Bootstrapping a Crosslingual Semantic Parser

Tom Sherborne, Yumo Xu and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
{tom.sherborne,yumo.xu}@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Recent progress in semantic parsing scarcely considers languages other than English but professional translation can be prohibitively expensive. We adapt a semantic parser trained on a single language, such as English, to new languages and multiple domains with minimal annotation. We query if machine translation is an adequate substitute for training data, and extend this to investigate bootstrapping using joint training with English, paraphrasing, and multilingual pre-trained models. We develop a Transformer-based parser combining paraphrases by ensembling attention over multiple encoders and present new versions of ATIS and Overnight in German and Chinese for evaluation. Experimental results indicate that MT can approximate training data in a new language for accurate parsing when augmented with paraphrasing through multiple MT engines. Considering when MT is inadequate, we also find that using our approach achieves parsing accuracy within 2% of complete translation using only 50% of training data.[1]

1 Introduction

Semantic parsing is the task of mapping natural language utterances to machine-interpretable expressions such as SQL or a logical meaning representation. This has emerged as a key technology for developing natural language interfaces, especially in the context of question answering (Kwiatkowski et al., 2013; Berant et al., 2013; Liang, 2016; Kollar et al., 2018), where a semantically complex question is translated to an executable *query* to retrieve an answer, or *denotation*, from a knowledge base.

Sequence-to-sequence neural networks (Sutskever et al., 2014) are a popular approach to semantic parsing, framing the task as *sequence transduction* from natural to formal languages (Jia and Liang, 2016; Dong and Lapata, 2016).

Recent proposals include learning intermediate logic representations (Dong and Lapata, 2018; Guo et al., 2019), constrained decoding (Yin and Neubig, 2017; Krishnamurthy et al., 2017; Lin et al., 2019), and graph-based parsing (Bogin et al., 2019; Shaw et al., 2019).

Given recent interest in semantic parsing and the data requirements of neural methods, it is unsurprising that many challenging datasets have been released in the past decade (Wang et al., 2015; Zhong et al., 2017; Iyer et al., 2017; Yu et al., 2018, 2019). However, these widely use English as synonymous for natural language. English is neither linguistically typical (Dryer and Haspelmath, 2013) nor the most widely spoken language worldwide (Eberhard et al., 2019), but is presently the *lingua franca* of both utterances and knowledge bases in semantic parsing. Natural language interfaces intended for international deployment must be adaptable to multiple locales beyond prototypes for English. However, it is uneconomical to create brand new datasets for every new language and domain.

In this regard, most previous work has focused on *multilingual* semantic parsing i.e., learning from multiple natural languages in parallel assuming the availability of multilingual training data. Examples of multilingual datasets include GeoQuery (Zelle and Mooney, 1996), ATIS (Dahl et al., 1994) and NLMaps (Haas and Riezler, 2016) but each is limited to one domain. For larger datasets, professional translation can be prohibitively expensive and require many man-hours from experts and native speakers. Recently, Min et al. (2019) reproduced the public partitions of the SPIDER dataset (Yu et al., 2018) into Chinese, but this required three expert annotators for verification and agreement. We posit there exists a more efficient strategy for expanding semantic parsing to a new language.

In this work, we consider *crosslingual semantic parsing*, adapting a semantic parser trained on English, to another language. We expand executable

---

[1]Our code and data can be found at github.com/tomsherborne/bootstrap.

semantic parsing to new languages and multiple domains by bootstrapping from in-task English datasets, task-agnostic multilingual resources, and publicly available machine translation (MT) services, in lieu of expert translation of training data. We investigate a core hypothesis that MT can provide a noisy, but reasonable, approximation of training data in a new source language. We further explore the benefit of augmenting noisy MT data using pre-trained models, such as BERT (Devlin et al., 2019), and multilingual training with English. Additionally, we examine approaches to ensembling multiple machine translations as approximate paraphrases. This challenge combines both *domain adaptation* and *localization*, as a parser must generalize to the locale-specific style of queries using only noisy examples to learn from.

For our evaluation, we present the first *multi-domain*, executable semantic parsing dataset in three languages and an additional locale for a *single-domain* dataset. Specifically, we extend ATIS (Dahl et al., 1994), pairing Chinese (ZH) utterances from Susanto and Lu (2017a) to SQL queries and create a parallel German (DE) human-translation of the full dataset. Following this, we also make available a new version of the multi-domain Overnight dataset (Wang et al., 2015) where only development and test sets are translations from native speakers of Chinese and German. This is representative of the real-world scenario where a semantic parser needs to be developed for new languages without gold-standard training data.

Our contributions can be summarized as follows: (1) new versions of ATIS (Dahl et al., 1994) and Overnight (Wang et al., 2015) for generating executable logical forms from Chinese and German utterances; (2) a combined encoder-decoder attention mechanism to ensemble over multiple Transformer encoders; (3) a cost-effective methodology for bootstrapping semantic parsers to new languages using minimal new annotation. Our proposed method overcomes the paucity of gold-standard training data using pre-trained models, joint training with English, and paraphrasing through MT engines; and (4) an investigation into practical minimum gold-standard translation requirements for a fixed performance penalty when MT is unavailable.

2 Related Work

Across logical formalisms, there have been several proposals for multilingual semantic parsing which employ multiple natural languages in parallel (Jones et al., 2012; Andreas et al., 2013; Lu, 2014; Susanto and Lu, 2017b; Jie and Lu, 2018).

Jie and Lu (2014) ensemble monolingual parsers to generate a single parse from < 5 source languages for GeoQuery (Zelle and Mooney, 1996). Similarly, Richardson et al. (2018) propose a polyglot automaton decoder for source-code generation in 45 languages. Susanto and Lu (2017a) explore a multilingual neural architecture in four languages for GeoQuery and three languages for ATIS by extending Dong and Lapata (2016) with multilingual encoders. Other work focuses on multilingual representations for semantic parsing based on universal dependencies (Reddy et al., 2017) or embeddings of logical forms (Zou and Lu, 2018).

We capitalize on existing semantic parsing datasets to bootstrap from English to another language, and therefore, do not assume that multiple languages are available as parallel input. Our work is closest to Duong et al. (2017), however they explore how to parse both English and German simultaneously using a multilingual corpus. In contrast, we consider English data only as an augmentation to improve parsing in Chinese and German and do not use "real" utterances during training. Recently, Artetxe et al. (2020) studied MT for crosslingual entailment, however, our results in Section 5 suggest these prior findings may not extend to semantic parsing, owing to the heightened requirement for factual consistency across translations.

Our work complements recent efforts in crosslingual language understanding such as XNLI for entailment (Conneau et al., 2018), semantic textual similarity (Cer et al., 2017) or the XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020) benchmarks. There has also been interest in parsing into interlingual graphical meaning representations (Damonte and Cohen, 2018; Zhang et al., 2018), spoken language understanding (Upadhyay et al., 2018) and λ-calculus expressions (Kwiatkowski et al., 2010; Lu and Ng, 2011; Lu, 2014). In contrast, we focus on logical forms grounded in knowledge-bases and therefore do not consider these approaches further.

