Exploring the Reasons for Non-generalizability of KBQA systems

Sopan Khosla[+]
AWS AI Labs
sopankh@amazon.com

Ritam Dutt[*+]
Carnegie Mellon University
rdutt@andrew.cmu.edu

Vinayshekhar Bannihatti Kumar
AWS AI Labs
vinayshk@amazon.com

Rashmi Gangadharaiah
AWS AI Labs
rgangad@amazon.com

Abstract

Recent research has demonstrated impressive generalization capabilities of several Knowledge Base Question Answering (KBQA) models on the GrailQA dataset. We inspect whether these models can generalize to other datasets in a zero-shot setting. We notice a significant drop in performance and investigate the causes for the same. We observe that the models are dependent not only on the structural complexity of the questions, but also on the linguistic styles of framing a question. Specifically, the linguistic dimensions corresponding to explicitness, readability, coherence, and grammaticality have a significant impact on the performance of state-of-the-art KBQA models. Overall our results showcase the brittleness of such models and the need for creating generalizable systems.

1 Introduction

The task of Question Answering over Knowledge Bases (KBQA) involves answering a natural language question by querying a predefined knowledge base (KB). While progress in KBQA research has addressed several challenges like answering complex questions, multi-hop reasoning (Lan and Jiang, 2020; Ren et al., 2021), conversational QA (Kacupaj et al., 2021), and multi-lingual KBQA (Zhou et al., 2021), most of the prior work in this field has been restricted to an i.i.d. setting (Yih et al., 2016; Talmor and Berant, 2018a).

In a real-world setting, a KBQA system should be well-equipped to handle users' queries that were unseen during training. To motivate research along this front, Gu et al. (2021a) proposed a dataset (GrailQA) with an associated leaderboard to benchmark the generalizability of KBQA methods to new compositions, and unseen schema items (Zero-shot). Multiple state-of-the-art models (Ye et al., 2021; Yu et al., 2022; Gu and Su, 2022; Shu et al., 2022) have achieved remarkable performance on the Zero-shot split giving the impression that KBQA generalization might be a solved problem.

However, a cross-dataset evaluation of the models trained on GrailQA reveals that they do not transfer well even for the more simpler one or two-hop questions. We observe that while these models achieve impressive performance on the GrailQA Zero-shot (GrailQA Z) split, they fail to generalize to questions from other datasets like WebQSP (Yih et al., 2016), GraphQ (Su et al., 2016), and ComplexWebQuestions (Talmor and Berant, 2018b) even though they are built upon the same Knowledge Base (i.e. Freebase). In this work we closely inspect the reasons for this drop. We analyse the structural and linguistic differences between questions from the different publicly available KBQA benchmark datasets.

We observe that while structural complexity somewhat explains the performance variations across questions within the same dataset, it does not explain the performance drop when testing on other datasets. Our analysis shows that the linguistic differences like explicitness and length of questions, grammaticality, readability, and coherence account for the degradation in performance. Although WebQSP and GrailQA share the same underlying KB, the substantial differences in the annotation process manifests as samples having different linguistic properties. We find that these linguistic variations act as an additional dimension for evaluating the generalizability and real-world usefulness of KBQA systems.

2 Datasets

In order to understand the zero-shot efficacy of the state-of-the-art KBQA models, we look at their performance on the following datasets:

GrailQA (Gu et al., 2021b) contains questions across 86 domains and covers more than 3500 Free-

[*]Work conducted during an internship at Amazon.
[+] denotes equal contribution

The Fourth Workshop on Insights from Negative Results in NLP, pages 88–93

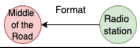
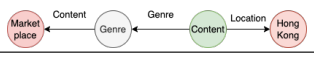
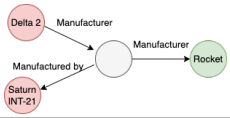
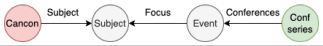
RP-Code	RP Instances	Question
RP-0		"what radio station uses the middle of the road format?"
RP-1		"what ship designer designed a ship that is designed by pete melvin?"
RP-2		"which powers do both catbus and rocky the flying squirrel have?"
RP-3		"genres of marketplace can be found in what broadcast content in hong kong?"
RP-4		"what other rocket did the manufacturer of saturn int-21 and delta 2 create?"
RP-5		"can-con has which conference series that focuses on it?"

Table 1: Example natural-language questions from GrailQA dev-set and their corresponding RP (relation path) categories. Red and green nodes in the graph correspond to the constraints (entities and literals), and the answer respectively.

base relations. It's development and test sets have three splits to independently measure the i.i.d, compositional and zero-shot capabilities of KBQA systems. We leverage their publicly available training and dev sets for our experiments.

