Error-Aware Interactive Semantic Parsing of OpenStreetMap

Michael Staniek
Computational Linguistics
Heidelberg University
Germany
staniek@cl.uni-heidelberg.de

Stefan Riezler
Computational Linguistics & IWR
Heidelberg University
Germany
riezler@cl.uni-heidelberg.de

Abstract

In semantic parsing of geographical queries against real-world databases such as Open-StreetMap (OSM), unique correct answers do not necessarily exist. Instead, the truth might be lying in the eye of the user, who needs to enter an interactive setup where ambiguities can be resolved and parsing mistakes can be corrected. Our work presents an approach to interactive semantic parsing where an explicit error detection is performed, and a clarification question is generated that pinpoints the suspected source of ambiguity or error and communicates it to the human user. Our experimental results show that a combination of entropy-based uncertainty detection and beam search, together with multi-source training on clarification question, initial parse, and user answer, results in improvements of 1.2% F1 score on a parser that already performs at 90.26% on the NLMaps dataset for OSM semantic parsing.

1 Introduction

Semantic Parsing has the goal of mapping natural language questions into formal representations that can be executed against a database. If real-world large-scale databases such as OpenStreetMap (OSM)[1] need to be accessed, the creation of gold standard parses by humans can be complicated and requires expert knowledge, and even reinforcement learning from answers might be impossible since unique correct answers to OSM queries do not necessarily exist. Instead, uncertainties can arise due to open-ended lists (e.g., of restaurants), fuzzily defined geo-positional objects (e.g., objects "near" or "in walking distance" of other objects), or by ambiguous mappings of natural language to OSM tags[2], with the truth lying in the eye of the beholder

who asked the original question. Semantic parsing against OSM thus asks for an interactive setup where an end-user inter-operates with a semantic parsing system in order to negotiate a correct answer, or to resolve parsing ambiguities and to correct parsing mistakes, in a dialogical process.

Previous work on interactive semantic parsing (Labutov et al., 2018; Yao et al., 2019; Elgohary et al., 2020) has put forward the following dialogue structure: i) the user poses a natural language question to the system, ii) the system parses the user question and explains or visualizes the parse to the user, iii) the user generates natural language feedback, iv) the parser tries to utilize the user feedback to improve the parse of the original user question. In most cases, the "explanation" produced by the system is restricted to a rule-based reformulation of the parse in a human intelligible form, whereas the human user has to take guesses about where the parse went wrong or is ambiguous.

The goal of our paper is to add an explicit step of error detection on the parser side, resulting in an automatically produced clarification question that pinpoints the suspected source of ambiguity or error and communicates it to the human user. Our experimental results show that a combination of entropy-based uncertainty detection and beam search for differences to the top parse yield concise clarification questions. We create a dataset of 15k clarification questions that are answered by extracting information from gold standard parses, and complement this with a dataset of 960 examples where human users answer the automatically generated questions. Supervised training of a multi-source neural network that adds clarification questions, initial parses, and user answers to the input results in improvements of 1.2% F1 score on

---

[1] www.openstreetmap.org

[2] For example, *recreation grounds* can map to tags reserved for *leisure* purposes or for official *landuse* registration; *bars*

map to tags *bar* and *pub* that differ in that only the latter sells food; *off-license* shops can have licenses to sell only *wine* or all kinds of *alcohol*.

a parser that already performs at 90.26% on the NLMaps dataset for OSM semantic parsing.

2 Related Work

Yao et al. (2019) interpret interactive semantic parsing as a slot filling task, and present a hierarchical reinforcement learning model to learn which slots to fill in which order. They claim the automatic production of clarification questions by the agent as a main feature of their approach, however, what is actually used in their work is a set of 4 predefined templates. Elgohary et al. (2020) show an interpretation of the parse that is understandable for laypeople with a template-based approach, and present different approaches to utilize the user response to improve the parser. In their work, the explantion on the parser side is purely template-based, whereas our work explicitly informs the clarification question by possible sources of parse ambiguities or errors.