3 Problem Formulation

Throughout this work, we consider the real-world scenario where a typical developer wishes to develop a semantic parser to facilitate question answering from an existing commercial database to

Noun/Adjective Ambiguity ("first-class fares" is a noun object)	
EN	Show me the first class fares from Baltimore to Dallas
DE$_{MT}$	Zeigen Sie mir die **erstklassigen** Tarife von Baltimore nach Dallas
DE$_H$	Zeige mir die Preise in der <u>ersten Klasse</u> von Baltimore nach Dallas

Entity Misinterpretation (Airline names aren't preserved)	
EN	Which Northwest and United flights go through Denver before noon?
DE$_{MT}$	Welche **Nordwesten** und **Vereinigten** Flüge gehen durch Denver vor Mittag
DE$_H$	Welche <u>Northwest</u> und <u>United</u> Flüge gehen durch Denver vor Mittag

Question to Statement Mistranslation (rephrased as "You have a...")	
EN	Do you have an 819 flight from Denver to San Francisco?
ZH$_{MT}$	你<u>有</u>一个从丹佛到旧金山的819 航班
ZH$_H$	<u>有没有</u>从丹佛到旧金山的819 航班

Contextual Misinterpretation ("blocks" translated to "街区" [street blocks)])	
EN	What seasons did Kobe Bryant have only three blocks?
ZH$_{MT}$	什么季节科比布莱恩特只有三个<u>街区</u>

Referential Ambiguity (他[he] refers to either players or Kobe Bryant)	
EN	Which players played more games than Kobe Bryant the seasons he played?
ZH$_{MT}$	在<u>他</u>打球的那些赛季中,哪些球员比科比布莱恩特打得更多

Table 1: Examples from ATIS (Dahl et al., 1994) and Overnight (Wang et al., 2015). Utterances are translated into Chinese and German using both machine translation (L$_{MT}$) and crowdsourcing with verification (L$_H$). We highlight issues with the noisy MT data (underlined and bolded) compared to improved human translations (underlined) for ATIS.

customers in a new locale. For example, an engineer desiring to extend support to German speakers for a commercial database of USA flights in English. Without the resources of high-valued technology companies, costs for annotation and machine learning resources must be minimized to maintain commercial viability. To economize this task, the developer must minimize new annotation or professional translation and instead bootstrap a system with public resources. At a minimum, a test and development set of utterances from native speakers are required for evaluation. However, the extent of annotation and the utility of domain adaptation for training are unknown. Therefore, our main question is *how successfully can a semantic parser learn with alternative data resources to generalize to novel queries in a new language?*

Crosslingual semantic parsing presents a unique challenge as an NLU task. It demands the generation of precise utterance semantics, aligned across languages while ensuring an accurate mapping between logical form and the idiomatic syntax of questions in every language under test. In com-

parison to NLU classification tasks such as XNLI (Conneau et al., 2018), our challenge is to **preserve** and **generate** meaning, constrained under a noisy MT channel. The misinterpretation of entities, relationships, and relative or numerical expressions can all result in an incorrect parse.

Lexical translation in MT, however accurate it may be, is insufficient alone to represent queries from native speakers. For example, the English expression "dinner flights" can be directly translated to German as "Abendessenflug" [dinner flight], but "Flug zur Abendszeit" [evening flight] better represents typical German dialogue. This issue further concerns question phrasing. For example, the English query "do you have X?" is often mistranslated to a statement "你有一个X" [you have one X] but typical Chinese employs a positive-negative pattern ("有没有一个X?" [have not have one X?]) to query possession. Our parser must overcome each of these challenges without access to gold data.

3.1 Neural Semantic Parsing

We approach our semantic parsing task using a SEQ2SEQ architecture *Transformer* encoder-decoder network (Vaswani et al., 2017). The encoder computes a contextual representation for each input token through *multi-head self-attention* by combining parallel dot-product attention weightings, or "heads", over the input sequence. The decoder repeats this self-attention across the output sequence and incorporates the source sequence through multi-head attention over the encoder output. A Transformer layer maps input $X = \{x_i\}_{i=0}^{N}$, where $x_i \in \mathbb{R}^{d_x}$, to output $Y = \{y_i\}_{i=0}^{N}$ using attention components of Query $\mathbf{Q}$, Key $\mathbf{K}$ and Value $\mathbf{V}$ in H attention heads:

$$\mathbf{e}_i^{(h)} = \frac{\mathbf{Q}W_Q^{(h)}\left(\mathbf{K}W_K^{(h)}\right)^T}{\sqrt{d_x/H}}; \; \mathbf{s}_i^{(h)} = \text{softmax}\left(\mathbf{e}_i^{(h)}\right) \quad (1)$$

$$\mathbf{z}_i^{(h)} = \mathbf{s}_i^{(h)}\left(\mathbf{V}W_V^{(h)}\right); \; \mathbf{z}_i = \text{concat}\{\mathbf{z}_i^{(h)}\}_{h=1}^{H} \quad (2)$$

$$\hat{\mathbf{y}}_i = \text{LayerNorm}\left(X + \mathbf{z}_i\right) \quad (3)$$

$$\mathbf{y}_i = \text{LayerNorm}\left(\hat{\mathbf{y}}_i + \text{FC}\left(\text{ReLU}\left(\text{FC}\left(\hat{\mathbf{y}}_i\right)\right)\right)\right) \quad (4)$$

Following Wang et al. (2019), Equation 1 describes attention scores between Query (Q) and Key (K), $\mathbf{z}_i^h$ is the hth attention head, applying scores $\mathbf{s}_i^{(h)}$ to value (V) into the multi-head attention function $\mathbf{z}_i$ with $W_{\{Q,K,V\}}^{(h)} \in \mathbb{R}^{d_x \times (d_x/H)}$. Output prediction $\mathbf{y}_i$ combines $\mathbf{z}_i$ with a residual connection and two fully-connected (FC) layers, ReLU nonlinearity, and layer normalization (Ba et al., 2016). The encoder computes self-attention through query, key, and value all equal to the input, $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} = X$. Decoder layers use self-attention over output sequence, $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} = Y_{out}$, followed by attention over the encoder output E ($\mathbf{Q} = Y_{out}$ and $\{\mathbf{K}, \mathbf{V}\} = E$) to incorporate the input encoding into decoding.

3.2 Crosslingual Modeling

Consider a parser, $\text{SP}(x)$, which transforms utterances in language x_L, to some executable logical form, y. We express a dataset in some language L as $\mathcal{D}^L = \left(\{x_n^L, y_n, d_n\}_{n=1}^N, KB\right)$, for N examples where x^L is an utterance in language L, y is the corresponding logical form and d is a denotation from knowledge base, $d = KB(y)$. The MT approximation of language L is described as J; using MT from English, $x^J = \text{MT}(x^{\text{EN}})$. Our hypothesis is that $J \approx L$ such that prediction $\hat{y} = \text{SP}(x^L)$ for test example x^L approaches gold logical form, y_{gold},
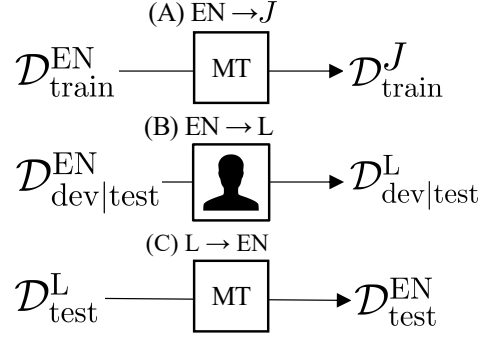

Figure 1: (A) Machine Translation (MT) from English into some language, L, for training data. J is the MT approximation of this language to be parsed. (B) Human translation of the development and test sets from English into language L. (C) Translation from language L into English using MT. Any system parsing language L must perform above this "back-translation" baseline to justify development.

conditioned upon the quality of MT. An ideal parser will output non-spurious prediction, $\hat{y}$, executing to return an equal denotation to $KB(y_{\text{gold}}) = d_{\text{gold}}$. The proportion of predicted queries which retrieve the correct denotation defines the *denotation accuracy*. Generalization performance is always measured on real queries from native speakers e.g. $\mathcal{D}^J = \{\mathcal{D}_{\text{train}}^J, \mathcal{D}_{\text{dev}}^L, \mathcal{D}_{\text{test}}^L\}$ and $\mathcal{D}_{\text{dev|test}}^J = \emptyset$.