WebQSP (Yih et al., 2016) contains question-answer pairs from non-experts collected using the Google Suggest API, and uses Amazon Mechanical Turk to get the answers for the obtained questions.

GraphQ (Su et al., 2016) has varying question characteristics that include complexity along the semantic structure, qualitative analysis over answer space, topic of the question, and the number of possible answers for the questions.

ComplexWebQuestions (CWQ) (Talmor and Berant, 2018a) builds on top of WebQSP and automatically creates complex questions that include phenomena such as function composition, conjunctions, superlatives and comparatives.

We consider these datasets for our experiments as all of them use Freebase as their underlying KB.

Creating zero-shot splits: We categorize questions in the test/dev splits of the corresponding dataset into (i) Non Zero-shot (I.I.D. + Compositional) and (ii) Zero-shot similar to the categories proposed by Gu et al. (2021a). Specifically, zero-shot instances have at least one schema item (class or relation) that were not seen during training in the original GrailQA dataset. We note the criteria to be a bit lenient for relations whose corresponding inverse relation occurred during training (ex: inventors.inventions as opposed to inventions.invented_by). Consequently, we update the zero-shot criterion to exclude questions where either the relation or it's corresponding inverse relation was observed during training.

	GrailQA		GraphQ		WebQSP		CWQ	
RP	All	Z	All	Z	All	Z	All	Z
RP-0	4950	2809	976	292	892	239	0	0
RP-1	1179	559	503	237	343	177	1188	602
RP-2	349	135	185	33	53	6	965	468
RP-3	128	18	70	31	14	3	1680	1347
RP-4	93	61	39	39	190	136	0	0
RP-5	62	22	33	33	1	0	856	608

Table 2: Data statistics. Distribution of different reasoning paths over the entire test/dev set (All) and the Zero-shot split (Z) for the different datasets.

Reasoning Paths: We characterize the complexity of the questions for different datasets based on the notion of reasoning paths as defined in Das et al. (2022). A reasoning path (hereforth RP) represents the sequence of actions (specifically relations traversed from the starting constraint(s) in the query graph) to reach the final answer. They provide a unified way to measure the complexity in terms of the number of hops and the number of constraints (examples shown in Table 1). Table 2 presents the most salient reasoning paths that occur in the dev split of the original GrailQA dataset and we thus restrict our analysis to these specific RPs on the other datasets. We further note the distribution of these RPs for the different datasets in Table 2.

3 Performance on Other KBQA Datasets

Experimental Setup: In this work, we explore the generalizability of four semantic-parsing based systems. These include (i) RNG-KBQA (Ye et al.,

	GrailQA				GraphQ				WebQSP				CWQ			
Model	EM	F1	EM(Z)	F1(Z)	EM	F1	EM(Z)	F1(Z)	EM	F1	EM(Z)	F1(Z)	EM	F1	EM(Z)	F1(Z)
RNG-KBQA	83.4	86.7	83.5	86.0	61.9	69.3	44.4	55.8	34.6	39.9	22.6	29.0	20.5	35.8	18.4	33.4
ArcaneQA	80.3	84.6	76.7	80.6	45.7	56.2	30.4	45.1	12.4	17.6	8.0	14.2	14.2	30.2	11.2	26.6
BERT-Ranker	66.7	72.2	69.6	74.4	43.9	50.1	32.3	40.1	35.7	43.9	25.0	37.1	13.3	28.3	10.3	25.0
BERT-Transducer	50.6	53.8	42.5	44.9	21.3	24.9	15.6	19.0	15.5	19.5	10.5	13.0	1.8	6.1	1.0	4.7

Table 3: EM and F1 scores for different KBQA baselines for the different KBQA datasets built on top of Freebase KB (with gold entities). Z refers to the Zero-shot subset.

