A Pilot Study for Chinese SQL Semantic Parsing

Qingkai Min , Yuefeng Shi and **Yue Zhang**

School of Engineering, Westlake University, China
Institute of Advanced Technology, Westlake Institute for Advanced Study
`{minqingkai, shiyuefeng, zhangyue}@westlake.edu.cn`

Abstract

The task of semantic parsing is highly useful for dialogue and question answering systems. Many datasets have been proposed to map natural language text into SQL, among which the recent Spider dataset provides cross-domain samples with multiple tables and complex queries. We build a Spider dataset for Chinese, which is currently a low-resource language in this task area. Interesting research questions arise from the uniqueness of the language, which requires word segmentation, and also from the fact that SQL keywords and columns of DB tables are typically written in English. We compare character- and word-based encoders for a semantic parser, and different embedding schemes. Results show that word-based semantic parser is subject to segmentation errors and cross-lingual word embeddings are useful for text-to-SQL.

1 Introduction

The task of semantic parsing is highly useful for tasks such as dialogue (Chen et al., 2013; Gupta et al., 2018; Einolghozati et al., 2019) and question answering (Gildea and Jurafsky, 2002; Yih et al., 2015; Reddy et al., 2016). Among a wide range of possible semantic representations, SQL offers a standardized interface to knowledge bases across tasks (Astrova, 2009; Xu et al., 2017; Dong and Lapata, 2018; Lee et al., 2011). Recently, Yu et al. (2018b) released a manually labelled dataset for parsing natural language questions into complex SQL, which facilitates related research.

Yu et al. (2018b)'s dataset is exclusive for English questions. Intuitively, the same semantic parsing task can be applied cross-lingual, since SQL is a universal semantic representation and database interface. However, for languages other than English, there can be added difficulties parsing into SQL. Take Chinese for example, the additional challenges can be at least two-fold. First, structures of relational databases, in particular names and column names of DB tables, are typically represented in English. This adds to the challenges to question-to-DB mapping. Second, the basic semantic unit for denoting columns or cells can be words, but word segmentation can be erroneous. It is also interesting to study the influence of other linguistic characteristics of Chinese, such as zero-pronoun, on its SQL parsing.

We investigate parsing Chinese questions to SQL by creating a first dataset, and empirically evaluating a strong baseline model on the dataset. In particular, we translate the Spider (Yu et al., 2018b) dataset into Chinese. Using the model of Yu et al. (2018a), we compare several key model configurations.

Results show that our human-translated dataset is significantly more reliable compared to a dataset composed of machine-translated questions. In addition, the overall accuracy for Chinese SQL semantic parsing can be comparable to that for English. We found that cross-lingual word embeddings are useful for matching Chinese questions with English table columns and keywords and that language characteristics have a significant influence on parsing results. We release our dataset named CSpider and code at https://github.com/taolusi/chisp.

2 Related Work

Existing datasets for semantic parsing can be classified into two major categories. The first uses logic for semantic representation, including ATIS (Price, 1990; Dahl et al., 1994) and GeroQuery (Zelle and Mooney, 1996). The second and dominant category of datasets uses SQL, which includes Restaurants (Tang and Mooney, 2001; Popescu et al., 2003), Academic (Iyer et al., 2017), Yelp and IMDB (Yaghmazadeh et al., 2017), Ad-

vising (Finegan-Dollak et al., 2018) and the recently proposed WikiSQL (Zhong et al., 2017) and Spider (Yu et al., 2018b). One salient difference between Spider and prior work is that Spider uses different databases across domains for training and testing, which can verify the generalization power of a semantic parsing model. Compared with WikiSQL, Spider further has multiple tables in each database and correspondingly more complex queries. We thus consider Spider for sourcing our dataset. Existing semantic parsing datasets for Chinese include a small corpus for assigning semantic roles (Sun and Jurafsky, 2004) and SemEval-2016 Task 9 for Chinese semantic dependency parsing (Che et al., 2012), but these data are not related to SQL. To our knowledge, we are the first to release a Chinese SQL semantic parsing dataset.

