LREC-COLING 2024

ELRA ICCL

RENMIN UNIVERSITY OF CHINA

Diversifying Question Generation over Knowledge Base via External Natural Questions

Shasha Guo[1, 2], Jing Zhang[1, 2], Xirui Ke[1, 2],

Cuiping Li[1, 2], and Hong Chen[1, 2]

[1] School of Information, Renmin University of China
[2] Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education, RUC

Problem Definition

- Generate questions based on **a set of formatted facts** from a knowledge base.

Example

Input: Triples & Answer(underlined)

<Aruba, capital, Oranjestad>,
<Aruba, currency_used, Aruban florin>

Output: Question

What is the name of the money used in the country where Oranjestad is the capital ?

Current Challenge

- **Limited diversity** in generated questions with traditional models.

Example

- Given a set of triples with the answer(underlined), each method returns top-3 questions, where the various surface forms are marked in **different colors**.

Triples:
<Aruba, capital, Oranjestad>, <Aruba, currency_used, Aruban florin>
Answer:
Aruban florin
Ground truth:
What is the name of the money used in the country where Oranjestad is the capital?

Questions generated by BART:
Q1: What type of money is used where Oranjestad is the capital?
Q2: The country with the capital of Oranjestad uses what type of money?
Q3: The country with the capital of Oranjestad uses what type of money?

Questions generated by BART+Paraphrase:
Q1: What kind of money is used by the country 's capital, Oranjestad?
Q2: What currency is used in the country with Oranjestad as its capital?
Q3: What currency is used in the country with Oranjestad as its capital?

Questions generated by ours:
Q1: What type of money is used in the country with Oranjestad as its capital?
Q2: The country with the capital of Oranjestad uses what type of money?
Q3: What currency is used in the country with capital of Oranjestad?

Reevaluating Diversity Metrics

- Limitations of Current Metrics

Focus on n-gram uniqueness, which may not accurately reflect true semantic diversity.

- Proposed Diversity Metric

***Diverse@k*:**

$$Diverse@k = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Diverse(S_i, S_j),$$

$$Diverse(S_i, S_j) = \frac{|\mathcal{T}_i - \mathcal{T}_j| + |\mathcal{T}_j - \mathcal{T}_i|}{|\mathcal{T}_i \cup \mathcal{T}_j|},$$

$$R(S_i, S) \geq \alpha \text{ and } R(S_j, S) \geq \alpha$$

A novel metric that assesses diversity **across top-k generated questions** while **ensuring relevance**.

Pilot Study

- Preliminary Experiment

Use an advanced paraphrase model to paraphrase ground truth questions.

- Experiment Results

Model	Diverse@10
BART	21.50
BART+Paraphrase	**29.05**
Gain	**7.55**

- Questions generated by injecting paraphrased ground truth are more diverse than those generated solely from the original ground truth.
- This results indicate that paraphrasing has a positive effect on enhancing diversifying question generation.