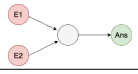
RP	RP-instance	GrailQA Z					GraphQ Z					WebQSP Z					CWQ Z				
		EM	F1	#Z	#W	#N	EM	F1	#Z	#W	#N	EM	F1	#Z	#W	#N	EM	F1	#Z	#W	#N
RP-0		87.1	88.0	2.9	10.6	4.3	41.8	53.8	2.0	8.9	2.9	31.4	38.3	2.0	6.8	2.1	-	-	-	-	-
RP-1		81.9	85.1	4.5	14.3	6.1	54.8	59.7	3.7	10.1	3.5	9.6	14.7	4.0	6.2	1.8	52.5	57.7	3.1	13.3	2.9
RP-2		74.8	86.2	4.7	15.7	6.2	63.6	87.9	5.0	12.3	3.3	0.0	38.3	2.6	8.2	2.6	25.0	45.6	3.2	12.5	2.3
RP-3		5.6	44.8	5.2	19.1	7.6	48.4	98.9	5.5	12.3	3.9	0.0	13.3	5.2	7.7	2.3	9.2	29.4	5.0	12.6	2.3
RP-4		9.8	47.6	7.0	13.0	3.6	17.9	32.7	4.9	12.9	4.5	25.7	31.1	5.3	7.2	2.6	-	-	-	-	-
RP-5		0.0	1.5	5.5	10.6	4.2	0.0	0.0	3.6	11.5	4.1	-	-	-	-	-	0.0	9.0	4.8	14.0	2.9

Table 4: EM and F1 scores for RNG-KBQA, and the mean # zero-shot items (#Z), # words (#W), # common nouns (#N) per question on the zero-shot splits of GrailQA, GraphQ, WebQSP, and CWQ.

2021), (ii) ArcaneQA (Gu and Su, 2022), (iii) BERT-Ranker (Gu et al., 2021a), and (iv) BERT-Transducer. We follow the exact inference setting mentioned in their Github repositories, and evaluate them in terms of EM and F1 scores. All experiments are carried out on a single RTX-1080Ti GPU with 12GB RAM. We use gold entities to control for the confounding caused by entity linking errors.

Overall Results: As shown in Table 3, both RnG-KGQA and ArcaneQA achieve F1 scores of more than 80% on GrailQA zero-shot split with gold entities. We observe that this comes from the near perfect performance on the simpler (RP-0,1,2) questions that make up more than 98% of GrailQA Z. BERT-Ranker also achieves a respectable F1 score of 74.4%, while BERT-Tranducer performs poorly with an F1 of 44.9%.

However, we observe that all models significantly suffer while transferring to other datasets. This is true for both zero-shot and non zero-shot splits, as the overall performance drops by more than half even for samples that do not contain any zero-shot schema items (Table 3). For the simpler 1-hop (RP-0) zero-shot questions, RnG-KGQA's F1 drops by more than 30% (Table 4). ArcaneQA, a seq2seq model, suffers even more. For 2-hop questions (RP-1), while RnG-KGQA scores a respectable 60% F1 on GraphQ Z, its performance on WebQSP Z is severely low (below 15% F1). Overall, we find that the state-of-the-art KBQA models trained on GrailQA are not able generalize to other datasets, despite the presence of gold entities, even though they are built on the same KB.

Number of zero-shot schema items (#Z): Previous works (Gu et al., 2021a; Ye et al., 2021) have shown a degradation in performance of KBQA systems when exposed to unseen schema items. We thus compare the number of zero-shot schema items in the questions across the datasets.

We observe that the zero-shot splits of the diferent datasets contain similar or fewer zero-shot schema items than GrailQA Z across the different reasoning paths (Table 4, 5). For example, the mean for WebQSP Z lies between 2 and 5 for the different RPs. Compare this with GrailQA Z, where this goes as high as 7 (RP-4). GraphQ Z is closer to GrailQA Z with an overall mean of 3.2, and with its bias towards more complex questions CWQ Z has a mean of 4.0 zero-shot items.

Controlling for RPs, none of the other datasets have significantly more zero-shot items than GrailQA Z, suggesting that these questions are not necessarily *more difficult*, and the non-generalizability of the evaluated systems cannot be solely attributed to this factor.

4 Analyzing Linguistic Variation

In this section, we explore whether the regression in performance can be explained via the linguistic variation among the different KBQA datasets. We analyze the questions in these datasets using the dimensions discussed below:

Dimension	GrailQA		GraphQ		WebQSP		CWQ	
	All	Z	All	Z	All	Z	All	Z
# Zero-shot items	1.87 ± 1.72	3.3 ± 0.97	1.46 ± 1.77	3.19 ± 1.67	1.65 ± 0.93	3.41 ± 0.76	2.95 ± 1.09	4.0 ± 0.72
# Words	10.96 ± 4.08	11.41 ± 3.58	9.35 ± 3.00	10.03 ± 2.94	6.64 ± 1.55	6.71 ± 1.61	13.19 ± 3.16	13.00 ± 3.12
# Common Nouns	4.32 ± 1.84	4.72 ± 1.75	3.22 ± 1.30	3.39 ± 1.30	2.12 ± 1.00	2.17 ± 1.00	2.6 ± 1.24	2.6 ± 1.25
Grammaticality	0.71 ± 0.45	0.7 ± 0.46	0.85 ± 0.36	0.83 ± 0.38	0.68 ± 0.47	0.73 ± 0.44	0.78 ± 0.41	0.75 ± 0.43
Complexity	0.02 ± 0.13	0.01 ± 0.11	0.01 ± 0.07	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.05	0.01 ± 0.08
Coherency	-5.96 ± 0.90	-5.99 ± 0.90	-5.54 ± 1.00	-5.54 ± 1.00	-5.7 ± 1.00	-5.65 ± 1.00	-4.96 ± 0.92	-5.04 ± 0.94
Formality	0.14 ± 0.24	0.16 ± 0.26	0.12 ± 0.23	0.13 ± 0.25	0.01 ± 0.02	0.01 ± 0.02	0.99 ± 0.08	0.99 ± 0.09
Readability	65.34 ± 30.91	60.46 ± 26.85	66.46 ± 31.4	71.85 ± 25.71	79.75 ± 23.89	77.23 ± 25.19	74.03 ± 22.57	71.57 ± 21.60

Table 5: Mean and std. dev. scores for All and Zero-shot (Z) questions across different KBQA datasets on the various analysis dimensions.

Sentence Length (#W): Firstly, we compare the length of the natural language questions in each dataset. We find that WebQSP seems to have the shortest questions (Table 4, 5). WebQSP questions consistently contain 6-8 words regardless of the complexity of the reasoning path. Compare this to GrailQA that contains more than double of that (19 words) in its RP-3 questions. Furthermore, CWQ that was built by combining different WebQSP samples also contains longer question statements.

Common Nouns (#N): We also investigate the frequency of common nouns across the dataset questions. We use NLTK's POS-tagger and consider words corresponding to "NN" and "NNS" tags as common nouns. We compute the mean distrbution of common nouns (#N) across the datasets.

We observe that #N is twice as large in GrailQA compared to WebQSP and CWQ (Table 4, 5). While this phenomenon is seen for very simple questions (RP-0,2), it is magnified more for questions with hidden nodes (RP-1,3). We attribute this difference to the explicit language used in GrailQA, where hidden classes in the graph query are also sometimes mentioned in the question statement.

Grammaticality & Complexity: Linjordet and Balog (2022) demonstrates a significant drop in performance of KBQA models in presence of more natural questions. The authors measure "naturalness" of questions along the lines of grammaticality, fluency, and complexity. We thus investigate whether the different datasets are similar in distribution along these aforementioned dimensions.

We use the BLIMP (Warstadt et al., 2020) and COLA corpora (Warstadt et al., 2019) to fine-tune a BERT-base-uncased model to detect grammaticality. We observe high scores for WebQSP and CWQ and low for GraphQ and GrailQA which ties in with previous findings. We also analyse whether the questions in the different datasets have varying degrees of complexity, for which we use the dataset of Iavarone et al. (2021). We observe that none of the four datasets are very complex, with GrailQA All achieving the highest mean score of 0.02.

Readability: We use the Flesch-reading score to characterize how easy it is to comprehend a given question in each of these datasets. We observe that GraphQ has a very similar score to GrailQA in that they are less readable, whereas WebQSP and CWQ have much higher readability (Table 5).

Formality: To quantify the formality in the writing style, we pass the questions through a RoBERTa based classifier trained on GYAFC and take the softmax outputs as the formality score. We find that WebQSP questions have the least mean formality (0.01) while CWQ questions have the highest (0.99). GrailQA and GraphQ questions are also on the informal side (Table 5).

Coherency: To measure the differences in the coherency, we use a reference free metric called CTR-LEval (Ke et al., 2022). We observe that GrailQA is not as coherent as WebQSP (Table 5). We hypothesize this to be the case because of the mention of the hidden classes in GrailQA question statements. On the other hand, WebQSP questions are more natural as they are scrapped from the Google Suggest API. We also observe that both CWQ and GraphQ have much higher coherency scores when compared to both GrailQA and WebQSP.

5 Discussion

Overall, our results show that systems trained on GrailQA seem to transfer the best to GraphQ which has similar linguistic properties to GrailQA i.e., higher sentence lengths and number of common nouns, medium formality scores, and lower readablity. This is inline with the similarity in their

https://huggingface.co/s-nlp/roberta-base-formality-ranker

annotation processes that requires annotators to refer to a query graph to arrive at a NL question, which might bias them to include hidden nodes in the reasoning path. The questions in GrailQA are more explicit (highest #N) than GraphQ.