There has been a line of work improving the model of Yu et al. (2018a) since the release of the Spider dataset (Guo et al., 2019; Bogin et al., 2019; Lin et al., 2019). At the time of our investigation, however, the models are not published. We thus chose the model of Yu et al. (2018a) as our baseline. The choice of more different neural models is orthogonal to our dataset contribution, but can empirically give more insights about the conclusions.

3 Dataset

We translate all English questions in the Spider dataset into Chinese.[1] The work is undertaken by 2 NLP researchers and 1 computer science student. Each question is first translated by one annotator, and then cross-checked and corrected by a second annotator. Finally, a third annotator verifies the original and corrected versions. Statistics of the dataset are shown in Table 1. There are originally 10181 questions from Spider, but only 9691 for the training and development sets are publicly available. We thus translated these sentences only. Following the database split setting of Yu et al. (2018b), we make training, development and test sets split in a way that no database overlaps in them as shown in Table 1.

The translation work is performed on a database to database basis. For each database, the same translator translates relevant inquiries sentence by

---

[1]Note that we do not translate the database schema (i.e., column names) into Chinese because in practice databases have English schema and Chinese contents in the industry.

		# Q	# SQL	# DB	# Table/DB
English	all	10181	5693	200	5.1
Chinese	all	9691	5263	166	5.28
	train	6831	3493	99	5.38
	dev	954	589	25	4.16
	test	1906	1193	42	5.69

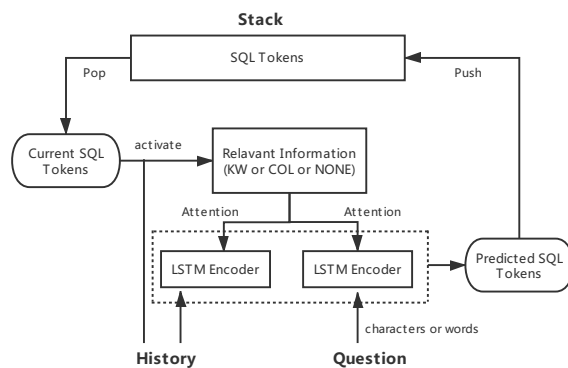Table 1: Comparisons between Spider and Chinese Spider datasets.

Figure 1: Overall structure of the Model.

sentence. The translator is asked to read the original question as well as the SQL query before making its Chinese translation. If the literal translation is possible, the translator is asked to stick to the original sentence style as much as feasible. For complex questions, the translator is allowed to rephrase the English question, so that the most natural Chinese translation is made. In addition, we keep the diversity of style in the English dataset by matching different English expressions to different Chinese expressions. A sample of our dataset is shown in Table 2. Our dataset is named CSpider.

4 Model

We use the neural semantic parsing method of Yu et al. (2018a) as the baseline model, which can be regarded as a sequence-to-tree model. In particular, the input question is encoded using an LSTM sequence encoder, and the output is a SQL query in its syntactic tree form. The tree is generated incrementally top-down, in a pre-order traversal sequence. Tree nodes include keyword nodes (e.g., SELECT, WHERE, EXCEPT) and table column name nodes (e.g., ID, City, Surname, which are defined in specific tables), which are represented in respective embedding spaces. Each keyword or column is generated by attention to the embedding space using the question representation as a key. A stack is used for incremental decoding, where

