DSGA 1011: Assignment 4

Full Name: Shengduo Li
Net ID: sl11793

Q0. 1.

Please provide a link to your github repository, which contains the code for both Part I and Part II.

https://github.com/sdl1013/hw4.git

Q2. 1.

Describe your transformation of dataset.

My transformation combines both synonym replacement and typos. For synonym replacement, with a certain probability, a word is replaced by one of its WordNet synonyms, which preserves the original semantic meaning while changing the actual wording. For typos, with a small probability, one character in a word is replaced by a neighboring key on the keyboard, simulating common human typing mistakes. This transformation is reasonable because people often choose different words to express the same idea or make small typing errors, so the transformed text reflects realistic input variations.

Q3. 1

Report & Analysis

- Report the accuracy values for both the original and transformed test data evaluations.

  The accuracy for original test data evaluation is 0.93116, and the accuracy for transformed test data evaluation is 0.88312.

- Analyze and discuss the following: (1) Did the model's performance on the transformed test data improve after applying data augmentation? (2) How did data augmentation affect the model's performance on the original test data? Did it enhance or diminish its accuracy?

  (1)Yes, the model's performance on the transformed test data did improve after applying data augmentation, and the accuracy increases from 0.88312 to 0.90928. (2) Data augmentation increases the model' s performance on the original test data a little bit, and the accuracy is enhanced from 0.93116 to 0.9314.

- Offer an intuitive explanation for the observed results, considering the impact of data augmentation on model training.

  Data augmentation improves the model's generalization by allowing it to learn from both original and transformed examples. The model becomes less sensitive to small changes in wording or spelling and focuses more on semantic meaning rather than exact token patterns. Thus, it becomes more robust to these transformations, leading to improved performance on both the original and transformed test sets.

- Explain one limitation of the data augmentation approach used here to improve performance on out-of-distribution (OOD) test sets.

  One limitation is that the augmented data only reflects the specific transformations we designed, which are synonym replacement and typos, so the model improves

only for these types of changes. It may still fail on other OOD variations that are not covered by the augmentation.

Part II. Q4

Statistics Name	Train	Dev
Number of examples	4225	466
Mean sentence length	10.956923076923077	10.905579399141631
Mean SQL query length	60.901775147928994	58.896995708154506
Vocabulary size (natural language)	868	444
Vocabulary size (SQL)	644	393

Table 1: Data statistics before any pre-processing. You need to at least provide the statistics listed above, and can add new entries.

Statistics Name	Train	Dev
T5 fine-tuned model		
Mean sentence length (tokens)	23.097278106508877	23.070815450643778
Mean SQL query length (tokens)	217.37254437869822	211.05364806866953
Vocabulary size (natural language)	32100	32100
Vocabulary size (SQL)	32100	32100

Table 2: Data statistics after pre-processing. You need to at least provide the statistics listed in Table 1 (except for the number of lines), and can add new entries.

Q5

Design choice	Description
Data processing	I added the prefix "translate English to SQL:" to each nl query before passing it into the T5 tokenizer.
Tokenization	I used the default `T5TokenizerFast` to tokenize both the encoder and decoder inputs. For the decoder, I shift the target SQL sequence to the right and insert the tokenizer's `<pad>` token as the start token. All sequences were dynamically padded using `pad_sequence()` to produce the `encoder_ids`, `decoder_inputs`, and `decoder_targets`.
Architecture	I fine-tune the entire model.
Hyperparameters	I trained the model using the following hyperparameter configuration: a learning rate of 5e-5, weight decay of 0.01, and a maximum of 30 epochs with early stopping after 5 patience epochs. I used 3 warmup epochs and a cosine learning rate scheduler. The batch size was 8 for training and 16 for evaluation.

Table 3: Details of the best-performing T5 model configurations (fine-tuned)

Q6.

Quantitative Results:

System	Query EM	F1 score
Dev Results		
T5 fine-tuned Full model	1.93	71.27
Test Results		
T5 fine-tuning	0.00	67.02

Table 4: Development and test results. Use this table to report quantitative results for both dev and test results.

Qualitative Error Analysis:

Error Type	Example Of Error	Error Description	Statistics
no such column	error message: no such column:flight_stops_1.flight_id, SQL: `SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service airport_service_1, city city_1, airport_service airport_service_2, city city_2 WHERE flight_1.from_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'BOSTON' AND(flight_1.to_airport = airport_service_2.airport_code AND airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'SAN FRANCISCO' AND flight_1.flight_id = flight_stops_1.flight_id AND flight_stop_1.stops = 0 AND 1 = 1)`	The model generates column names that do not exist in the database schema.	20/466
near ")": syntax error	error message: near ")": syntax error, SQL: `SELECT DISTINCT ground_service_1.transport_type FROM ground_service ground_service_1, city city_1, days days_1, date_day date_day_1 WHERE ground_service_1.city_code = city_1.city_code AND city_1.city_name = 'PHILADELPHIA' AND ground_service_1.airport_code = airport_1.airport_code AND airport_1.airport_code = airport_1.airport_code AND airport_service_1.city_code = city_2.city_code AND city_2.city_name = 'PHILADELPHIA' AND ground_service_1.day_name = date_day_1.day_name AND date_day_1.year = 1991 AND date_day_1.month_number = 4 AND date_day_1.day_number = 23)`	The model generates SQL queries where the number of ")" exceeds the number of "(", causing a syntax error.	25/466
incomplete input error	error message: incomplete input, SQL: `SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service airport_service_1, city city_1, airport_service airport_service_2, city city_2 WHERE flight_1.airline_code = 'DL' AND(flight_1.from_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'DENVER' AND(flight_1.to_airport = airport_service_2.airport_code AND airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'DALLAS' AND flight_1.departure_time > 5)`	The model generates SQL queries where the number of "(" exceeds the number of ")", which causes the SQL parser to treat the query as incomplete.	4/466

Table 5: Use this table for your qualitative analysis on the dev set.

Q7.

Provide a link to a google drive which contains a model checkpoint used to generate outputs you have submitted.

`https://drive.google.com/file/d/1AECxTwaijNJhAsQPIN0CM_Zpz5nlbEfO/view?usp=sharing`

Extra Credit:

If you are doing extra credit assignment, please describe your system here, as well as provide a link to a google drive which contains a model checkpoint used to generate outputs you have submitted. Optional TODO