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 $2^{\{7,\ldots 10\}}$, the number of layers between $\{2,\ldots 8\}$, the number of heads between $\{4,\ldots 8\}$ and the shuffling probability for the MT-Ensemble model between $p_{\text{shuffle}} = \{0.1,\ldots 0.5\}$. 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 translation 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