3653

Sample 1: applying only one table in one database.
SQL Query
SELECT area FROM state WHERE state_name = "New Mexico";
English Question
What is the size of New Mexico?
Translated Chinese Question
新墨西哥州的面积是多少?
Sample 2: applying multiple tables in one database.
SQL Query
SELECT T2.star_rating_description FROM HOTELS AS T1 JOIN Ref_Hotel_Star_Ratings AS T2 ON T1.star_rating_code = T2.star_rating_code WHERE T1.price_range >10000;
English Question
Give me the star rating descriptions of the hotels that cost more than 10000.
Translated Chinese Question
给出费用超过10000的酒店星级的描述。
Sample 3: with a nested SQL query.
SQL Query
SELECT T1.staff_name , T1.staff_id FROM Staff AS T1 JOIN Fault_Log AS T2 ON T1.staff_id = T2.recorded_by_staff_id EXCEPT SELECT T3.staff_name, T3.staff_id FROM Staff AS T3 JOIN Engineer_Visits AS T4 ON T3.staff_id = T4.contact_staff_id";
English Question
What is the name and ID of the staff who recorded the fault log but has not contacted any visiting engineers?
Translated Chinese Question
那些记录了错误报告但没有联系任何到访工程师的职工的姓名和ID是什么?

Table 2: Example questions corresponding to SQL.

the whole output history is leveraged as a feature for deciding the next term. This method gives the current released state-of-the-art results while submitting this paper. We provide a visualization of the model in Figure 1.

5 Experiments

We focus on comparing different word segmentation methods and different embedding representations. As discussed above, column names are selected by attention over column embeddings using sentence representation as a key. Hence there must be a link between the embeddings of columns and those of the questions. Since columns are written in English and questions in Chinese, we consider two embedding methods. The first method is to use two separate sets of embeddings for Chinese and English, respectively. We use Glove (Pennington et al., 2014)[2] for embeddings of English keywords, column names etc., and Tencent embeddings (Song et al., 2018)[3] for Chinese. The second method is to directly use the cross-lingual word embeddings. To this end, the Tencent multilingual embeddings are chosen, which contain both Chinese and English words in a multi-lingual embedding matrix.

Evaluation Metrics. We follow Yu et al. (2018b), evaluating the results using two major

types of metrics. The first is exact matching accuracy, namely the percentage of questions that have exactly the same SQL output as its reference. The second is component matching F1, namely the F1 scores for SELECT, WHERE, GROUP BY, ORDER BY and all keywords, respectively.

Hyperparameters. Our hyperparameters are mostly taken from Yu et al. (2018a), but tuned on the Chinese Spider development set. We use character and word embeddings from Tencent embedding; both of them are not fine-tuned during model training. Embedding sizes are set to 200 for both characters and words. For the different choices of keywords and column names embeddings, sizes are set to 200 and 300, respectively. Adam (Kingma and Ba, 2014) is used for optimization, with a learning rate of 1e-4. Dropout is used for the output of LSTM with a rate of 0.5.

For word-based models, segmentation is necessary. We take two segmentors with different performances, including the Jieba segmentor and the model of Yang et al. (2017), which we name Jieba and YZ, respectively. To verify their accuracy, we manually segment the first 100 sentences from the test set. Jieba and YZ give F1 scores of 89.8% and 91.7%, respectively.

5.1 Overall Results

The overall exact matching results are shown in Table 3. In this table, ENG represents the results of Yu et al. (2018a)'s model on their English dataset but under our split. HT and MT denote human translation and machine translation of questions, respectively. Both HT and MT results are evaluated on human translated questions. C-ML and C-S denote the results of our Chinese models based on characters with multi-lingual embeddings and monolingual embeddings, respectively, while WY-ML, WY-S denote the word-based models applying YZ segmentor with multi-lingual embeddings and monolingual embeddings, respectively. Finally, WJ-ML and WJ-S denote the word model with multi-lingual embeddings and monolingual embeddings with the Jieba segmentor, respectively.