Considerable effort has been invested in the creation of large datasets for parsing into SQL representations. Yu et al. (2018) created a dataset called Spider which is a complex, cross-domain semantic parsing and text-to-SQL dataset. Their annotation process was very extensive, and involved 11 computer science students who invested a total of 1,000 hours into asking natural language queries and creating the corresponding SQL query. Extensions of the Spider dataset, SParC (Yu et al., 2019b), or Co-SQL (Yu et al., 2019a) involved even more computer science students. Our work attempts an automatic construction of concise clarification questions, allowing for faster dataset construction.

3 (Multi-Source) Neural Machine Translation

Our work employs as a semantic parser a sequence-to-sequence neural network (Sutskever et al., 2014) that is based on an recurrent encoder and decoder architecture with attention (Bahdanau et al., 2015).

Given a corpus of aligned data $D = \{(x_n, y_n)\}_{n=1}^N$ of user queries x and semantic parses y, standard supervised training is performed by minimizing a Cross-Entropy objective $-\frac{1}{N}\sum_{n=1}^N \sum_{t=1}^T \log p(y_{n,t}|y_{n,<t}, x_n)$, where the probability of the full output sequence $y = y_1, y_2, ..., y_n$ is calculated by the product of the probability for every timestep where $p(y|x) = \prod_{t=1}^T p(y_t|y_{<t}, x)$.

This model can be easily extended to multi-source learning (Zoph and Knight, 2016) by using not only one, but multiple encoders. This means that there are actually multiple sequences of hidden states. The decoder hidden state is consequently initialized by a linear projection of the average of the last hidden states of all encoders $c = \frac{1}{N}\sum_{i=1}^N h_i W_l$, and needs to implement a separate attention mechanism for every encoder.

To be able to fine-tune a model with feedback from a user, the standard cross-entropy objective cannot be used because the desired target is not a gold parse, but a parse $\tilde{y}$ predicted by the system, that has been annotated with positive and negative markings by a human user. This can be formalized as assigning a reward δ_t that is either positive or negative to every token in the parse ($\delta_{t+} = 0.5$ and $\delta_{t-} = -0.5$). It is then possible to maximize the likelihood of the correct parts of the parse by optimizing a weighted supervised learning objective $\sum_{x,\tilde{y}}\sum_{t=1}^T \delta_t \log p(\tilde{y}_t|x, y_{<t})$. (Petrushkov et al., 2018)

4 Neural Semantic Parsing of OSM

4.1 Data

Our work is based on the NLmaps v2 dataset.[3] NLmaps builds on the Overpass API which allows the querying of the OSM database with natural language queries. This dataset includes template-based expansions leading to duplicates in train and test sets. However, these expansions introduced problematic features into the data in that OSM tags were inserted which, according to the documentation in the OSM developer wiki, should not be used:

- Is there Recreation Grounds in Marseille
→ `query(area(keyval('name','Marseille')),`
  `nwr(keyval('`**`leisure`**`','recreation_ground'),`
  `qtype(least(topx(1))))`
- Recreation Ground in Frankfurt am Main
→ `query(area(keyval('name','Frankfurt`
  `am Main')),`
  `nwr(keyval('`**`landuse`**`','recreation_ground')),`
  `qtype(latlong))`

While *leisure=recreation_ground* certainly exists as a tag[4], its use is heavily discouraged[5]. Furthermore, several mistakes were introduced in the

---

[3] www.cl.uni-heidelberg.de/ statnlpgroup/nlmaps/
[4] https://wiki.openstreetmap.org/wiki/ Tag:leisure%3Drecreation_ground.
[5] https://wiki.openstreetmap.org/wiki/ Tag:landuse%3Drecreation_ground.

data by the augmentation with the help of a wordlist. For example, an automatically generated natural language question based on this wordlist asks for bars, whereas the gold parse associated to that question asks for pubs instead:

- Where Bars in Bradford
→ `query(area(keyval('name','Bradford')),`
`nwr(keyval('amenity','`**`pub`**`')),`
`qtype(latlong))`

Conceptually, bars and pubs may not be that different to each other, but OSM advises a strict distinction between bars and pubs [6]. While a pub sells alcohol on premise, a pub also sells food, the athmosphere is more relaxed and the music is quieter compared to a bar.

