Semantic Parsing Multilingual Extension: German-to-SQL

Alex Hobmeier **Pietro Marini** **Tanay Sonthalia**
Georgia Institute of Technology
{ahobmeier3, pmarini3, tsonthalia}@gatech.edu

Abstract

The task of converting a natural language question into an executable SQL query, known as Text-to-SQL, is an important branch of semantic parsing. Relational databases are ubiquitous and store a great amount of structured information. The interaction with databases often requires expertise on writing structured code like SQL, which is not friendly for users who are not proficient in query languages. Existing methods of converting English questions to SQL queries exist and are readily available, but models for other languages are often unexplored. In this paper, we compare two models. The first translates German text to English and uses the existing English-to-SQL model. The second retrains the Text-to-SQL model with a German to SQL dataset. We find that the former is a more successful method of querying a database in German.

1 Introduction

Relational databases store a significant amount of the world's data. To query this Structured Data we usually use SQL (Structured Query Language). There has been growing interest in automatically converting natural language questions into executable SQL queries.

The goal of our project is the following: understand how to extend the Text-to-SQL framework to more languages. We are going to compare two different methods:

1. Machine Translation and English2SQL
2. Retrain a German2SQL Model on a Dataset of German Questions

The main issue of Text-to-SQL extension to different languages is the fact that there are many low-resourced languages and there is a great imbalance of availability of training data, with English being resource rich, and other languages having much less data, probably insufficient to train a state of the art semantic parsing model. If the MT-EN2SQL Pipeline outperforms German2SQL re-trained model we would solve this data unavailability problem and we could assume that it is feasible to translate natural language questions into structured queries potentially from every Machine Translatable language. Moreover, this could be interesting for other Semantic Parsing applications, as Xia and Monti (2021) show in their work.

2 Related Work

2.1 Portuguese-to-SQL

A different approach to extending Text-to-SQL to other languages has been investigated by José and Cozman (2021). Their work exposes that fine-tuning the multilingual BART model on a double-sized dataset (English and Portuguese) can deliver a very good performance, making inferences on the Portuguese-only test set.

Their experiment shows that multilingual pre-trained transformers can be extremely effective with languages other than English. In addition to that, if we integrate English processing with the other languages of interest, we can leverage the higher similarity of English language questions to simplify inference.

In our German2English2SQL model we are only training the model on the dataset of German questions, and not a double-sized dataset, this could be a good idea to improve performance. However, DE-EN Machine Translation should output English questions that will probably account for the similarity of input questions to output queries.

2.2 Query Validity

SQL queries, in order to retrieve the data we are looking for, must satisfy:

- syntactic validity
- semantic accuracy

A query must be syntactically valid to be executed without syntax errors. The seq2seq model might output a sequence with wrong arrangement of keywords. This can be avoided by making predictions for one slot conditioned on the values of the slots it depends on. Two techniques to make predictions

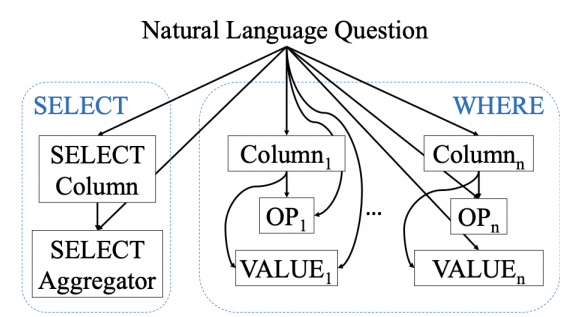

Figure 1: Sketch of syntax dependencies

base on dependencies are sequence-to-set and column attention (Xu et al.).

Semantic Accuracy is more difficult to measure, we will explain some metrics that have been developed in section 6.1.

3 WikiSQL and Multilingual Data Augementation

The pretrained English2SQL model that we used was finetuned using the WikiSQL dataset which was initially created by Zhong et al. (2017). WikiSQL is a large crowd-sourced dataset of English questions and their corresponding SQL queries.

3.1 Creating a German to SQL Dataset

The adaptation to languages other than English demands some preliminary work to train the model on the low-resourced language: translating the English questions in the WikiSQL Dataset to German. To accomplish this task, we use the state-of-the-art DeepL Translation API[1]. It is important to notice that we are going to use a different model for back-translation, because using the same one might make the translation aspect trivial.

More techniques for Multilingual Data Augmentation to generate high quality synthetic data have been investigated recently and showed promising results (Liu et al., 2021).

After this translation, a sample entry in the dataset looks like Table 1.

[1]More information on the DeepL API can be found here: https://www.deepl.com/en/docs-api/.

Language	Data
EN	What school did the player that has been in Toronto from 2010-2012 go to?
DE	Auf welche Schule ist der Spieler gegangen, der von 2010-2012 in Toronto war?
SQL	SELECT School/Club Team FROM table WHERE Years in Toronto = 2010-2012

Table 1: Example entry of modified WikiSQL dataset

4 Model 1: German2English2SQL

4.1 Translate

The Machine Translation model we used was the Marian: an Encoder-Decoder Attention Transformer with 6 layers in each component, described in Junczys-Dowmunt et al. (2018). The encoder-decoder framework used, allowed to integrate dual encoders and hard-attention without changes to beam-search or ensembling mechanisms.

4.2 English to SQL

The English2SQL model[2] we use in this paper is finetuned on Google's T5 model created in Raffel et al. (2019). The T5 model demonstrates the effectiveness of transfer-learning. Instead of creating a complete network and training on it, we can just finetune the T5 model to our use case, which in this case is Text2SQL.