First, compared to the best results of human translation (C-ML and WY-ML), machine translation results show a large disadvantage (e.g. 7.1% vs 12.1% using C-ML). We further did a manual inspection of 100 randomly picked machine-translated sentences. Out of the 100 translated

---

[2]https://nlp.stanford.edu/projects/glove/
[3]https://ai.tencent.com/ailab/nlp/embedding.html

		Easy	Medium	Hard	Extra Hard	All
	ENG	31.8%	11.3%	9.5%	2.7%	14.1%
HT	C-ML	**27.3%**	**9.9%**	7.5%	**2.3%**	**12.1%**
	C-S	23.1%	7.7%	6.2%	1.7%	9.9%
	WY-ML	21.4%	8.1%	**8.0%**	1.7%	10.0%
	WY-S	20.2%	6.4%	6.7%	2.0%	8.9%
	WJ-ML	19.8%	8.6%	5.0%	1.3%	9.2%
	WJ-S	20.1%	5.0%	5.7%	1.7%	8.2%
MT	C-ML	18.1%	4.6%	5.2%	0.3%	7.9%
	WY-ML	17.9%	4.7%	4.5%	0.3%	7.6%

Table 3: Accuracy of Exact Matching on test set.

Figure 2: Component Matching Comparisons.

sentences, 42 have translation mistakes such as semantic changes (28 sentences) and grammar errors (14 sentences). Both of these facts indicate that data by machine-translation is not reliable for semantic parsing research.

Second, comparisons among C-ML, WY-ML and WJ-ML, and among C-S, WY-S and WJ-S show that multi-lingual embeddings give superior results compared to monolingual embeddings, which is likely because they bring a better connection between natural language questions and database columns.

Third, comparisons between WY-ML and WJ-ML, and WY-S and WJ-S indicate that better segmentation accuracy has a significant influence on question parsing. Word-based methods are subject to segmentation errors.

Moreover, with the current segmentation accuracy of 92%, a word-based model underperforms a character-based model. Intuitively, since words carry more direct semantic information as compared with database columns and keywords, improved segmentation may allow a word-based model to outperform a character-based model.

Finally, for easy questions, the character-based model shows strong advantages over the word-based models. However, for medium to extremely hard questions, the trend becomes less obvious, which is likely because the intrinsic semantic complexity overwhelms the encoding differences.

Our best Chinese system gives an overall accuracy of 12.1%, [4] which is less but comparable to the English results. This shows that Chinese semantic parsing may not be significantly more challenging compared to English with text to SQL.

Component matching. Figure 2 shows F1 scores of several typical components, including SELN (SELECT NO AGGREGATOR), WHEN

(WHERE NO OPERATOR) and GBN (GROUP BY NO HAVING), applying the superior multi-lingual embeddings. The trends are consistent with the overall results.

The detailed results are shown in Table 4. Specifically, the char-based methods achieve around 41% on SELN and SEL (SELECT), which are about 5% higher compared to the word-based methods. This result may be due to the fact that word-based models are sensitive to the OOV words (Zhang and Yang, 2018; Li et al., 2019). Unlike other components, SEL and SELN are confronted with more severe OOV challenges caused by recognizing the unseen schema during testing.

In addition, the models using multi-lingual embedding overperform the models using separate embeddings on both WHEN and OB (ORDERBY), which further demonstrates that embeddings in the same dimension distribution benefit to strengthen the connection between the question and the schema.

Contrary to the overall result, the models employing the jieba segmentor perform better than those using the YZ segmentor on OB. The reason is that the jieba segmentor has different word segmentation results in terms of the superlative of adjectives. For example, the word "最高" (the highest) is segmented as "最"(most) and "高"(high) by YZ segmentor but "最高" in jieba segmentor. This again demonstrates the influence of word segmentation. Finally, for GB (GROUPBY) there is not a regular contrast pattern between different models, which can be likely because of the lack of sufficient training data.

