonstrate that an absence of semantic parsing data can be overcome through aligning latent representations and extend this to examine languages also unseen during our multi-task alignment training. In the future, we plan to explore additional objectives, such as translation, in our model and consider larger and more diverse corpora for our auxiliary objectives.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *ArXiv*, abs/1812.10464.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. 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.

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

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

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 *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 43–48, Stroudsburg, PA, USA.

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.

Li Dong and Mirella Lapata. 2016. Language to Logical Form with Neural Attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33–43, Stroudsburg, PA, USA.

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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora.

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*, pages 379–389, Vancouver, Canada.

Arash Einolghozati, Abhinav Arora, Lorena Sainz-Maza Lecanda, Anuj Kumar, and Sonal Gupta. 2021. El volumen louder por favor: Code-switching in task-oriented semantic parsing. In *EACL2021*. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-Resource Translation with Multi-Lingual Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Stroudsburg, PA, USA.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *ArXiv*, abs/1803.07640.

Ofer Givoli and Roi Reichart. 2019. Zero-shot semantic parsing for instructions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4454–4464, Florence, Italy. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-SQL in cross-domain database with intermediate representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4524–4535, Florence, Italy. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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, Stroudsburg, PA, USA.

Xiaodong He, Li Deng, Dilek Hakkani-Tür, and Gokhan Tur. 2013. Multi-style adaptive training for robust cross-lingual spoken language understanding. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Dan Hendrycks and Kevin Gimpel. 2020. Gaussian error linear units (gelus).

Daniel Hershcovich, Zohar Aizenbud, Leshem Choshen, Elior Sulem, Ari Rappoport, and Omri Abend. 2019. Sem-Eval-2019 task 1: Cross-lingual semantic parsing with UCCA. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1–10, Minneapolis, Minnesota, USA.

Jonathan Herzig and Jonathan Berant. 2018. Decoupling Structure and Lexicon for Zero-Shot Semantic Parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1619–1629, Brussels, Belgium.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit Alignment Objectives for Multilingual Bidirectional Encoders. In *NAACL2021*.

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.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973, Vancouver, Canada.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. 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, Stroudsburg, PA, USA.

Zhanming Jie and Wei Lu. 2014. Multilingual Semantic Parsing : Parsing Multiple Languages into Semantic Representations. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1291–1301, Dublin, Ireland.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Thomas Kollar, Danielle Berry, Lauren Stuart, Karolina Owczarzak, Tagyoung Chung, Lambert Mathias, Michael Kayser, Bradford Snow, and Spyros Matsoukas. 2018. The Alexa meaning representation language. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 177–184, New Orleans - Louisiana.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob

Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2011. Lexical Generalization in CCG Grammar Induction for Semantic Parsing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1512–1523, Edinburgh, Scotland, UK.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised Machine Translation Using Monolingual Corpora Only. In *ICLR2018*. arXiv.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark.

Percy Liang. 2016. Learning executable semantic parsers for natural language understanding. *Commun. ACM*, 59(9):68–76.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation.

Kevin Lin, Ben Bogin, Mark Neumann, Jonathan Berant, and Matt Gardner. 2019. Grammar-based neural text-to-sql generation. *CoRR*, abs/1905.13326.

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

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.

Wei Lu. 2014. Semantic parsing with relaxed hybrid trees. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1308–1318, Doha, Qatar.

Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *International Conference on Learning Representations*.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, 24(C):109–165.

Mehrad Moradshahi, Giovanni Campagna, Sina Semnani, Silei Xu, and Monica Lam. 2020. Localizing open-ontology QA semantic parsers in a day using machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. 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.

A. See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Peter Shaw, Philip Massey, Angelica Chen, Francesco Piccinno, and Yasemin Altun. 2019. Generating logical forms from graph representations of text and entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 95–106, Florence, Italy. Association for Computational Linguistics.

Tom Sherborne, Yumo Xu, and Mirella Lapata. 2020. Bootstrapping a crosslingual semantic parser. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 499–517, Online. Association for Computational Linguistics.

Alane Suhr, Ming-Wei Chang, Peter Shaw, and Kenton Lee. 2020. Exploring unexplored generalization challenges for cross-database semantic parsing. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8372–8388, Online. Association for Computational Linguistics.

