This folder contains the old CSpider data, embeddings and pretrained models for the EMNLP 2019 paper: A Pilot Study for Chinese SQL Semantic Parsing.

Citation

@inproceedings{Min&al.19.emnlp.chisp,

title = {A Pilot Study for Chinese SQL Semantic Parsing},

author = {Qingkai Min and Yuefeng Shi and Yue Zhang},

booktitle = {2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing},

address = {Hong Kong},

year = {2019}

}

It contains the following files:

- data/

- char/

- train.json

- dev.json

- test.json

- word/

- train.json

- dev.json

- test.json

- tables.json

- test\_gold.sql

- database/

- embedding/

- char\_emb.txt

- word\_emb.txt

- models/

- char/

- saved\_models\_multi/

- saved\_models\_mono/

- word/

- saved\_models\_multi/

- saved\_models\_mono/

- README.txt

The CSpider dataset is translated from Spider dataset. 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.

For the format of each json file, please refer to Spider github page https://github.com/taoyds/spider.

In database/, there is one folder for each database which contains the [db\_name].sqlite file.

In embedding/, char\_emb.txt and word\_emb.txt corresponding to the character- and word-based methods are generated from the Tencent multilingual embeddings(Song et al., 2018). They are for the cross-lingual word embeddings schema, if to use two separate sets of embeddings for Chinese and English, additionally download the pretrained Glove(https://nlp.stanford.edu/data/wordvecs/glove.42B.300d.zip)(Pennington et al., 2014), and put it as `chisp/embedding/glove.%dB.%dd.txt`

@inproceedings{yu2018spider,

title={Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task},

author={Yu, Tao and Zhang, Rui and Yang, Kai and Yasunaga, Michihiro and Wang, Dongxu and Li, Zifan and Ma, James and Li, Irene and Yao, Qingning and Roman, Shanelle and others},

booktitle={Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing},

pages={3911--3921},

year={2018}

}

@inproceedings{song2018directional,

title={Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings},

author={Song, Yan and Shi, Shuming and Li, Jing and Zhang, Haisong},

booktitle={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},

year={2018}

}

@inproceedings{pennington2014glove,

title={Glove: Global vectors for word representation},

author={Pennington, Jeffrey and Socher, Richard and Manning, Christopher},

booktitle={Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)},

pages={1532--1543},

year={2014}

}