Distribution Shifts Are Bottlenecks: Extensive Evaluation for Grounding Language Models to Knowledge Bases

Yiheng Shu, Zhiwei Yu

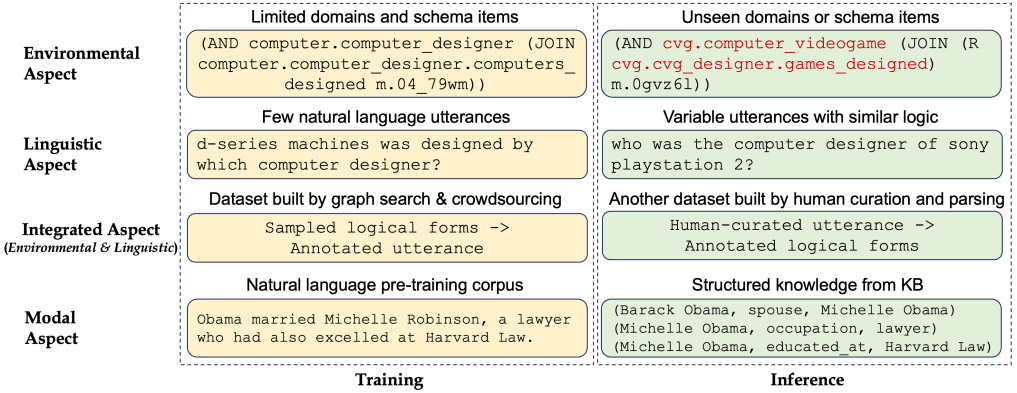
南京大学 NANJING UNIVERSITY Microsoft

Motivation

- Existing KBQA benchmarks may not fully represent the diverse scenario
  - KBs are enormous, structured and only partially observable (cannot be fully encoded by LMs)
  - Robustness concerns
- We aim to bridge this gap by
  - exploring the limitations of current KBQA **benchmarks**
  - proposing more comprehensive evaluation **protocols**

Challenges from Distribution Shifts

- Robustness is closely related to data distribution (Hendrycks et al., 2020)
- Training and inference using LMs face different distributions

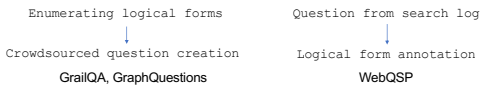


Environmental Aspect

- Schema-level generalization
  - The majority of KBQA benchmarks have i.i.d. schema generalization
- Evaluation Protocols (many **unseen** schema)
  - KBQA: GrailQA / GraphQuestions
  - Relation linking: SimpleQuestions-Balance

Integrated Aspect

- Unknown schema and linguistic distribution based on user query
  - Evaluating the pre-trained models on the unseen human-curated WebQSP dataset, where the questions are derived from search logs



Linguistic Aspect

- Adaptability to paraphrases
  - Natural language can be expressed in a variety of forms
  - **A new metric**, the standard deviation (**std**) of EM/F1 scores for questions of each logical form template.

$$Std_{F1} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(\frac{\sum_{j=1}^{m}(F1_{i,j} - \bar{F1}_i)^2}{m}\right)}$$

Modal Aspect

- In-context learning for KB modality using LLM without fine-tuning
  - LLMs are mainly trained with texts rather than KB contexts

Experiments

Augmentation Approach

- Data Augmentation for LMs
  - **G**raph se**a**rch and question generatio**n** **(GAIN)**
    - Graph search
    - Training question generator
    - Verbalization using question generator
    - Expanding training data
  - The sample size and schema distribution are extended
- Retrieval Augmentation for LLMs
  - Retrieving similar questions (k-shot)
  - Retrieving KB contexts for k samples and the input question
    - contexts: *entities*, *logical forms* and *schema items* relevant to the question

Setup

- Compared models
  - Models on GrailQA leaderboard
- TIARA (Shu et al., 2022) as the base model for GAIN
  - Due to its strong performance on zero-shot schema

Analyses

- Environmental Aspect
  - Effectiveness of synthesis and scaling up
  - Fine-tuning is better than few-shot learning in performance
- Linguistic Aspect
  - Improvements are linguistic biased
- Integrated Aspect
  - Difficult transfer across datasets
  - Causes from different data collection
- Modal Aspect
  - Context alone is insufficient
    - Notably, GPT often simply copies the logical forms in the retrieved contexts