Our Approach

- Main Idea

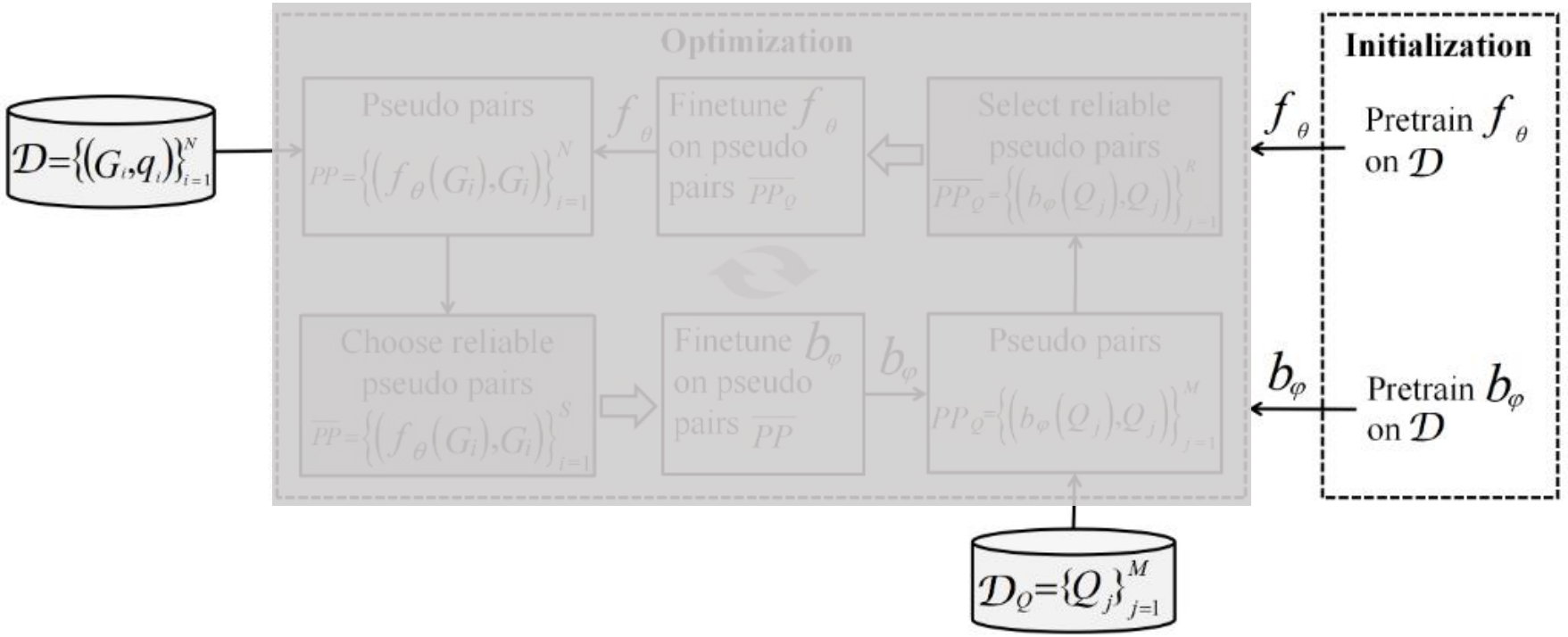
Employ external natural question to diversify question generation.

- Overview

Forward model aims to generate questions according to given a set of triples.

Backward model aims to help the forward model capture diverse question expressions.

