INSPIRED: A Large-Scale Dataset and Simulation Framework for Exploring Interactive Learning in Knowledge-Based Question Answering

Lingbo Mo, Ashley Lewis, Huan Sun, Michael White

The Ohio State University

Overview

Motivation

We aim to enhance transparency of the parsing process, increase user confidence in the final answer, and improve accessibility to knowledge bases for novice users. We design an interactive framework that demonstrates the question-answering process in a step-by-step fashion and allows the user to make corrections to individual steps in natural language.

Our Contributions

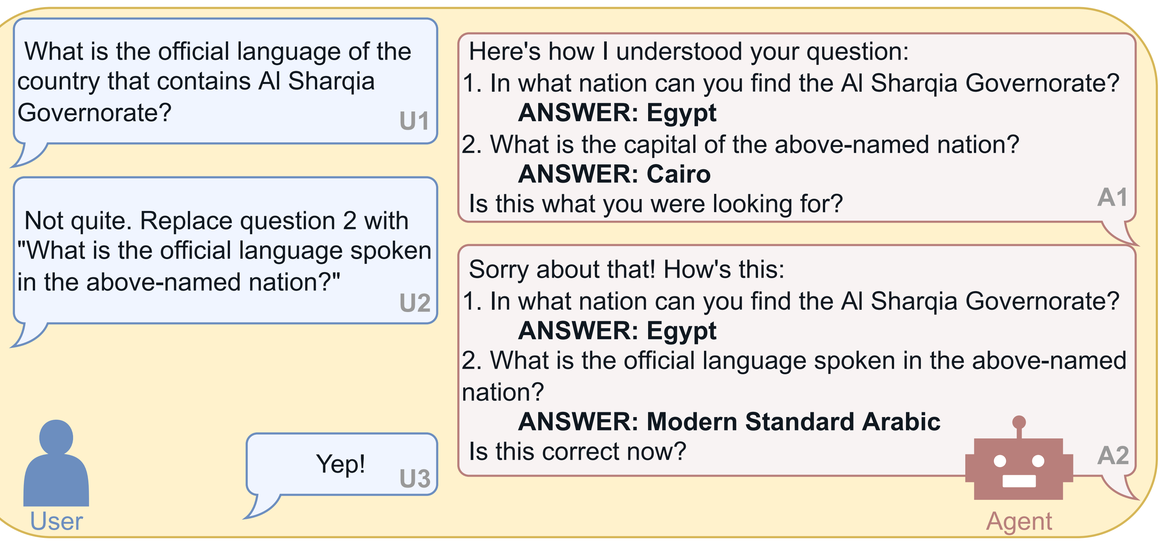
1. **A transparent interactive semantic parsing framework** that explains to a user how a complex question is answered step by step and allows them to make corrections in natural language.

2. **A high-quality dialogue dataset** using our framework to support research in interactive semantic parsing for KBQA.