Model on GrailQA Test Set	Overall EM	Overall F1	I.I.D. EM	I.I.D. F1	Compositional EM	Compositional F1	Zero-shot EM	Zero-shot F1
Fine-tuned Models								
BERT + Ranking (Gu et al., 2021)	50.6	58.0	59.9	67.0	45.5	53.9	48.6	55.7
RnG-KBQA (Ye et al., 2022)	68.8	74.4	86.2	89.0	63.8	71.2	63.0	69.2
TIARA (T5-base) (Shu et al., 2022)	73.0	78.5	87.8	90.6	69.2	76.5	68.0	73.9
DecAF (FiD-3B) (Yu et al., 2022)	68.4	78.8	84.8	89.9	73.4	81.8	58.6	72.3
Pangu (BERT-base) (Gu et al., 2022a)	73.7	79.9	82.6	87.1	74.9	81.2	69.1	76.1
Pangu (T5-large) (Gu et al., 2022a)	74.8	81.4	82.5	87.3	**75.2**	**82.2**	71.0	78.4
Pangu (T5-3B) (Gu et al., 2022a)	75.4	**81.7**	84.4	88.8	74.6	81.5	71.6	**78.5**
Codex-driven Models								
KB-BINDER (6)-R (Li et al., 2023)	53.2	58.5	72.5	77.4	51.8	58.3	45.0	49.9
Pangu (Codex) (Gu et al., 2022a)	56.4	65.0	67.5	73.7	58.2	64.9	50.7	61.1
GAIN-augmented Models								
TIARA + **GAIN** (T5-base)	75.1	80.6	88.3	91.0	73.0	79.6	69.9	76.4
TIARA + **GAIN** (T5-3B)	**76.3**	81.5	**88.5**	**91.2**	73.7	80.0	**71.8**	77.8
GPT-3.5-turbo (5-shot)	66.6	71.4	82.7	85.3	60.5	66.3	61.9	67.2

EM and F1 scores (%) on the hidden test set of GrailQA

Model on WebQSP	F1	Hits@1
TIARA♣ (T5-base) (Shu et al., 2022)	28.5	27.6
TIARA*♣ (T5-base) (Shu et al., 2022)	33.5	31.5
BERT + Ranking* (Gu et al., 2021)	**43.0**	-
TIARA + **GAIN** (T5-base)	29.1	28.2
TIARA + **GAIN** (T5-3B)	29.8	28.7
TIARA* + **GAIN** (T5-base)	33.9	31.8
TIARA* + **GAIN** (T5-3B)	34.5	32.3

F1 and Hits@1 scores (%) on WebQSP without fine-tuning on it; all models are trained on large-scale GrailQA; * denotes oracle entity annotations

Model on GraphQuestions	F1(↑)	Std(↓)
GraphQuestions on Freebase 2013-07		
UDepLambda (Reddy et al., 2017)	17.7	-
PARA4QA (Dong et al., 2017)	20.4	-
SPARQA (Sun et al., 2020)	21.5	-
BERT + Ranking (Gu et al., 2021)	25.0	-
ArcaneQA (Gu and Su, 2022)	31.8	-
TIARA♣ (T5-base) (Shu et al., 2022)	37.9	**0.141**
KB-BINDER (6) (Li et al., 2023)	39.5	-
TIARA + **GAIN** (T5-base)	45.5	0.153
TIARA + **GAIN** (T5-3B)	**48.7**	0.180
GraphQuestions on Freebase 2015-08-09		
BERT + Ranking (Gu et al., 2021)	27.0	-
ArcaneQA (Gu and Su, 2022)	34.3	-
TIARA♣ (T5-base) (Shu et al., 2022)	41.2	**0.157**
Pangu (Codex) (Gu et al., 2022a)	44.3	-
Pangu (T5-3B) (Gu et al., 2022a)	**62.2**	-
TIARA + **GAIN** (T5-base)	49.5	0.170
TIARA + **GAIN** (T5-3B)	53.0	0.200

F1 scores (%) and average std of F1 scores for each paraphrase set on the test set of GraphQuestions

Conclusion

- Call for further research into better **evaluation protocols** and enhancing the **robustness** of multiple aspects
- Results indicate that the existing methodologies for grounding LLMs are yet to prove their efficacy and superiority
- Future research issues include
  - collecting more balanced environment-specific corpora
  - improving the LLM learning paradigms
  - Our experiments show that the data augmentation techniques deserve further research.