We evaluate parsing on two languages to compare transfer learning from English into varied locales. We investigate German, a similar Germanic language, and Mandarin Chinese, a dissimilar Sino-Tibetan language, due to the purported quality of existing MT systems (Wu et al., 2016) and availability of native speakers to verify or rewrite crowd-sourced annotation. Similar to Conneau et al. (2018), we implement a "back-translate into English" baseline wherein the test set in ZH/DE is machine translated into English and a semantic parser trained on the source English dataset predicts logical forms. Figure 1 indicates how each dataset is generated. To maintain a commercial motivation for developing an in-language parser, any proposed system must perform above this baseline. Note that we do not claim to be investigating semantic parsing for low-resource languages since, by virtue, we require adequate MT into each language of interest. We use Google Translate (Wu et al., 2016) as our primary MT system and complement this with systems from other global providers. The selection and use of MT is further discussed in Appendix C.
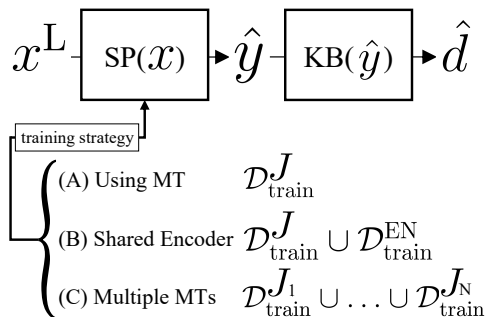
$$x^{\mathrm{L}} - \boxed{\mathrm{SP}(x)} \blacktriangleright \hat{y} - \boxed{\mathrm{KB}(\hat{y})} \blacktriangleright \hat{d}$$

training strategy

$$
\begin{cases}
\text{(A) Using MT} & \mathcal{D}^{J}_{\mathrm{train}} \\[4pt]
\text{(B) Shared Encoder} & \mathcal{D}^{J}_{\mathrm{train}} \cup \mathcal{D}^{\mathrm{EN}}_{\mathrm{train}} \\[4pt]
\text{(C) Multiple MTs} & \mathcal{D}^{J_1}_{\mathrm{train}} \cup \ldots \cup \mathcal{D}^{J_{\mathrm{N}}}_{\mathrm{train}}
\end{cases}
$$

Figure 2: The semantic parser (SP) predicts a logical form, $\hat{y}$, from an utterance in language L, x^{L}. A knowledge base (KB) executes the logical form to predict a denotation, $\hat{d}$. Approaches to crosslingual modeling involve: (A) using machine translation (MT) to approximate training data in language L; (B) training SP on both MT data and source English data; (C) using multiple MT systems to improve the approximation of L.

3.3 Feature Augmentation

Beyond using MT for in-language training data, we now describe our approach to further improve parsing using external resources and transfer learning. These approaches are described in Figure 2.

Pre-trained Representations Motivated by the success of contextual word representations for semantic parsing of English by Shaw et al. (2019), we extend this technique to Chinese and German using implementations of BERT from Wolf et al. (2019). Rather than learning embeddings for the source language *tabula rasa*, we experiment with using pre-trained 768-dimensional inputs from BERT-base in English, Chinese and German[2], as well as the multilingual model trained on 104 languages. To account for rare entities which may be absent from pre-trained vocabularies, we append these representations to learnable embeddings. Representations for logical form tokens are trained from a random initialisation, as we lack a BERT-style pre-trained model for meaning representations (i.e., $\lambda-$DCS or SQL queries). Early experiments considering multilingual word representations (Conneau et al., 2017; Song et al., 2018) yielded no significant improvement and these results are omitted for brevity.

Multilingual "Shared" Encoder Following Duong et al. (2017) and Susanto and Lu (2017a), we experiment with an encoder trained with batches from multiple languages as input. Errors in the MT data are purportedly mitigated through the

---

[2]deepset.ai/german-bert

model observing an equivalent English utterance for the same logical form. The joint training dataset is described as $\mathcal{D}^{\mathrm{EN}+J}_{\mathrm{train}} = \mathcal{D}^{\mathrm{EN}}_{\mathrm{train}} \cup \mathcal{D}^{J}_{\mathrm{train}}$ for $J = \{\mathrm{ZH}, \mathrm{DE}\}$. Consistent with Section 3.2, we measure validation and test performance using only utterances from native speakers, $\mathcal{D}^{\mathrm{L}}_{\mathrm{dev|test}}$, and ignore performance for English. This is similar to the `All` model from Duong et al. (2017), however, our objective is biased to maximize performance on one language rather than a balanced multilingual objective.

Machine Translation as Paraphrasing Paraphrasing is a common augmentation for semantic parsers to improve generalization to unseen utterances (Berant and Liang, 2014; Dong et al., 2017; Iyer et al., 2017; Su and Yan, 2017; Utama et al., 2018). While there has been some study of multilingual paraphrase systems (Ganitkevitch and Callison-Burch, 2014), we instead use MT as a paraphrase resource, similar to Mallinson et al. (2017). Each MT system will have have different outputs from different language models and therefore we hypothesize that an ensemble of multiple systems, $(J_1, \ldots J_N)$, will provide greater linguistic diversity to better approximate L. Whereas prior work uses back-translation or beam search, a developer in our scenario lacks the resources to train a NMT system for such techniques. As a shortcut, we input the same English sentence into m public APIs for MT to retrieve a set of candidate paraphrases in the language of interest (we use three APIs in experiments).

We experiment with two approaches to utilising these pseudo-paraphrases. The first, MT-Paraphrase, aims to learn a single, robust language model for L by uniformly sampling one paraphrase from $(J_1, \ldots J_N)$ as input to the model during each epoch of training. The second approach, MT-Ensemble, is an ensemble architecture similar to Garmash and Monz (2016) and Firat et al. (2016) combining attention over each paraphrase in a single decoder. For N paraphrases, we train N parallel encoder models, $\{e_n\}^{N}_{n=1}$, and ensemble across each paraphrase by combining N sets of encoder-decoder attention heads. For each encoder output, $E_n = e_n(X_n)$, we compute multi-head attention, $\mathbf{z}_i$ in Equation 2, with the decoder state, D, as the query and E_n as the key and value (Equation 5). Attention heads are combined through a combination function (Equation 6) and output $\mathbf{m}_{i\varepsilon}$ replaces $\mathbf{z}_i$ in Equation 3.

We compare ensemble strategies using two combination functions: the mean of heads (Equation 7a) and a gating network (Garmash and Monz 2016; Equation 7b) with gating function $\mathbf{g}$ (Equation 8) where $W_{\mathbf{g}} \in R^{N \times |V|}, W_h \in R^{|V| \times N|V|}$. We experimentally found the gating approach to be superior and we report results using only this method.

$$\mathbf{m}_n = \text{MultiHeadAttention}\left(D, E_n, E_n\right) \quad (5)$$

$$\mathbf{m}_{i\varepsilon} = \text{comb}\left(\mathbf{m}_1, \dots \mathbf{m}_N\right) \quad (6)$$

$$\text{comb} = \begin{cases} \frac{1}{N}\sum_n^N \mathbf{m}_n & \text{(a)} \\ \sum_n^N \mathbf{g}_n \mathbf{m}_n & \text{(b)} \end{cases} \quad (7)$$

$$\mathbf{g} = \text{softmax}\left(W_{\mathbf{g}}\text{tanh}\left(W_h[\mathbf{m}_n, \dots \mathbf{m}_N]\right)\right) \quad (8)$$

Each expert submodel uses a shared embedding space to exploit similarity between paraphrases. During training, each encoder learns a language model specific to an individual MT source, yielding diversity among experts in the final system. However, in order to improve robustness of each encoder to translation variability, inputs to each encoder are shuffled by some tuned probability p_{shuffle}. During prediction, the test utterance is input to all N models in parallel. In initial experiments, we found negligible difference in MT-Paraphrase using random sampling or round-robin selection of each paraphrase. Therefore, we assume that both methods use all available paraphrases over training. Our two approaches differ in that MT-Paraphrase uses all paraphrases sequentially whereas MT-Ensemble uses paraphrases in parallel. Previous LSTM-based ensemble approaches propose training full parallel networks and ensemble at the final decoding step. However, we found this was too expensive given the non-recurrent Transformer model. Our hybrid mechanism permits the decoder to attend to every paraphrased input and maintains a tractable model size with a single decoder.