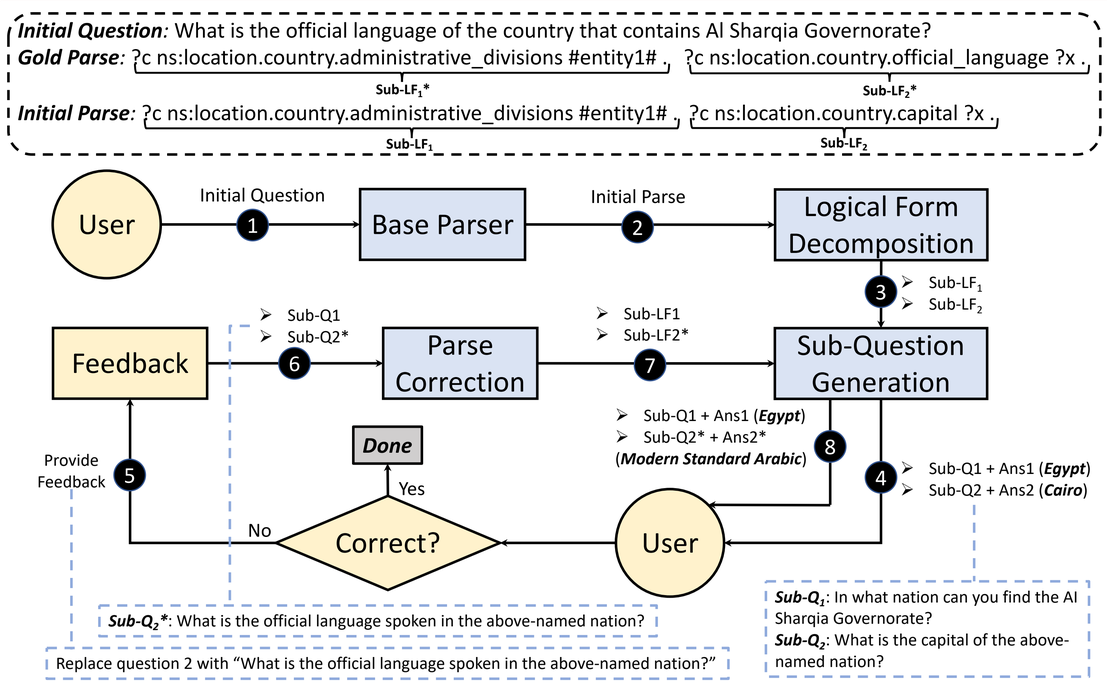
3. **Baseline models for two core subtasks**: Sub-Question Generation and Parse Correction.

4. **A simulation pipeline** to simulate user feedback, allowing us to study the promise of our framework to correct errors from other semantic parsers.

Scan code!



Interactive Semantic Parsing Framework

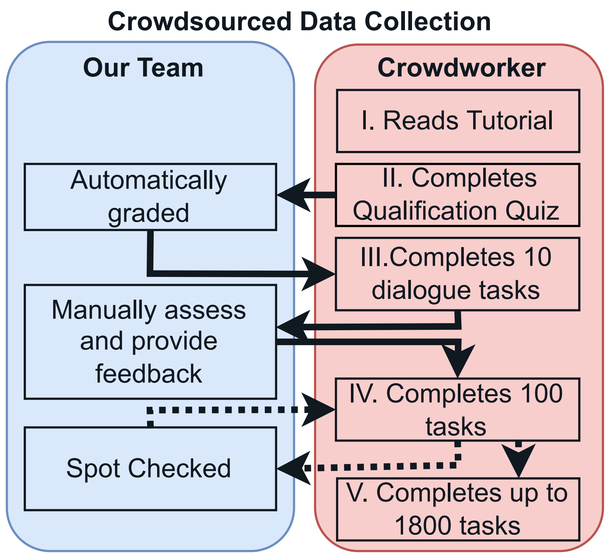


Interactive Semantic Parsing Framework for KBQA

We design an interactive framework that explains to the user how a complex question is answered **step by step** and enables them to make corrections in natural language, hence enhancing transparency and user trust. The main components in our framework include:

- **Logical Form Decomposition**: decompose the initial parse into sub-LFs.
- **Sub-Question Generation**: translate each sub-LF into a natural-language question.
- **Human Feedback**: human provides feedback in natural language to correct each step.
- **Parse Correction**: decode a new sub-LF given human feedback.

The INSPIRED Dataset



Data Quality

- Crowdworkers read a detailed tutorial, pass multiple rounds of qualification tests to ensure their proficiency at the task, and have their work spot checked in an ongoing fashion.
- We use a semi-automatic data cleaning method to identify inaccurate paraphrases for manual repair (see Appendix B.1 in the paper for details)
- Crowdworkers are encouraged to incorporate contextual information of a given sub-question into their rephrasings to enhance the contextual richness of the dataset (see Section 3 of the paper).