Lastly, ambiguity was introduced because natural language words now map to multiple different OSM tags. This leads to the following data occurrences:

- shop Off Licenses in Birmingham
→ `query(area(keyval('name','Birmingham')),`
`nwr(keyval('shop','`**`alcohol`**`')),`
`qtype(findkey('shop')))`

- How many closest Off License from Wall Street in Glasgow
→ `query(around(center(area(`
`keyval('name','Glasgow')),`
`nwr(keyval('name','Wall Street'))),`
`search(nwr(keyval('shop','`**`wine`**`'))),`
`maxdist(DIST_`
`INTOWN),`
`topx(1)),qtype(count))`

The previous examples show that for the same keyword "Off License" both *shop=alcohol* and *shop=wine* are valid interpretations.

Finally, since the data was augmented first, and only afterwards split into train, development and test sets, there is a lot of overlap between the train and test data. This is problematic because a proper evaluation should also test for overfitting, which does not work if data is shared between different splits, as shown in the following examples:

- Train: cinema in Nantes

- Dev: cinema in Paris

- Test: cinemas in Paris

We applied a dataset de-duplication by removing all datapoints from the development and test sets which are identical to training datapoints when

---

System	F1
Lawrence (2018)	80.36
Lawrence (2018)+NER	90.09
token-based	83.43
character-based	93.77

Table 1: F1 results of single-source models on the original NLmaps v2 dataset.

location (e.g., *Paris*) and POI (e.g., *cinema*) are masked. This results in the dataset described in table 3.

4.2 Semantic Parsing

We use the Joey NMT (Kreutzer et al., 2019) as framework to build a baseline parser. The basic Joey NMT architecture is modified to allow for a multi-source setup (see Figure 3 in the appendix) and for learning from markings.[7]

As evaluation metrics we use exact match accuracy, defined as $\frac{1}{N} \sum_{n=1}^{N} \delta(\text{predicted, gold})$ of a predicted parse and the gold parse. Furthermore, we report F1 score as harmonic mean of recall, defined as the percentage of fully correct answers divided by the set size, and precision, defined as the percentage of correct answers out of the set of answers with non-empty strings.

A character-based Joey NMT semantic parser is able to improve the results reported in Lawrence and Riezler (2018) on the dataset without de-duplication, as shown in Table 1. All results presented in the following are relative improvements over our own baseline parser, reported on the de-duplicated dataset for which no external baseline is available.

5 Generation of Clarification Questions

On of the goals of error-aware interactive semantic parsing is to alert to user about suspected sources of ambiguity and error by initiating a dialogue. The parser thus needs to detect uncertainty in its output, and generate a clarification questions on the detected source of uncertainty. We use entropy-based uncertainty measures. Firstly, entropy per timestep t is measured as $-\sum_{\tilde{y}_t} p(\tilde{y}_t|x, y_{<t}) \log p(\tilde{y}_t|x, y_{<t})$. This is employed to calculate the entropy of a token as the mean of the character entropies for each of a

---

Figure 1: Annotation setup for human interaction study.

System	Accuracy	F1
baseline	83.50	90.26
baseline + hyps	83.66	90.85
baseline + dia	84.74	92.02
baseline + hyps + dia	84.84	91.47
baseline + hyps + dia + log	85.01	91.61

Table 2: Results of the multi-source models compared to the single-source model taking only the source into account on the modified test data.

token's characters.[8] Based on entropy information, we generate simple questions by employing a template-based method which incorporates the least certain token: "Did you mean $token?". Furthermore, we offer alternative answers for the user based on beam search of size 2. This heuristic is justified experimentally since always taking the first beam yields an accuracy of 92.7%, while another 5% of accuracy can be gained by choosing the second beam. This verifies the usefulness of proposing entries in the second beam as alternative in clarification questions: "Did you mean $token or $alternative?".

---

[8]A visualization of entropy is reported in the appendix.