4 Data

We consider two datasets in this work. Firstly, we evaluate our hypothesis that MT is an adequate proxy for "real" utterances using ATIS (Dahl et al., 1994). This *single-domain* dataset contains 5,418 utterances paired with SQL queries pertaining to a US flights database. ATIS was previously translated into Chinese by Susanto and Lu (2017a)

for semantic parsing into λ-calculus, whereas we present these Chinese utterances aligned with SQL queries from Iyer et al. (2017). In addition, we translate ATIS into German following the methodology described below. We use the split of 4,473/497/448 examples for train/validation/test from Kwiatkowski et al. (2011).

We also examine the *multi-domain* Overnight dataset (Wang et al., 2015), which contains 13,682 English questions paired with λ−DCS logical forms executable in SEMPRE (Berant et al., 2013). Overnight is $2.5\times$ larger than ATIS, so a complete translation of this dataset would be uneconomical for our case study. As a compromise, we collect human translations in German and Chinese only for the test and validation partitions of Overnight. We argue that having access to limited translation data better represents the crosslingual transfer required in localizing a parser. We define a fixed development partition of a stratified 20% of the training set for a final split of 8,754/2,188/2,740 for training/validation/testing. Note we consider only Simplified Mandarin Chinese for both datasets.

Crowdsourcing Translations The ATIS and Overnight datasets were translated to German and Chinese using Amazon Mechanical Turk, following best practices in related work (Callison-Burch, 2009; Zaidan and Callison-Burch, 2011; Behnke et al., 2018; Sosoni et al., 2018).

We initially collected three translations per source sentence. Submissions were restricted to Turkers from Germany, Austria, and Switzerland for German and China, USA, or Singapore for Chinese. Our AMT interface barred empty submissions and copying or pasting anywhere within the page. Any attempts to bypass these controls triggered a warning message that using MT is prohibited. Submissions were rejected if they were > 80% similar (by BLEU) to references from Google Translate (Wu et al., 2016), as were nonsensical or irrelevant submissions.

In a second stage, workers cross-checked translations by rating the best translation from each candidate set, including an MT reference, with a rewrite option if no candidate was satisfactory. We collected three judgements per set to extract the best candidate translation. Turkers unanimously agreed on a single candidate in 87.8% of the time (across datasets). Finally, as a third quality filter, we recruited bilingual native speakers to verify, rewrite, and break ties between all top candidates. Annota-

	DE	ZH
Back-translation to EN	53.9	57.8
+BERT-base	56.4	58.9
SEQ2SEQ	66.9	66.2
+BERT (de/zh)	67.8	67.4
Shared Encoder	69.3	68.3
+BERT-ML	**69.5**	**68.9**

(a) training on gold-standard data

	DE (MT)	ZH (MT)
Back-translation to EN	57.8	53.9
+BERT-base	58.9	56.4
SEQ2SEQ	61.0	55.2
+BERT-(de/zh)	64.8	57.3
Shared Encoder	64.1	58.7
+BERT-ML	66.4	59.9
MT-Paraphrase	62.2	64.5
+BERT-ML	67.8	65.0
+Shared Encoder	66.6	68.1
MT-Ensemble	63.9	62.2
+BERT-ML	64.8	65.5
+Shared Encoder	**68.5**	**68.3**

(b) training on machine translated (MT) data

Table 2: Test set denotation Accuracy for ATIS in German (DE) and Chinese (ZH).

tors chose to rewrite best candidates in only 3.2% of cases, suggesting our crowdsourced dataset is well representative of utterances from native speakers. Example translations from annotators and MT are shown in Table 1. Further details of our crowdsourcing methodology and a sample of human-translated data can be found in Appendix C.

Machine Translation All machine translation systems used in this work were treated as a blackbox. For most experiments, we retrieved translations from English to the target language with the Google Translate API (Wu et al., 2016). We use this system owing to the purported translation quality (Duong et al., 2017) and the API public availability. For ensemble approaches, we used Baidu Translate and Youdao Translate for Mandarin, and Microsoft Translator Text and Yandex Translate for German (see Appendix C).

5 Results and Analysis

We compare the neural model defined in Section 3.1 (SEQ2SEQ) to models using each augmentation outlined in Section 3.3, a combination thereof, and the back-translation baseline. Table 2(a) details experiments for ATIS using human translated training data, contrasting to Table 2(b) which substitutes MT for training data in ZH and DE. Similar results for Overnight are then presented in Table 3. Finally we consider partial translation in Figure 3. Optimization, hyperparameter settings and reproducibility details are given in Appendix A. To the best of our knowledge, we present the first results for executable semantic parsing of ATIS and Overnight in any language other than English. While prior multilingual work using $\lambda-$calculus logic is not comparable, we compare to similar results for English in Appendix B.

ATIS Table 2(a) represents the ideal case of human translating the full dataset. While this would be the least economical option, all models demonstrate performance above back-translation with the best improvement of +13.1% and +10.0% for DE and ZH respectively. This suggests that an in-language parser is preferable over MT into English given available translations. Similar to Shaw et al. (2019) and Duong et al. (2017), we find that pre-trained BERT representations and a shared encoder are respectively beneficial augmentations, with the best system using both for ZH and DE. However, the latter augmentation appears less beneficial for ZH than DE, potentially owing to decreased lexical overlap between EN and ZH (20.1%) compared to EN and DE (51.9%). This could explain the decreased utility of the shared embedding space. The accuracy of our English model is 75.4% (see Appendix B), incurring an upper-bound penalty of -6.1% for DE and -6.5% for ZH. Difficulty in parsing German, previously noted by Jie and Lu (2014), may be an artefact of comparatively complex morphology. We identified issues similar to Min et al. (2019) in parsing Chinese, namely word segmentation and dropped pronouns, which likely explain weaker parsing compared to English.

Contrasting to back-translation, the SEQ2SEQ model without BERT in Table 2(b), improves upon the baseline by +3.2% for DE and +1.3% for ZH. The translation approach for German supersedes back-translation for all models, fulfilling the minimum requirement as a useful parser. However for

	DE (MT)									ZH (MT)								
	Ba.	Bl.	Ca.	Ho.	Pu.	Rec.	Res.	So.	Avg.	Ba.	Bl.	Ca.	Ho.	Pu.	Rec.	Res.	So.	Avg.
Back-translation to EN	17.6	44.1	11.3	37.0	20.5	23.1	27.4	34.0	26.9	18.2	33.6	7.7	30.2	24.2	26.9	22.3	29.4	24.1
+BERT-base	59.1	51.6	28.6	38.6	29.8	37.0	32.2	60.0	42.1	47.1	33.6	33.9	34.4	33.5	36.6	27.4	52.9	37.4
SEQ2SEQ	76.5	47.4	70.8	51.3	67.1	70.4	62.3	73.1	64.9	78.5	51.6	55.4	64.0	62.7	69.0	66.6	73.1	65.1
+BERT-(de/zh)	74.2	56.6	**80.4**	60.8	65.8	73.6	70.8	79.2	70.2	**84.7**	48.6	64.9	73.0	68.9	68.5	70.5	78.3	69.7
Shared Encoder	72.9	58.6	75.0	60.8	**76.4**	73.1	63.6	75.9	69.5	78.0	46.1	61.3	67.7	65.2	70.4	63.6	76.5	66.1
+BERT-(de/zh)	80.8	60.4	78.6	61.4	71.4	**78.2**	66.9	**79.8**	72.2	81.1	51.4	66.7	71.4	65.2	67.6	**74.7**	77.5	69.4
MT-Paraphrase	79.5	53.4	73.8	58.7	69.6	73.1	66.9	72.4	68.4	76.0	48.6	59.5	66.7	**69.6**	63.9	66.9	76.5	65.9
+BERT-ML	82.4	55.4	73.8	**67.2**	69.6	75.9	79.2	76.7	72.5	82.4	50.4	63.7	74.6	67.7	69.9	70.5	77.4	69.6
+Shared Encoder	**82.6**	60.7	78.6	66.1	72.0	77.3	75.0	79.2	73.9	81.3	50.9	**69.6**	75.7	65.8	72.2	69.0	77.9	70.3
MT-Ensemble	72.1	55.8	74.1	54.4	67.9	70.2	64.9	68.6	66.0	71.1	45.8	58.3	62.2	61.5	62.0	61.1	71.4	61.7
+BERT-ML	81.0	57.3	73.9	62.2	68.3	74.2	**81.1**	77.6	72.0	83.6	50.2	64.3	72.1	62.1	67.1	71.4	78.0	68.6
+Shared Encoder	81.1	**66.7**	77.9	65.9	74.4	73.1	80.4	77.5	74.6	84.1	**52.9**	69.0	74.1	65.4	**73.6**	71.1	78.3	**71.1**