Dataset Statistics

Number of	Train	Dev	Test	Overall
Complex Questions	3,492	3,441	3,441	10,374
- Composition	1,196	1,532	1,490	4,218
- Conjunction	1,796	1,503	1,553	4,852
- Comparative	253	217	207	677
- Superlative	247	189	191	627
Predicted Sub-Questions	1.7	2.0	1.9	1.9
Gold Sub-Questions	2.2	2.1	2.1	2.1
Range of the number of predicted sub-questions				0 - 5
Range of the number of gold sub-questions				2 - 4
Average number of edits				1.4
Dialogues with 0 edits				5,016

The above table shows statistics of the INSPIRED dataset, including a breakdown of the questions of the reasoning types in the ComplexWebQuestion dataset (Talmor and Berant, 2018), from which our initial questions are drawn. We show the average number of sub-questions of the predicted and gold parses and the average number of edits per parse.

Experiments

Part 1: Parse Correction (Sub-Question => Sub-LF)

Context	Dialog-level EM Accuracy	Turn-1 (3441)	Turn-2 (3441)	Turn-3 (345)	Turn-4 (56)
		Without Feedback			
w/o Correction	52.3	-	-	-	-
		With Feedback			
BART-large					
w/o Context	71.3	84.6	81.5	85.5	53.6
+ h_q	72.2	84.7	82.2	89.3	**100.0**
+ h_{lf}	72.0	84.3	82.1	89.3	**100.0**
+ h_q & h_{lf}	**73.5**	**86.4**	**83.2**	**91.0**	**100.0**

Observations:

- Incorporating human feedback can substantially improve the parse accuracy.
- Adding contexts (h_{lf} and h_q denotes the dialogue history of sub-LFs and sub-questions respectively) into the input improves the correction accuracy.
- As the number of turns goes up, context contributes more to the correction process, which indicates that including the full dialogue history in the input leads to better results.

Part 2: Sub-Question Generation (Sub-LF => Sub-Question)

Context	BLEU-2	BLEU-4	BERTScore
BART-large^t			
w/o Context	32.4	16.2	94.2
+ h_{q^t}	33.3	16.5	94.6
+ Q	33.4	16.6	94.6
+ Q & h_{q^t}	**34.1**	**17.1**	**94.8**

Observations:

- Using templated sub-questions h_{q^t} in the model input improves generation performance.
- Incorporating the original complex question Q generally leads to better results.

Part 3: Simulation Experiments

	BART-large	QGG
EM	60.9	-
EM (After correction)	**75.1**	-
F1	65.8	49.0
F1 (After correction)	**75.7**	**56.5**

Attempt	EM	F1
BART-large		
1	75.1	75.7
2	78.7	79.9
3	**79.0**	**80.1**

A simulation pipeline to correct errors made by other semantic parsers.

- Translate the LFs predicted by other semantic parsers (e.g., BART-large and QGG for KBQA) into natural questions using the sub-question generation model.
- Use oracle error detection and train a generator to simulate a human user's feedback.
- Correct erroneous parses using our parse correction model. **Left table** shows the results before and after the correction process.
- **Right Table:** Expand to include multiple attempts of correction to simulate situations in which the model does not repair the parse correctly on the first attempt

Discussion and Future Work

Interactive Learning via Simulation Pipeline.

- Our simulation pipeline with the INSPIRED dataset provides an efficient means of *automatic evaluation* in an interactive scenario without further human efforts.
- Simulated interactions can re-train parse correction or base parsers continuously to explore the balance between the simulated data volume and performance gains.

Human in the Loop.

- We expect to generalize the framework to handle other unlabeled questions for parsing and involve human users to provide feedback for correction in a timely manner.
- Mispredicted parses recognized by the user can serve as negative instances to enhance the training process via contrastive learning.

Acknowledgements: This research was partially supported by a collaborative open science research agreement between Facebook and The Ohio State University.