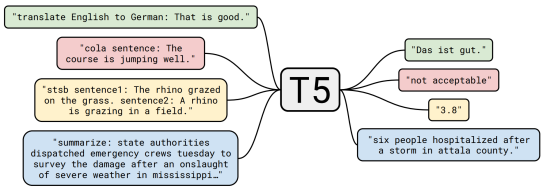

Figure 2: Diagram of Google's T5 text-to-text framework

Figure 2 demonstrates the text-to-text capabilities of the T5 model. This means that the input and output are always strings unlike BERT or other models where the output can be class labels or spans. This is a perfect application for Text2SQL because we need both a string input for the question and a string output for the SQL query.

[2]We use the following model from HuggingFace: https://huggingface.co/mrm8488/t5-base-finetuned-wikiSQL.

5 Model 2: German2SQL

The second model uses a more conventional approach by finetuning the T5 model with the modified WikiSQL dataset referenced in section 3.1. We use a multilingual tokenizer[3] to tokenize the German data. Then, we finetune the original English2SQL model on the new data for 5 epochs.

6 Results

6.1 Semantic Accuracy Metrics

We want our model to output an SQL query that is semantically correct: it must have the same denotation (meaning) as the gold query, so that it will extract the intended data from the database. There are several methods to measure semantic accuracy, we are going to focus on the most common two.

6.1.1 Exact String Match

This method only considers a single output sequence as correct, because it compares two strings: the output sequence and the gold query, they match if they are exactly the same. This will generate false negatives because there are equivalent semantically equivalent queries that differ in logical form as seen in Figure 3.

Exact String Match:

```
"SELECT NAME FROM People WHERE AGE > 34"
!=
"SELECT NAME FROM People WHERE AGE >= 35"
```

Figure 3: Describes evaluation metric difficulty with semantic equivalence but different logical form

6.1.2 Exact Set Match

For component and exact matching evaluation, instead of simply conducting string comparison between the predicted and gold SQL queries, we decompose each SQL query into several clauses, and conduct set comparison in each SQL clause. This can help handle the "ordering issue" (Xu et al.).

A potential improvement in measuring semantic accuracy - especially for hard queries - is Distilled Test Suites, an approach proposed by Zhong et al. (2020), which now is now the official metric of Spider (Yu et al., 2018), SParC (Yu et al., 2019b), and CoSQL (Yu et al., 2019a). This refinement of Exact Set Match helped identify a few concerns over ESM: the evaluation metric slightly underestimates model performance and tends not to reflect all improvements in semantic accuracy.

6.2 Model Evaluation

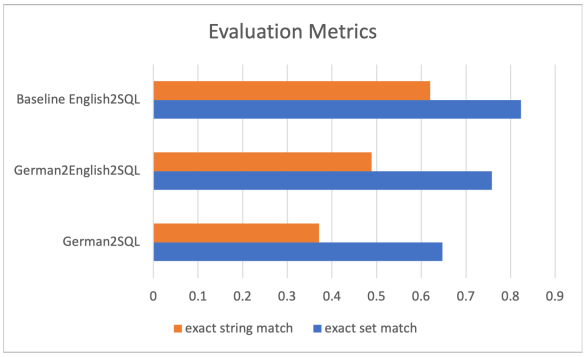

Figure 4: Evaluation results for different models discussed

The baseline English2SQL model from HuggingFace had a evaluation metric of $(0.8234, 0.6199)$ where the data is formatted as such: (exact set match, exact string match).

With the German2English2SQL model (i.e. Model 1), we received a metric of $(0.7582, 0.4886)$. This is quite good considering the loss of information during translation.

After retraining and finetuning the German2SQL model (i.e. Model 2) with the multilingual tokenizer, we received a metric of $(0.6476, 0.3713)$. It appears that the loss of English does significantly harm the model as detailed by José and Cozman (2021).

As seen in Figure 4, the German2English2SQL model far outperforms the German2SQL model.

7 Conclusion

Our exploration showed that DE-EN Machine Translated Natural Language questions can be converted in SQL queries with a semantic accuracy that is not severely impacted by Machine Translation.

The German2English2SQL model also leverages the higher similarity of English Language questions to SQL queries, something that the German2SQL retrained model can't do.

We also explored a very interesting multilingual Data Augmentation method to workaround the low-resourced languages problem for Semantic Parsing tasks. Potentially, this can be used in future research to go beyond merely Text-to-SQL towards Language Independent Semantic Parsing.

[3] We use the google/mt5-large tokenizer from HuggingFace: https://huggingface.co/google/mt5-large

References

Marcelo Archanjo José and Fábio Gagliardi Cozman. 2021. mrat-sql+gap: A portuguese text-to-sql transformer. In *Intelligent Systems - 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29 - December 3, 2021, Proceedings, Part II*, volume 13074 of *Lecture Notes in Computer Science*, pages 511–525. Springer.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2021. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. In *ACL/IJCNLP (1)*, pages 5834–5846.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Menglin Xia and Emilio Monti. 2021. Multilingual neural semantic parsing for low-resourced languages. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 185–194, Online. Association for Computational Linguistics.

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

Tao Yu, Rui Zhang, He Yang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter S Lasecki, and Dragomir Radev. 2019a. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases.

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.

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. 2019b. Sparc: Cross-domain semantic parsing in context.

Ruiqi Zhong, Tao Yu, and Dan Klein. 2020. Semantic evaluation for text-to-sql with distilled test suite. *ArXiv*, abs/2010.02840.

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