Table 3: Test set denotation accuracy for Overnight in German (DE) and Chinese (ZH) from training on machine translated (MT) data. Results are shown for individual domains and an eight-domain average (best results in bold). Domains are *Basketball*, *Blocks*, *Calendar*, *Housing*, *Publications*, *Recipes*, *Restaurants* and *Social Network*.

Chinese, the SEQ2SEQ approach requires further augmentation to perform above the 56.4% baseline. For ATIS, the MT-Ensemble model, with a shared encoder and BERT-based inputs, yields the best accuracy. We find that the MT-Paraphrase model performs similarly as a base model and with pre-trained inputs. As the former model has $3\times$ the encoder parameters, it may be that additional data, $\mathcal{D}_{\text{train}}^{\text{EN}}$, improves each encoder sufficiently for the MT-Ensemble to improve over smaller models. Comparing between gold-standard human translations, we find similar best-case penalties of -1.0% for DE and -0.6% for ZH using MT as training data. The model trained on MT achieves nearly the same generalization error as the model trained on the gold standard. Therefore, we consider the feasibility of our approach justified by this result.

Overnight We now extend our experiments to the multi-domain Overnight dataset, wherein we have only utterances from native speakers for evaluation, in Table 3. Whereas back-translation was competitive for ATIS, here we find a significant collapse in accuracy for this baseline. This is largely due to translation errors stemming from ambiguity and idiomatic phrasing in each locale, leading to unnatural English phrasing and dropped details in each query. Whereas Artetxe et al. (2020) found back-translation to be competitive across 15 languages for NLI, this is not the case for semantic parsing where factual consistency and fluency in parsed utterances must be maintained.

The SEQ2SEQ model with BERT outperforms the baseline by a considerable +28.1% for DE and +32.3% for ZH, further supporting the notion that an in-language parser is a more suitable strategy for the task. Our reference English parser attains an average 79.8% accuracy, incurring a penalty from crosslingual transfer of -14.9% for DE and -14.7% for ZH with the SEQ2SEQ model. Similar to ATIS, we find MT-Ensemble as the most performant system, improving over the baseline by +32.5% and +33.7% for DE and ZH respectively. The best model minimises the crosslingual penalty to -5.2% for DE and -8.7% for ZH. Across both datasets, we find that single augmentations broadly have marginal gain and combining approaches maximizes accuracy.

Challenges in Crosslingual Parsing We find several systematic errors across our results. Firstly, there are orthographic inconsistencies between translations that incur sub-optimal learned embeddings. For example, "5" can be expressed as "五" or "five". This issue also arises for Chinese measure words which are often mistranslated by MT. Multilingual BERT inputs appear to mostly mitigate this error, likely owing to pre-trained representations for each fragmented token.

Secondly, we find that multilingual training improved entity translation errors e.g. resolving translations of "the Cavs" or "coach", which are ambiguous terms for "Cleveland Cavaliers" and "Economy Class". We find that pairing the training logical form with the source English utterance allows a system to better disambiguate and correctly
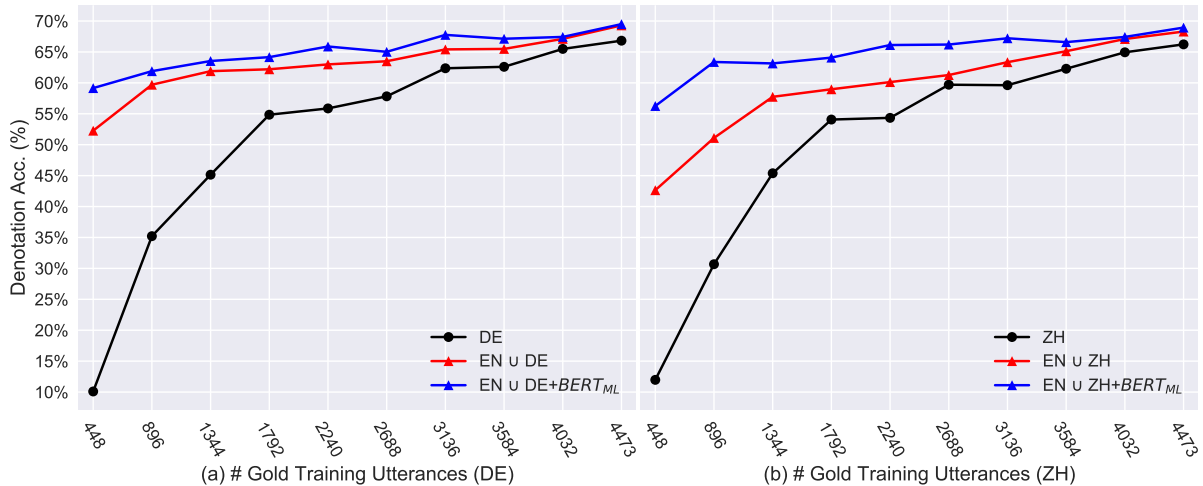
Figure 3: Denotation Accuracy against number of training examples in (a) German and (b) Chinese. Augmenting the training data with English, $EN \cup L$, uses all 4,473 English training utterances (y axis shared between figures). Each point averages results on three random splits of the dataset.

translate rare entities from DE/ZH. This disparity arises during inference because human translators are more likely to preserve named entities but this is often missed by MT with insufficient context.

Finally, paraphrasing techniques benefit parsing expressions in DE/ZH equivalent to peculiar, or KB-specific, English phrases. For example, the *Restaurants* domain heavily discusses "dollar-sign" ratings for price and "star sign" ratings for quality. There is high variation in how native speakers translate such phrases and subsequently, the linguistic diversity provided through paraphrasing benefits parsing of these widely variable utterances.

Partial Translation Our earlier experiments explored the utility of MT for training data, which assumes the availability of adequate MT. To examine the converse case, without adequate MT, we report performance with partial human-translation in Figure 3. Parsing accuracy on ATIS broadly increases with additional training examples for both languages, with accuracy converging to the best case performance outlined in Table 2(a). When translating 50% of the dataset, the SEQ2SEQ model performs -10.9% for DE and -13.1% for ZH below the ideal case. However, by using both the shared encoder augmentation and multilingual BERT ($EN \cup L + BERT_{ML}$), this penalty is minimized to -1.5% and -0.7% for DE and ZH, respectively. While this is below the best system using MT in Table 2(b), it underlines the potential of crosslingual parsing without MT as future work.

6   Conclusions

We presented an investigation into bootstrapping a crosslingual semantic parser for Chinese and German using only public resources. Our contributions include a Transformer with attention ensembling and new versions of ATIS and Overnight in Chinese and German. Our experimental results showed that a) multiple MT systems can be queried to generate paraphrases and combining these with pre-trained representations and joint training with English data can yield competitive parsing accuracy; b) multiple encoders trained with shuffled inputs can outperform a single encoder; c) back-translation can underperform by losing required details in an utterance; and finally d) partial translation can yield accuracies $< 2\%$ below complete translation using only 50% of training data. Our results from paraphrasing and partial translation suggest that exploring semi-supervised and zero-shot parsing techniques is an interesting avenue for future work.

