	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 black-box. 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× the encoder parameters, it may be that additional data, $\mathcal{D}^{\text{EN}}_{\text{train}}$, 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

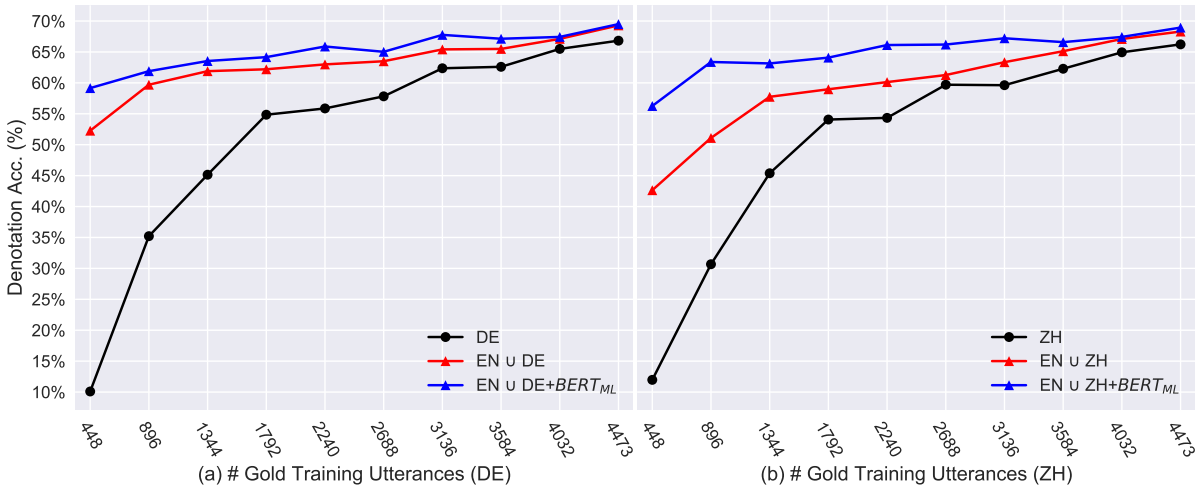

Figure 3: Denotation Accuracy against number of training examples in (a) German and (b) Chinese. Augmenting the training data with English, $EN \cup \text{L}$, uses all 4,473 English training utterances (y axis shared between figures). Each point averages results on three random splits of the dataset.

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 ($EN \cup \text{L} + BERT_{ML}$), 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.

Acknowledgements The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1; Sherborne) and the European Research Council (award number 681760; Lapata).

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.