5.2 Case study

Figure 3 shows the negative influence of segmentation errors. In particular, the incorrect segmentation of the word "店名" (shop name) leads to incorrect SQL for the whole sentence, since the

---

[4]Note that the results are lower than those reported by Yu et al. (2018a) under their split due to different training/test splits. Our split has less training data and more test instances in the "Hard" category and less in "Easy" and "Medium".
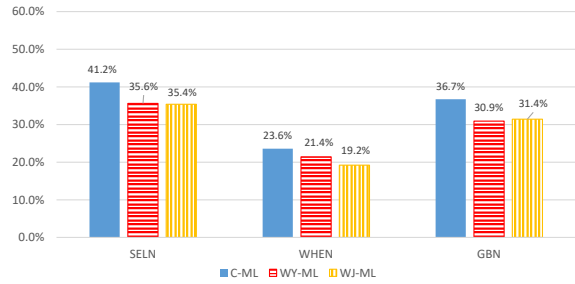
		SEL	SELN	WHE	WHEN	GB	GBN	OB
	ENG	47.3%	48.2%	19.9%	24.4%	35.0%	40.6%	57.6%
HT	C-ML	40.7%	41.2%	19.9%	23.6%	33.6%	36.7%	53.8%
	C-S	40.6%	41.0%	15.3%	17.3%	29.2%	32.9%	51.7%
	WY-ML	34.8%	35.6%	18.1%	21.4%	26.7%	30.9%	49.8%
	WY-S	34.5%	35.6%	16.5%	19.8%	30.2%	34.2%	46.9%
	WJ-ML	34.7%	35.4%	15.8%	19.2%	27.9%	31.4%	52.5%
	WJ-S	35.7%	36.8%	15.9%	19.6%	24.4%	26.8%	48.0%
MT	C-ML	36.5%	37.2%	11.3%	14.2%	29.1%	33.4%	50.7%
	WY-ML	32.1%	32.8%	11.3%	13.4%	24.8%	27.5%	49.1%

Table 4: F1 scores of Component Matching on test set.

Word segmentation error	Predicted query
哪些 商店 的 产品 数量 高于 平 均 水平 ？ 把 店 名 给 我 。 Which shops' number products is above the average? Give me the shop names.	SELECT Manager_name FROM shop WHERE Number_products > (SELECT AVG(Number_products) FROM shop)
哪些 商店 的 产品 数量 高于 平均 水平 ？ 把 店名 给 我 。 Which shops' number products is above the average? Give me the shop names.	SELECT name FROM shop WHERE Number_products > (SELECT AVG(Number_products) FROM shop)

Figure 3: Word segmentation error.

character "店" (shop) can typically be associated with "店长" (shop manager).

Figure 4 shows the sensitivity of our model to sentence patterns. In particular, the word-based model gives incorrect predictions for many question sentences frequently. As shown in the first row, the word "where" confuses the system for making a choice between "ORDER BY" and "GROUP BY". When we manually change the sentence pattern into "List the most common hometown of teachers", the parser gives the correct keyword. In contrast, the character-based model is less sensitive to question sentences, which is likely because characters are less sparse compared with words. More training data or contextualized embeddings may alleviate the issue for the word-based method, which we leave for future work.

Figure 5 shows the sensitivity of the model to Chinese linguistic patterns. In particular, the first sentence has a zero pronoun "各党的" (in each party), which is omitted later. As a result, a semantic parser cannot tell the correct database columns from the sentence. We manually add the correct entity for the zero pronoun, resulting in the second sentence. The parser can correctly identify both the column name and the table name for this corrected sentence. Since zero-pronouns are frequent

Sentence patterns	Predicted query
最 常 见 的 教师 的 家乡 是 哪里 ？ What is the most common hometowns for teachers?	SELECT Hometown FROM teacher ORDER BY Age DESC LIMIT 1
列出 最 常 见 的 教师 的 家乡 。 List the most common hometown of teachers.	SELECT Hometown FROM teacher GROUP BY Hometown ORDER BY COUNT(*) DESC LIMIT 1

Figure 4: Sentence pattern.