References

Jacob Andreas, Andreas Vlachos, and Stephen Clark. 2013. Semantic parsing as machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 47–52, Sofia, Bulgaria.

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

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450.*

Maximiliana Behnke, Antonio Valerio Miceli Barone, Rico Sennrich, Vilelmini Sosoni, Thanasis Naskos, Eirini Takoulidou, Maria Stasimioti, Menno Van Zaanen, Sheila Castilho, Federico Gaspari, Panayota Georgakopoulou, Valia Kordoni, Markus Egg, and Katia Lida Kermanidis. 2018. Improving Machine Translation of Educational Content via Crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 3343–3347, Miyazaki, Japan.

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.

Jonathan Berant and Percy Liang. 2014. Semantic Parsing via Paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Stroudsburg, PA, USA.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain.

Ben Bogin, Matt Gardner, and Jonathan Berant. 2019. Representing Schema Structure with Graph Neural Networks for Text-to-SQL parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 4560–4565, Florence, Italy. Association for Computational Linguistics.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore.

Ruisheng Cao, Su Zhu, Chen Liu, Jieyu Li, and Kai Yu. 2019. Semantic parsing with dual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 51–64, Florence, Italy. Association for Computational Linguistics.

Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu. 2020. Unsupervised dual paraphrasing for two-stage semantic parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada.

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, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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.

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, 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.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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.

David Eberhard, Gary Simons, and Charles. 2019. Languages of the World. *Ethnologue: Languages of the World. Twenty-Second Edition*, 22.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving Text-to-SQL Evaluation Methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360, Melbourne, Australia.

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.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4276–4283, Reykjavik, Iceland.

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.

Ekaterina Garmash and Christof Monz. 2016. Ensemble Learning for Multi-Source Neural Machine Translation. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*.

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.

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.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Benjamin V Hanrahan, Jeffrey P Bigham, and Chris Callison-Burch. 2019. Worker Demographics and Earnings on Amazon Mechanical Turk: An Exploratory Analysis. *CHI'19 Late Breaking Work*.

Jonathan Herzig and Jonathan Berant. 2017. Neural Semantic Parsing over Multiple Knowledge-bases. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 623–628, Stroudsburg, PA, USA.

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.

Huseyin A. Inan, Gaurav Singh Tomar, and Huapu Pan. 2019. Improving semantic parsing with neural generator-reranker architecture. *ArXiv*, abs/1909.12764.

Srinivasan Iyer, Alvin Cheung, and Luke Zettlemoyer. 2019. Learning Programmatic Idioms for Scalable Semantic Parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5425–5434, Hong Kong, China.

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.

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.

Zhanming Jie and Wei Lu. 2018. Dependency-based Hybrid Trees for Semantic Parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2431–2441, Brussels, Belgium.

Bevan Keeley 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: Long Papers - Volume 1*, pages 488–496, Stroudsburg, PA, USA.

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.

Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. 2017. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1526, Copenhagen, Denmark.

Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. 2013. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1556, Seattle, Washington, USA.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1223–1233, Stroudsburg, PA, USA.

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.

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

Percy Liang, Michael I Jordan, and Dan Klein. 2013. Learning Dependency-based Compositional Semantics. *Comput. Linguist.*, 39(2):389–446.

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.

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.

Wei Lu and Hwee Tou Ng. 2011. A probabilistic forest-to-string model for language generation from typed lambda calculus expressions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1611–1622, Edinburgh, Scotland, UK.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain.

Qingkai Min, Yuefeng Shi, and Yue Zhang. 2019. A Pilot Study for Chinese SQL Semantic Parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3643–3649, Stroudsburg, PA, USA.

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*.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark.

Kyle Richardson, Jonathan Berant, and Jonas Kuhn. 2018. Polyglot Semantic Parsing in APIs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana.

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.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, New Orleans, Louisiana.

Vilelmini Sosoni, Katia Lida Kermanidis, Maria Stasimioti, Thanasis Naskos, Eirini Takoulidou, Menno Van Zaanen, Sheila Castilho, Panayota Georgakopoulou, Valia Kordoni, and Markus Egg. 2018. Translation Crowdsourcing: Creating a Multilingual Corpus of Online Educational Content. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation)*, Miyazaki, Japan.

Yu Su and Xifeng Yan. 2017. Cross-domain semantic parsing via paraphrasing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1235–1246, Copenhagen, Denmark.

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.

S. Upadhyay, M. Faruqui, G. Tür, H. Dilek, and L. Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.

P. Utama, N. Weir, F. Basik, C. Binnig, U. Cetintemel, B. Hättasch, A. Ilkhechi, S. Ramaswamy, and A. Usta. 2018. An End-to-end Neural Natural Language Interface for Databases. *ArXiv e-prints*.

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.

Chenglong Wang, Po-Sen Huang, Alex Polozov, Marc Brockschmidt, and Rishabh Singh. 2018. Execution-guided neural program decoding. *CoRR*, abs/1807.03100.

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.

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.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir Radev. 2019. SParC: Cross-domain semantic parsing in context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4511–4523, Florence, Italy. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229.

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, pages 1050–1055.

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, Brussels, Belgium.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.

Yanyan Zou and Wei Lu. 2018. Learning Cross-lingual Distributed Logical Representations for Semantic Parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 673–679, Melbourne, Australia.

7   Appendices

A   Experimental Setup

For ATIS, we implement models trained on both real and machine-translated utterances in German and Chinese. The former is our upper bound, representing the ideal case, and the latter is the minimal scenario for our developer. Comparison between these cases demonstrates both the capability of a system in the new locale and delineates the adequacy of MT for the task. Following this, we explore the multi-domain case of the Overnight dataset wherein there is no gold-standard training data in either language.

Preprocessing   Data are pre-processed by removing punctuation and lowercasing with NLTK (Bird and Loper, 2004), except for cased pre-trained vocabularies and Chinese. Logical forms are split on whitespace and natural language is tokenized using the `sentencepiece` tokeniser[3] to model language-agnostic subwords. We found this critical for Chinese, which lacks whitespace delimitation in sentences, and for German, to model word compounding. For ATIS, we experimented with the entity anonymization scheme from Iyer et al. (2017), however, this was found to be detrimental when combined with pre-trained input representations and was subsequently not used.

---

[3] github.com/google/sentencepiece

Evaluation and Model Selection   Neural models are optimized through a grid search between an embedding/hidden layer size of $2^{\{7,\dots 10\}}$, the number of layers between $\{2,\dots 8\}$, the number of heads between $\{4,\dots 8\}$ and the shuffling probability for the MT-Ensemble model between $p_{\text{shuffle}} = \{0.1,\dots 0.5\}$. The best hyperparameters had 6 layers for encoder and decoder, an embedding/hidden layer size of 128, 8 attention heads per layer, a dropout rate of 0.1 and for MT-Ensemble models, we show results for the gated combination approach, which was superior in all cases, and the optimal shuffling probability was 0.4. Models range in size from 4.2-5.7 million parameters. All weights are initialized with Xavier initialization (Glorot and Bengio, 2010) except pre-trained representations which remain frozen. Model weights, θ, are optimized using sequence cross-entropy loss against gold-standard logical forms as supervision.

Each experiment trains a network for 200 epochs using the Adam Optimizer (Kingma and Ba, 2014) with a learning rate of 0.001. We follow the Noam learning rate scheduling approach with a warmup of 10 epochs. Minimum validation loss is used as an early stopping metric for model selection, with a patience of 30 epochs. We use teacher forcing for prediction during training and beam search, with a beam size of 5, during inference.