Compare this with the extremely poor performance on WebQSP, which can be explained by the stark differences in the language used in this dataset i.e., lesser (i) number of words in question sentences, (ii) number of common nouns and (iii) formality, and higher readability. This follows from WebQSP containing real-world non-expert queries collected from a search engine.

Finally, despite CWQ having longer questions like GrailQA, it does not contain as many #N suggesting that the annotators do not rely on introducing hidden classes in the NL question while merging the simpler WebQSP questions. Higher formality, readability, and coherency scores for CWQ show that the paraphrasing step used by the authors creates more *natural* and *readable* questions, as compared to GrailQA. We believe that these linguistic differences atleast partially explain the drop in performance for models when tested on CWQ.

We posit that the higher explicitness of GrailQA questions might provide some additional signal to KBQA systems during training that helps them in deciding the best relations/ nodes among the possible options. Systems' over-reliance on this signal might not transfer well to other datasets (as shown in this work) thus rendering them less useful.

6 Conclusion

Recent KBQA systems have demonstrated impressive performance on the GrailQA leaderboard that evaluates them for their zero-shot genearlizability. In this work, we show that these systems that are trained on GrailQA do not transfer to other KBQA datasets built on top of the same KB. Our analysis shows that despite controlling for structural complexity of the questions, there is a drop in performance across datasets. We observe that this can be explained by the difference in annotation processes and the resulting variations in the linguistic properties of these questions. Our work showcases that linguistic variation is an important dimension for evaluating the generalizability of KBQA systems in real-world scenarios.

References

Rajarshi Das, Ameya Godbole, Ankita Naik, Elliot Tower, Manzil Zaheer, Hannaneh Hajishirzi, Robin Jia, and Andrew Mccallum. 2022. Knowledge base question answering by case-based reasoning over subgraphs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4777–4793. PMLR.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021a. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021b. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488.

Yu Gu and Yu Su. 2022. Arcaneqa: Dynamic program induction and contextualized encoding for knowledge base question answering. *arXiv preprint arXiv:2204.08109*.

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence complexity in context. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 186–199, Online. Association for Computational Linguistics.

Endri Kacupaj, Joan Plepi, Kuldeep Singh, Harsh Thakkar, Jens Lehmann, and Maria Maleshkova. 2021. Conversational question answering over knowledge graphs with transformer and graph attention networks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 850–862, Online. Association for Computational Linguistics.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. Ctrleval: An unsupervised reference-free metric for evaluating controlled text generation. *arXiv preprint arXiv:2204.00862*.

Yunshi Lan and Jing Jiang. 2020. Query graph generation for answering multi-hop complex questions from knowledge bases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 969–974, Online. Association for Computational Linguistics.

Trond Linjordet and Krisztian Balog. 2022. Would you ask it that way? measuring and improving question naturalness for knowledge graph question answering. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3090–3098, New York, NY, USA. Association for Computing Machinery.

Hongyu Ren, Hanjun Dai, Bo Dai, Xinyun Chen, Michihiro Yasunaga, Haitian Sun, Dale Schuurmans, Jure Leskovec, and Denny Zhou. 2021. Lego: Latent execution-guided reasoning for multi-hop question answering on knowledge graphs. In *International Conference on Machine Learning*, pages 8959–8970. PMLR.

Yiheng Shu, Zhiwei Yu, Yuhan Li, Börje F Karlsson, Tingting Ma, Yuzhong Qu, and Chin-Yew Lin. 2022. Tiara: Multi-grained retrieval for robust question answering over large knowledge bases. *arXiv preprint arXiv:2210.12925*.

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gür, Zenghui Yan, and Xifeng Yan. 2016. On generating characteristic-rich question sets for qa evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 562–572.

Alon Talmor and Jonathan Berant. 2018a. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018b. The web as a knowledge-base for answering complex questions. In *North American Chapter of the Association for Computational Linguistics*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. 2021. Rng-kbqa: Generation augmented iterative ranking for knowledge base question answering. *arXiv preprint arXiv:2109.08678*.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, Berlin, Germany. Association for Computational Linguistics.

Donghan Yu, Sheng Zhang, Patrick Ng, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Yiqun Hu, William Wang, Zhiguo Wang, and Bing Xiang. 2022. Decaf: Joint decoding of answers and logical forms for question answering over knowledge bases. *arXiv preprint arXiv:2210.00063*.

Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. 2021. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834, Online. Association for Computational Linguistics.