Split	Count
Train	15,658
Dev	961
Test	4156

Table 3: Statistics of dialogue-enriched data.

6 Experiments on Synthetic Dialogues

In a first experiment, we generated entire dialogues synthetically, that is, the clarification question from the parser and synthetic user answers. The latter were constructed by checking if either the original token or the alternative is contained in the given gold parse. Dataset statistics for train, development and test splits are given in Table 3.

Model training is performed by extending the character-based baseline model by additional encoders for the dialogue (question and answer) and the predicted parse hypothesis. Experiments show that the character-based multi-source model including hypothesis and dialogue as additional input (line 4) outperforms the baseline (line 1) by more than 1 point in accuracy and F1 score (Table 2). This difference is statistically significant with a p-value of 0.0483 determined by approximate randomization.

7 Human Interaction Study

We furthermore performed a small field study where human users interacted with the system. Parses for queries from both train and development parts of the dataset were generated and augmented with automatically created clarification questions based on the uncertainty model. Examples were then filtered to keep only those parses that contained a parse mistake or parse ambiguity. This resulted in a total of 930 annotation tasks.[9]

The annotation interface shown in Figure 1 illustrates the system-user interaction: Human annotators are presented with a natural language query ("closest Off License from Lyon"), the parse (shown below in linearized form), and the result of the generated parse (show as the map extract on top of the figure). In addition to the linearized form of the predicted parse, a human-intelligible list format of the key-value pairs in the parse[10] is presented, following the annotation interface of Lawrence and

---

[9]Both synthetic and user data will be publicly released.
[10]https://wiki.openstreetmap.org/wiki/Map_Features

56

User Question:
Is there Off Licence in Edinburgh?

Beam Search with the trained single-source model

query@3 area@1 keyval@2 name@0 Edinburgh@s nwr@1 keyval@2 shop@0 **wine**@s qtype@1 least@1 topx@1 1@0
query@3 area@1 keyval@2 name@0 Edinburgh@s nwr@1 keyval@2 shop@0 **alcohol**@s qtype@1 least@1 topx@1 1

Entropy analysis detects wine as uncertain

Detection of alternatives to generate a question

wine

Did you mean wine or alcohol?