Predicted logical forms are input to the knowledge base for ATIS, an SQL database, and Overnight, SEMPRE (Berant et al., 2013), to retrieve denotations. All results are reported as exact-match (hard) denotation accuracy, the proportion of predicted logical forms which execute to retrieve the same denotation as the reference query. Models are built using PyTorch (Paszke et al., 2017), AllenNLP (Gardner et al., 2018) and HuggingFace BERT models (Wolf et al., 2019). Each parser is trained using a cluster of 16 NVIDIA P100 GPUs with 16GB memory, with each model demanding 6-16 hours to train on a single GPU.

B   English Results

We compare our reference model for English to prior work in Table 5. Our best system for this language uses the SEQ2SEQ model outlined in Section 3.1 with input features from the pre-trained BERT-base model. We acknowledge our system performs below the state of the art for ATIS by -7.8% and Overnight by -3.9%, but this is most likely because we omit any English-specific fea-

DE	MT1	MT2	MT3	ZH	MT1	MT2	MT3
G	0.732	0.576	0.611	G	0.517	0.538	0.525
MT1	—	0.650	0.667	MT1	—	0.660	0.645
MT2	—	—	0.677	MT2	—	—	0.738

(a) ATIS

DE	MT1	MT2	MT3	ZH	MT1	MT2	MT3
MT1	—	0.570	0.513	MT1	—	0.614	0.604
MT2	—	—	0.585	MT2	—	—	0.653

(b) Overnight

Table 4: Corpus BLEU between gold-standard translations (G) and machine translations from sources 1–3 for (a) ATIS and (b) Overnight. For German (DE): MT1 is Google Translate, MT2 is Microsoft Translator Text and MT3 is Yandex. For Chinese (ZH): MT1 is Google Translate, MT2 is Baidu Translate and MT3 is Youdao Translate.

	ATIS	Overnight Ba.	Bl.	Ca.	Ho.	Pu.	Rec.	Res.	So.	Avg
Wang et al. (2015)	—	46.3	41.9	74.4	54.5	59.0	70.8	75.9	48.2	58.8
Su and Yan (2017)	—	88.2	62.7	82.7	78.8	80.7	86.1	83.7	83.1	80.8
Herzig and Berant (2017)	—	86.2	62.7	82.1	78.3	80.7	82.9	82.2	81.7	79.6
Iyer et al. (2017)	82.5	—	—	—	—	—	—	—	—	—
Wang et al. (2018)	77.9	—	—	—	—	—	—	—	—	—
Iyer et al. (2019)	**83.2**	—	—	—	—	—	—	—	—	—
Cao et al. (2019)	—	87.5	63.7	79.8	73.0	**81.4**	81.5	81.6	83.0	78.9
Inan et al. (2019)	—	**89.0**	**65.7**	**85.1**	**83.6**	**81.4**	**88.0**	**91.0**	**86.0**	**83.7**
Cao et al. (2020)	—	87.2	**65.7**	80.4	75.7	80.1	86.1	82.8	82.7	80.1
SEQ2SEQ	74.9	85.2	64.9	77.4	77.2	78.9	84.3	85.5	81.2	79.3
+BERT-base	75.4	87.7	65.4	81.0	79.4	71.4	85.6	85.8	82.0	79.8

Table 5: Test denotation accuracy on ATIS and Overnight for reference model for English. Best accuracy is bolded. Note that Inan et al. (2019) evaluate on ATIS, but use the non-executable $\lambda-$calculus logical form and are therefore not comparable to our results. Domains are *Basketball*, *Blocks*, *Calendar*, *Housing*, *Publications*, *Recipes*, *Restaurants*, and *Social Network*.

ture augmentation other than BERT. In comparison to prior work, we do not use entity anonymization, paraphrasing, execution-guided decoding or a mechanism to incorporate feedback for incorrect predictions from humans or neural critics. The closest comparable model to ours is reported by Wang et al. (2018), implementing a similar SEQ2SEQ model demonstrating 77.0% test set accuracy. However, this result uses entity anonymization for ATIS to replace each entity with a generic label for the respective entity type. Prior study broadly found this technique to yield improved parsing accuracy (Iyer et al., 2017; Dong and Lapata, 2016; Finegan-Dollak et al., 2018), a crosslingual implementation requires crafting multiple language-specific trans-lation tables for entity recognition. We attempted to implement such an approach but found it to be unreliable and largely incompatible with the vocabularies of pre-trained models.

C Data Collection

Translation through Crowdsourcing For the task of crosslingual semantic parsing, we consider the ATIS dataset (Dahl et al., 1994) and the Overnight dataset (Wang et al., 2015). The former is a single-domain dataset of utterances paired with SQL queries pertaining to a database of travel information in the USA. Overnight covers eight domains using logical forms in the $\lambda-$DCS formalism (Liang et al., 2013) which can be executed in

the SEMPRE framework (Berant et al., 2013).

ATIS has been previously translated into Chinese and Indonesian for the study of semantic parsing into $\lambda-$calculus logical forms (Susanto and Lu, 2017a), however Overnight exists only in English. To the best of our knowledge, there is presently no multi-domain dataset for executable semantic parsing in more than two languages. As previously mentioned in Section 4 , we consider Chinese and German in this paper to contrast between a language similar and dissimilar to English and also due to the reported availability of crowd-sourced workers for translation (Pavlick et al., 2014) and bilingual native speakers for verification.

To facilitate task evaluation in all languages of interest, we require a full parallel translation of ATIS in German, for comparison to the existing Chinese implementation, and a partial translation of Overnight in both German and Chinese. For task evaluation in all languages, we require a full parallel translation of ATIS to complement the existing Chinese translation from (Susanto and Lu, 2017a). As previously discussed, we translate only the development and test set of Overnight (Wang et al., 2015) into Chinese and German for assessment of crosslingual semantic parsing in a multi-domain setting. Therefore, we translate all 5,473 utterances in ATIS and 4,311 utterances in Overnight. The original Overnight dataset did not correct spelling errors from collected English paraphrases, however, we consider it unreasonable to ask participants in our task to translate misspelled words, as ambiguity in correction could lead to inaccurate translations. We subsequently identified and corrected spelling errors using word processing software.

We use Amazon Mechanical Turk (MTurk) to solicit three translations per English source sentence from crowdsourced workers (Turkers), under the assumption that this will collect at least one adequate translation (Callison-Burch, 2009). Our task design largely followed practices for translation without expert labels on MTurk (Zaidan and Callison-Burch, 2011; Post et al., 2012; Behnke et al., 2018; Sosoni et al., 2018). The task solicits translations by asking a Turker to translate 10 sentences and answer demographic questions concerning country of origin and native language. Submissions were restricted to Turkers from Germany, Austria and Switzerland or China, Singapore, and the USA for German and Chinese respectively. We built an AMT interface with quality controls which restricted Turkers from inputting whitespace and disabled copy/paste anywhere within the webpage. Attempting to copy or paste in the submission window triggered a warning that using online translation tools will result in rejection. Inauthentic translations were rejected if they held an >80% average BLEU to reference translations from Google Translate (Wu et al., 2016), as were nonsensical or irrelevant submissions. For the Chinese data collection, we also rejected submissions using Traditional Chinese Characters or Pinyin romanization. Instructions for the initial candidate collection task are given in Figure 4 and the ranking task in Figure 5. We found 94% of workers completed the optional demographic survey and that all workers reported their first language Chinese or German as desired. For Chinese, 94% of workers came from the USA and reported to have spoken Chinese for >20 years, and remaining workers resided in China. For German, all workers came from Germany and had spoken German for >25 years.