Chinese zero pronoun	Predicted query
代表 的 不同 党派 是 什么 ？ 显示 各党 的 党名 和 代表 人数 。 What are the different parties of representative? Show the party name and the number of representatives.	SELECT Date , COUNT(*) FROM election GROUP BY Seats
代表 的 不同 党派 是 什么 ？ 显示 各党 的 党名 和 各党 的 代表 人数 。 What are the different parties of representative? Show the party name and the number of representatives in each party.	SELECT Party , COUNT(*) FROM representative GROUP BY Party

Figure 5: Chinese zero pronoun.

for Chinese (Chen and Ng, 2016), they give added difficulty for its semantic parsing.

6 Conclusion

We constructed a first resource named CSpider for Chinese sentence to SQL, evaluating the performance of a strong English model on this dataset. Results show that the input representation, embedding forms and linguistic factors all have the influence on the Chinese-specific task. Our dataset can serve as a starting point for further research on this task, which can be beneficial to the investigation of Chinese QA and dialogue models.

Acknowledgments

References

Irina Astrova. 2009. Rules for mapping sql relational databases to owl ontologies. *Metadata and Semantics*, pages 415–424.

Ben Bogin, Jonathan Berant, and Matt Gardner. 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*, pages 4560–4565, Florence, Italy. Association for Computational Linguistics.

Wanxiang Che, Meishan Zhang, Yanqiu Shao, and Ting Liu. 2012. Semeval-2012 task 5: Chinese semantic dependency parsing. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 378–384. Association for Computational Linguistics.

Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 778–788.

Yun-Nung Chen, William Yang Wang, and Alexander I Rudnicky. 2013. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 120–125. IEEE.

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*, pages 43–48. Association for Computational Linguistics.

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.

Arash Einolghozati, Panupong Pasupat, Sonal Gupta, Rushin Shah, Mrinal Mohit, Mike Lewis, and Luke Zettlemoyer. 2019. Improving semantic parsing for task oriented dialog. *arXiv preprint arXiv:1902.06000*.

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.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

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.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2787–2792.

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.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Rubao Lee, Tian Luo, Yin Huai, Fusheng Wang, Yongqiang He, and Xiaodong Zhang. 2011. Ysmart: Yet another sql-to-mapreduce translator. In *2011 31st International Conference on Distributed Computing Systems*, pages 25–36. IEEE.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is word segmentation necessary for deep learning of chinese representations? In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3242–3252.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ana-Maria Popescu, Oren Etzioni, and Henry Kautz. 2003. Towards a theory of natural language interfaces to databases. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 149–157. ACM.

Patti J Price. 1990. Evaluation of spoken language systems: The atis domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Siva Reddy, Oscar Täckström, Michael Collins, Tom Kwiatkowski, Dipanjan Das, Mark Steedman, and Mirella Lapata. 2016. Transforming dependency structures to logical forms for semantic parsing.

Transactions of the Association for Computational Linguistics, 4:127–140.

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)*, volume 2, pages 175–180.

Honglin Sun and Daniel Jurafsky. 2004. Shallow semantc parsing of chinese. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Lappoon R Tang and Raymond J Mooney. 2001. Using multiple clause constructors in inductive logic programming for semantic parsing. In *European Conference on Machine Learning*, pages 466–477. Springer.

Xiaojun Xu, Chang Liu, and Dawn Song. 2017. Sqlnet: Generating structured queries from natural language without reinforcement learning. *arXiv preprint arXiv:1711.04436*.

Navid Yaghmazadeh, Yuepeng Wang, Isil Dillig, and Thomas Dillig. 2017. Sqlizer: query synthesis from natural language. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA):63.

Jie Yang, Yue Zhang, and Fei Dong. 2017. Neural word segmentation with rich pretraining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–849.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. 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 1321–1331.

Tao Yu, Michihiro Yasunaga, Kai Yang, Rui Zhang, Dongxu Wang, Zifan Li, and Dragomir Radev. 2018a. Syntaxsqlnet: Syntax tree networks for complex and cross-domain text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1663.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018b. 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*, pages 3911–3921.

John M Zelle and Raymond J Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the national conference on artificial intelligence*, pages 1050–1055.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.