Beamsearch with the trained multi-source model

Interaction with a user

alcohol

query@3 area@1 keyval@2 name@0 Edinburgh@s nwr@1 keyval@2 shop@0 **alcohol**@s qtype@1 least@1 topx@1 1@0
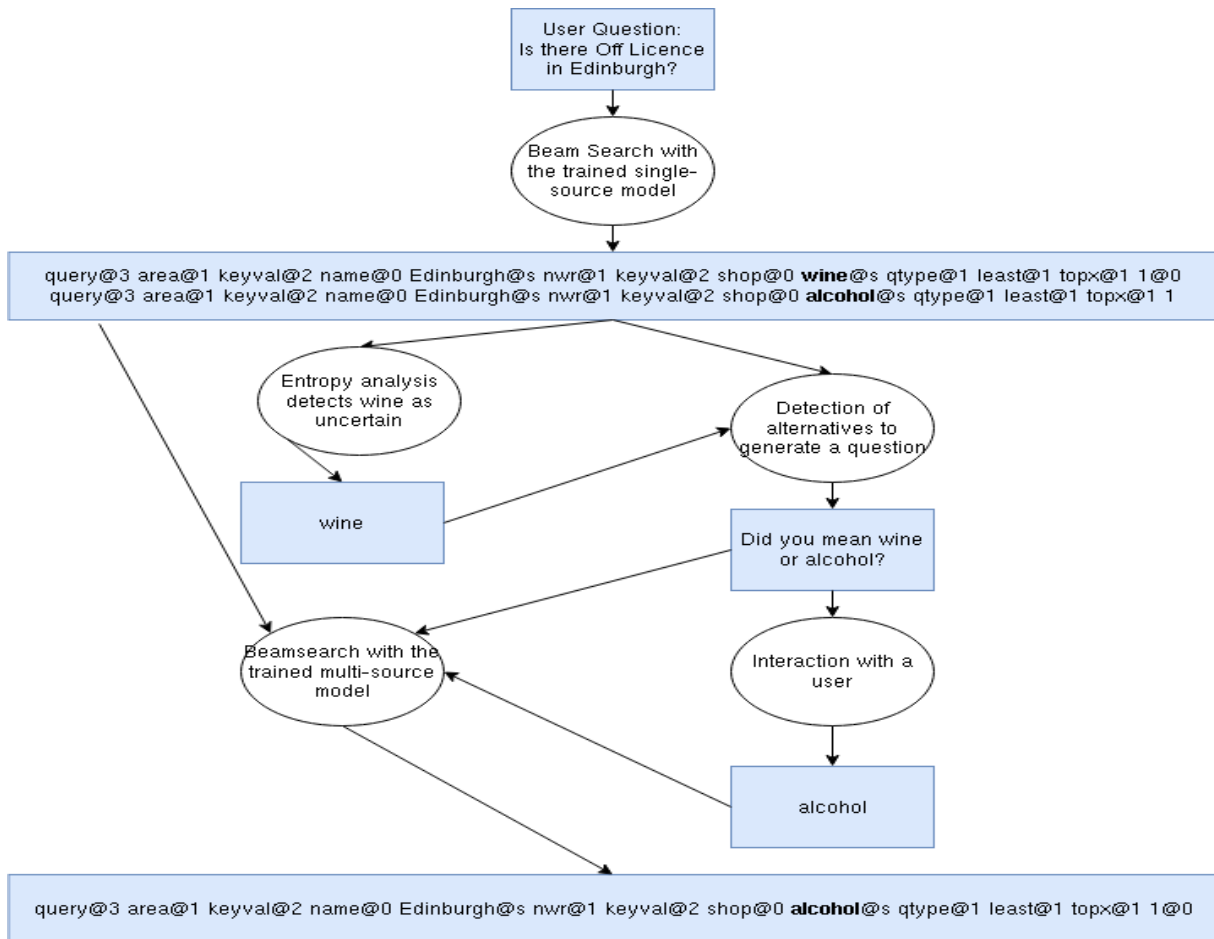
Figure 2: Workflow for the interaction process.

Riezler (2018). The task of the human users is to mark the errors in the list of keys and values, and to answer or correct the clarification question. The markings are used as feedback in the weighted fine-tuning objective of Petrushkov et al. (2018). As the outputs of the model are on character-level, the token-level reward of the annotations is distributed onto them for training. The final model is trained on the weighted objective in a multi-source fashion, taking parse hypothesis, clarification question, and logged user answer as additional inputs. Line 5 in Table 2 shows that fine-tuning a multi-source model that takes hypothesis, dialogue, and logged answer as additional input increases the sequence accuracy by another 0.15%. This difference is statistically significant with a p-value of 0.0027 determined by approximate randomization. The interaction process can be seen in Figure 2.[11]

8 Conclusion

Ambiguities or errors in real-world semantic OSM parsing arise because of different tagging preferences of developers and users, an issue that can only be solved by an interactive setup where a parser is aware of its errors, and a satisfactory answer is found by the user marking parse errors and communicating alternatives. Our current work is a first step towards precise communication and offline learning in interactive semantic parsing. An interesting future direction of work is to move to online learning in interactive semantic parsing.

---

[11]Additional experiments using the human annotations as test data are reported in the appendix.

57

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Ahmed Elgohary, Saghar Hosseini, and Ahmed Hassan Awadallah. 2020. Speak to your Parser: Interactive Text-to-SQL with Natural Language Feedback. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2065–2077.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A Minimalist NMT Toolkit for Novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 109–114.

Igor Labutov, Bishan Yang, and Tom Mitchell. 2018. Learning to Learn Semantic Parsers from Natural Language Supervision. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1676–1690.

Carolin Lawrence. 2018. *Response-Based and Counterfactual Learning for Sequence-to-Sequence Tasks in NLP*. Universitätsbibliothek Heidelberg, Heidelberg, Germany.