Turkers submitted 10 translations per task for $0.7 and $0.25 to rank 10 candidate translations, at an average rate to receive an equivalent full-time wage of $8.23/hour. This is markedly above the average wage for US workers of $3.01/hour discovered by Hara et al. (2019). To ensure data quality and filter disfluencies or personal biases from Turkers, we then recruited bilingual postgraduate students, native speakers of the task language, to judge if the best chosen translation from Turk was satisfactory or required rewriting. If an annotator was dissatisfied with the translation ranked best from Turk then they provided their own, which only occurred for 3.2% of all translations. Verifiers preferred the MT candidate over the Turk submissions for 29.5% of German rankings and 22.6% of Chinese rankings, however, this preference bias arose only in translations of small sentences (five or fewer words) where MT and the Turk translation were practically identical. We paid $12 an hour for this verification but to minimize cost, we did not collect multiple judgments per translation. We found that verification was completed at a rate of 60 judgments per hour, leading to an approximate cost of $2200 per language for Overnight and $2500 for ATIS into German. While this may be considered expensive, this is the minimum cost to permit comparable evaluation in every language. Sample translations for ATIS into German are given in Table 6 and sample translations for Overnight into

German and Chinese are given in Table 7.

Machine Translation   In this work, we evaluate the feasibility of using machine translation (MT) as a proxy to generate in-language training data for semantic parsing of two languages. All MT systems are treated as black-box models without inspection of underlying translation mechanics or recourse for correction. For most experiments in this work, we use translations from English to the target language using Google Translate (Wu et al., 2016). We use this system owing to the purported translation quality (Duong et al., 2017) and because the API is publicly available, contrasting to the closed MT used in Conneau et al. (2018).

Additionally, we explore two approaches to modeling an ensemble of translations from multiple MT sources. We expect, but cannot guarantee, that each MT system will translate each utterance differently for greater diversity in the training corpus overall. For this approach, we consider two additional MT systems each for Chinese and German. For Mandarin, we use Baidu Translate and Youdao Translate. For German, we use Microsoft Translator Text and Yandex Translate. To verify that the ensemble of multiple MT systems provides some additional diversity, we measure the corpus level BLEU between training utterances from each source. These scores for ATIS, with comparison to human translation, and Overnight are detailed in Table 4.

Overall, we find that each MT system provides a different set of translations, with no two translation sets more similar than any other. We also find that for ATIS in German, Wu et al. (2016) provides the most similar training dataset to the gold training data. However, we find that Microsoft Translator Text appears to narrowly improve translation into Chinese by +0.021 BLEU. This arises as an effect of a systematic preference for a polite form of Chinese question, beginning with "请" [please], preferred by the professional translator. Overall, we collected all training data using MT for < \$50 across both datasets and languages.

Translate all 10 sentences into Simplified Chinese

In this task, we ask you to provide a translation into Simplified Chinese of an English question.

You must be **native speaker of Chinese (Mandarin) and proficient in English** to complete this HIT.

We ask you to use **only Simplified Chinese characters** (简体汉字) and **do not use Pinyin** (汉语拼音).

Attempt to translate every word into Chinese. If this is difficult for rare words you do not understand, such as a person's name or place names, then please copy the English word into the translation.

You can assume all currency amounts are US Dollars and all measurements are in feet and inches.

In order to receive payment, you must complete all translations without using online translation services.

The use of online translation websites or software will be considered cheating.

Identified cheating will result in withheld payment and a ban on completing further HITs.

The demographic questionnaire is optional and you are welcome to complete as many HITs as you like.

Figure 4: Instructions provided to Turkers for the English to Chinese translation task of Overnight (Wang et al., 2015). We specify the requirement to answer in Simplified Chinese characters and specify the basis for rejection of submitted work. Instructions are condensed for brevity.

Select the best German translation for 10 English sentences

In this HIT, you will be presented with an English question and three candidate translations of this English sentence in German. We ask you to use your judgment as a native-speaker of German to select the best German translation from the three candidates.

If you consider all candidate translations to be inadequate, then provide your own translation.

You must be **native speaker of German and proficient in English** to complete this HIT.

We consider the best translation as one which asks the same question in the style of a native speaker of German, rather than the best direct translation of English. Occasionally, multiple candidates will be very similar, or identical, in this case select the first identical candidate.

You must complete all 10 to submit the HIT and receive payment.

You are welcome to submit as many HITs as you like.

Figure 5: Instructions provided to Turkers for the English to German translation ranking for both ATIS (Dahl et al., 1994) and Overnight(Wang et al., 2015). Instructions are condensed for brevity.

English	Translation into German
What ground transportation is available from the Pittsburgh airport to the town?	Welche Verkehrs Anbindung gibt es vom Pittsburgh Flughafen in die Stadt?
Could you please find me a nonstop flight from Atlanta to Baltimore on a Boeing 757 arriving at 7pm?	Könntest du für mich bitte einen Direktflug von Atlanta nach Baltimore auf einer Boeing 757 um 19 Uhr ankommend finden?
What is fare code QO mean?	Was bedeutet der ticketpreiscode QO?
Show me the cities served by Canadian Airlines International.	Zeige mir die Städte, die von den Canadian Airlines International angeflogen werden.
Is there a flight tomorrow morning from Columbus to Nashville?	Gibt es einen Flug morgen früh von Columbus nach Nashville?
Is there a Continental flight leaving from Las Vegas to New York nonstop?	Gibt es einen Continental-flug ohne Zwischenstopps, der von Las Vegas nach New York fliegt?
I would like flight information from Phoenix to Denver.	Ich hätte gerne Informationen zu Flügen von Phoenix nach Denver.
List flights from Indianapolis to Memphis with fares on Monday.	Liste Flüge von Indianapolis nach Memphis am Montag inklusive ticketpreisen auf.
How about a flight from Milwaukee to St. Louis that leaves Monday night?	Wie wäre es mit einem Flug von Milwaukee nach St. Louis, der Montag Nacht abfliegt?
A flight from St. Louis to Burbank that leaves Tuesday afternoon.	Einen Flug von St. Louis nach Burbank, der Dienstag Nachmittag abfliegt.

Table 6: Sample translations from English to German for the ATIS dataset (Dahl et al., 1994).

English	Translation into German
What kind of cuisine is Thai Cafe?	Welche Art von Küche bietet das Thai Café?
What neighborhood has the largest number of restaurants?	Welche Wohngegend hat die meisten Restaurants?
Which recipe requires the longest cooking time?	Welches Rezept benötigt die längste Kochzeit?
Which player had a higher number of assists in a season than Kobe Bryant?	Welcher Spieler hatte eine höhere Anzahl an Vorlagen in einer Saison als Kobe Bryant?
Housing with monthly rent of 1500 dollars that was posted on January 2?	Welche Wohnung hat eine monatliche Miete von 1500 Dollar und wurde am 2. Januar veröffentlicht?
What article is cited at least twice?	Welcher Artikel wurde mindestens zweimal zitiert?
What block is to the right of the pyramid shaped block?	Welcher Block befindet sich rechts neben dem pyramidenförmigen Block?
What is the birthplace of students who graduated before 2002?	Was ist der Geburtsort von Studenten, die vor 2002 ihren Abschluss gemacht haben?
Who is the shortest person in my network?	Wer ist die kleinste Person in meinem Netzwerk?
Find me the employee who quit between 2004 and 2010.	Welche Angestellten haben zwischen 2004 und 2010 gekündigt?

English	Translation into Chinese
Hotels that have a higher rating than 3 stars?	评级高于3星级的酒店
Thai restaurants that accept credit cards?	接受信用卡的泰式餐馆
Show me recipes posted in 2004 or in 2010?	告诉我2004年或2010年发布的食谱
Which player has played in fewer games than Kobe Bryant?	哪个球员比科比布莱恩特打得比赛少？
Meeting that has duration of less than three hours?	时长短于3小时的会议
Meetings in Greenberg Cafe that end at 10am?	在Greenberg咖啡厅举行并且在早上10点结束的会议
Housing units that are smaller than 123 Sesame Street?	比123芝麻街要小的房屋单元
Publisher of article citing Multivariate Data Analysis?	引用多变量数据分析的文章出版商
Block that is below at least two blocks?	在至少两个块以下的块
Find me all students who attended either Brown University or UCLA.	给我找到所有要么在布朗大学要么在UCLA上学的学生们

Table 7: Sample translations from English to German and Chinese for the Overnight dataset (Wang et al., 2015).