Raymond Hendy Susanto and Wei Lu. 2017a. 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, Stroudsburg, PA, USA.

Raymond Hendy Susanto and Wei Lu. 2017b. Semantic parsing with neural hybrid trees. In *AAAI Conference on Artificial Intelligence*, San Francisco, California, USA.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

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.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140, Ann Arbor, Michigan. Association for Computational Linguistics.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *Proceedings of the IEEE ICASSP*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2019. Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers.

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)*, volume 1, pages 1332–1342, Stroudsburg, PA, USA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Haoran Xu and Philipp Koehn. 2021. Zero-shot cross-lingual dependency parsing through contextual embedding transformation.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Pengcheng Yin and Graham Neubig. 2017. A syntactic neural model for general-purpose code generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 440–450, Vancouver, Canada.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2020. A closer look at few-shot crosslingual transfer: Variance, benchmarks and baselines.

Victor Zhong, Mike Lewis, Sida I. Wang, and Luke Zettlemoyer. 2020. Grounded adaptation for zero-shot executable semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6869–6882, Online. Association for Computational Linguistics.

Qile Zhu, Haidar Khan, Saleh Soltan, Stephen Rawls, and Wael Hamza. 2020. Don't parse, insert: Multilingual semantic parsing with insertion based decoding. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 496–506, Online. Association for Computational Linguistics.

A Additional Results

EN	Avg.	Ba.	Bl.	Ca.	Ho.	Pu.	Res.	Rec.	So.
D_{LF} only	80.5	**90.0**	**66.7**	82.7	76.7	75.8	87.7	83.3	80.9
$D_{\mathrm{LF}} + D_{\mathrm{NL}}$	81.3	88.5	60.9	83.3	**78.8**	**83.9**	**88.6**	83.8	82.6
$D_{\mathrm{LF}} + D_{\mathrm{NL}} + \mathrm{LP}$	**81.4**	87.7	64.4	**85.1**	77.8	80.1	88.0	**85.6**	**83.0**

Table 3: Denotation accuracy for Overnight (Wang et al., 2015) using the source English data. Domains are *Basketball*, *Blocks*, *Calendar*, *Housing*, *Publications*, *Recipes*, *Restaurants* and *Social Network*.

DE	Avg.	Ba.	Bl.	Ca.	Ho.	Pu.	Res.	Rec.	So.
Baseline									
Back-translation	60.1	75.7	50.9	61.0	55.6	**50.4**	69.9	46.3	71.4
Our model									
D_{LF} only	58.4	70.3	51.1	61.9	54.0	49.7	65.4	42.1	73.1
$D_{\mathrm{LF}} + D_{\mathrm{NL}}$	62.7	73.1	56.1	66.1	**58.7**	49.7	70.2	57.9	70.1
$D_{\mathrm{LF}} + D_{\mathrm{NL}} + \mathrm{LP}$	**64.3**	**78.8**	**56.6**	**68.5**	**58.7**	46.6	**70.8**	**59.3**	**75.5**

Table 4: Denotation accuracy for Overnight (Wang et al., 2015) using the German test set from Sherborne et al. (2020). Domains are *Basketball*, *Blocks*, *Calendar*, *Housing*, *Publications*, *Recipes*, *Restaurants* and *Social Network*.

ZH	Avg.	Ba.	Bl.	Ca.	Ho.	Pu.	Res.	Rec.	So.
Baseline									
Back-translation	48.1	**62.3**	39.6	49.8	43.1	**48.3**	51.4	**29.2**	61.2
Our model									
D_{LF} only	48.0	53.7	49.6	53.0	50.8	36.0	52.1	23.1	65.3
$D_{\mathrm{LF}} + D_{\mathrm{NL}}$	49.5	56.5	49.4	55.4	**56.1**	35.4	54.2	24.9	64.1
$D_{\mathrm{LF}} + D_{\mathrm{NL}} + \mathrm{LP}$	**52.7**	59.1	**50.1**	**67.9**	54.5	41.6	**59.6**	20.8	**67.6**

Table 5: Denotation accuracy for Overnight (Wang et al., 2015) using the Chinese test set from Sherborne et al. (2020). Domains are *Basketball*, *Blocks*, *Calendar*, *Housing*, *Publications*, *Recipes*, *Restaurants* and *Social Network*.