Carolin Lawrence and Stefan Riezler. 2018. Improving a Neural Semantic Parser by Counterfactual Learning from Human Bandit Feedback. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1820–1830.

Pavel Petrushkov, Shahram Khadivi, and Evgeny Matusov. 2018. Learning from chunk-based feedback in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Ilya Sutskever, Oriol Vinyals, and Quoc Le. 2014. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112. Curran Associates, Inc., Montreal, Canada.

Ziyu Yao, Xiujun Li, Jianfeng Gao, Brian M. Sadler, and Huan Sun. 2019. Interactive Semantic Parsing for If-Then Recipes via Hierarchical Reinforcement Learning. In *Proceedings of the 33rd Conference on Artificial Intelligence (AAAI)*, pages 2547–2554.

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 Lasecki, and Dragomir Radev. 2019a. CoSQL: A Conversational Text-to-SQL Challenge Towards Cross-Domain Natural Language Interfaces to Databases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, pages 1962–1979.

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*, pages 3911–3921.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Tan, 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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4511–4523.

Barret Zoph and Kevin Knight. 2016. Multi-Source Neural Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34.

A Supplementary Material for "Towards Error-Aware Interactive Semantic Parsing"

A.1 Hyperparameter Settings

A.2 Evaluation on the human annotated data

In an additional experiment, we evaluated the models that were trained on the synthetically generated dataset on the data resulting from the human interaction study. The result of comparing the baseline model with the multi-source model trained on parse hypothesis and synthetic dialogue as additional inputs is shown in Table 5. The astonishing gains of over 15% in F1 score can be explained by the fact that the data for human annotation set were filtered to include only examples for which the baseline parser did not match the gold standard parse (thus producing an accuracy score of 0).

System	Accuracy	F1
char	0	43.07
char+ hyps + dia	25.09	60.85

Table 5: Test results on human-annotated data.

A.3 Entropy visualization

The entropy of the parse of the sentence "How many Off License in Heidelberg" can be seen in Figure 4. The character-based model shows uncertainty with respect to the token *wine*. This is the desired result because the alternative for this position would be *alcohol*.

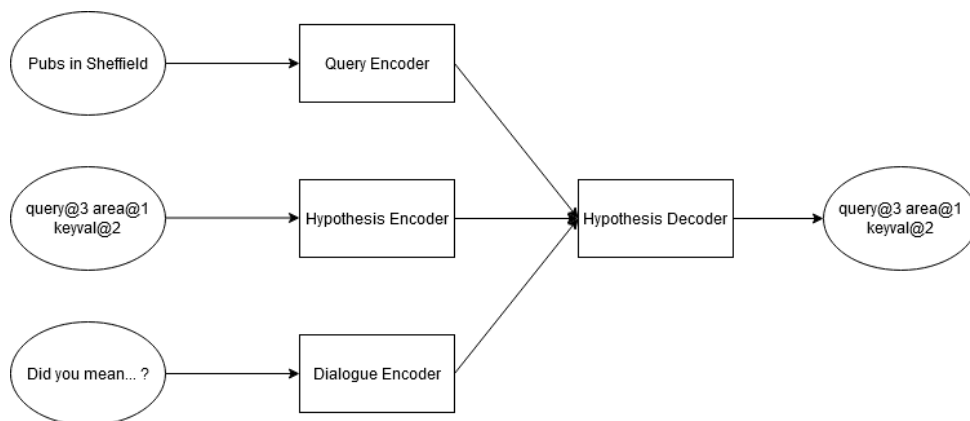| Parameter	Lawrence and Riezler (2018)	token-based	character-based
Attention mechanism	bahdanau	bahdanau	bahdanau
RNN type	gru	gru	gru
Embedding size	1000	620	620
Encoder layer count	1	1	1
Encoder hidden size	1024	400	400
Decoder layer count	1	1	1
Decoder hidden size	1024	800	800

Table 4: Parameter overview compared to Lawrence and Riezler (2018).

Figure 3: Multi-source semantic parsing.

Figure 4: Character entropy for parse of query "How many Off License in Heidelberg".