Our Approach



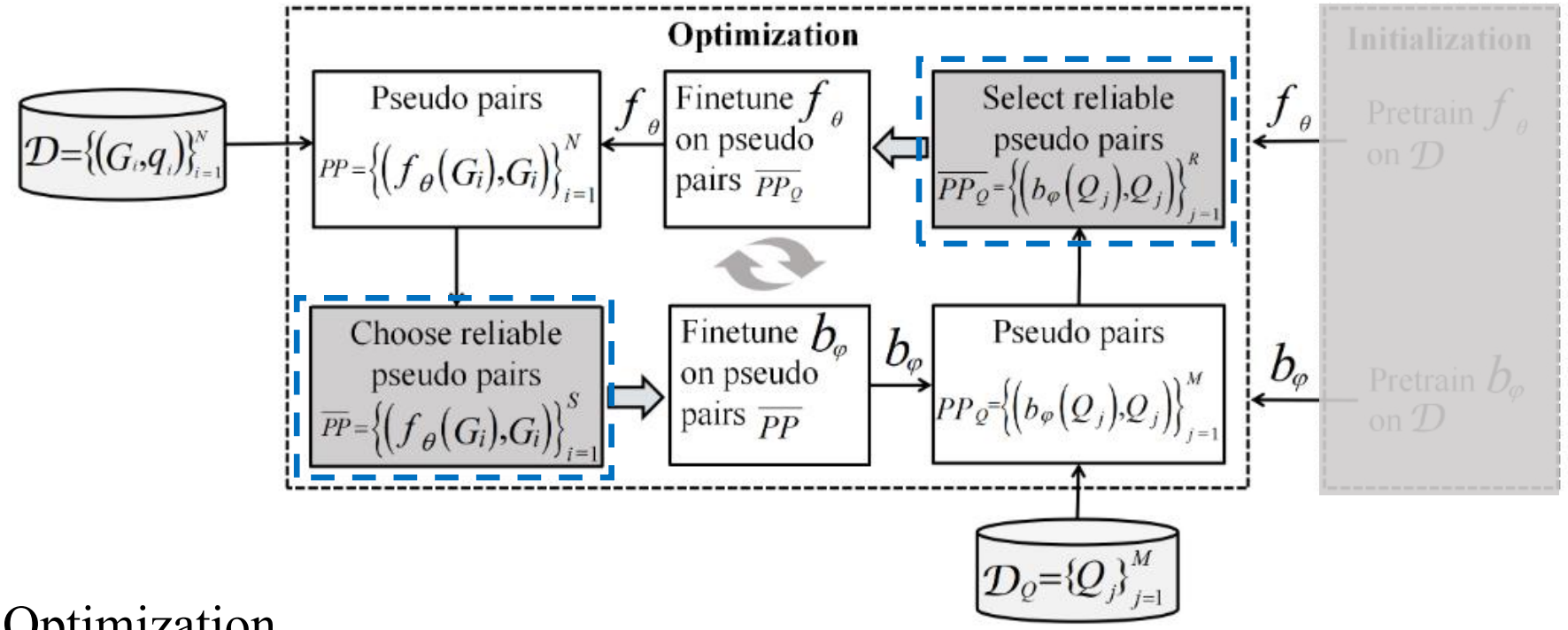
- Step 1: Initialization

  Use training data *D* to pre-train dual models.

Forward Model : $\mathcal{L}_f^{(0)} = \max_\theta \sum_{i=1}^{N} \log P_\theta\,(q_i|G_i)$

Backward Model : $\mathcal{L}_b^{(0)} = \max_\varphi \sum_{i=1}^{N} \log P_\varphi\,(G_i|q_i)$

Our Approach



- Step 2: Optimization

  Use **external questions D_Q** to finetune forward model and **training data D** to finetune backward model.

  ✓ Select reliable pairs based on the semantic similarity between $f_\theta(b_\varphi(Q_j))$ and Q_j

  $$\mathcal{L}_f = \max_\theta \sum_{j=1}^{M} \log P_\theta\left(Q_j | b_\varphi(Q_j)\right)$$

  ✓ Select reliable pairs based on their semantic relevance and diversity between $f_\theta(G_i)$ and q_i

  $$\mathcal{L}_b = \max_\varphi \sum_{i=1}^{N} \log P_\varphi\left(G_i | f_\theta(G_i)\right)$$

Experiments

- Overall Evaluation

Model	Top-3 Questions				Top-5 Questions				Top-10 Questions			
	simCSE	BLEU-1	Diverse@3	Dist-1	simCSE	BLEU-1	Diverse@5	Dist-1	simCSE	BLEU-1	Diverse@10	Dist-1
T5	87.04	52.35	22.50	34.67	86.75	52.30	25.57	23.67	86.16	52.47	29.80	14.67
BART	94.07	78.76	18.32	33.97	**93.37**	**77.76**	20.70	22.55	**92.06**	**76.21**	24.53	13.84
JointGT	**94.11**	**78.93**	18.66	33.96	93.28	77.74	21.26	22.69	91.99	76.05	25.22	14.01
T5+P	85.58	48.61	24.81	36.82	85.44	49.50	27.66	25.36	85.24	50.41	31.10	15.37
B+P	89.97	68.51	24.58	37.96	89.92	68.75	26.39	25.12	89.55	68.47	28.88	14.79
JointGT+P	90.09	68.89	24.44	38.18	89.97	68.75	26.76	25.28	89.56	68.30	28.91	14.66
Davinci003	77.06	39.33	28.14	38.95	76.86	39.32	30.18	27.27	76.94	39.38	31.46	16.97
ChatGPT	77.17	33.58	**29.87**	39.59	77.18	33.59	**32.04**	**28.68**	77.25	33.75	34.38	**18.04**
Ours	85.63	59.71	28.12	**40.62**	85.06	59.17	31.60	28.39	84.40	58.46	**35.85**	17.75

Table 2: Overall evaluation on PQ (%).

Model	Top-3 Questions				Top-5 Questions				Top-10 Questions			
	simCSE	BLEU-1	Diverse@3	Dist-1	simCSE	BLEU-1	Diverse@5	Dist-1	simCSE	BLEU-1	Diverse@10	Dist-1
T5	75.80	42.11	21.36	41.19	75.88	42.51	24.14	28.21	75.83	42.85	28.02	17.13
BART	82.42	51.64	16.88	42.46	82.30	51.59	18.87	29.50	82.02	51.18	21.50	18.22
JointGT	**82.64**	**52.01**	16.37	41.92	**82.52**	**51.90**	18.27	29.14	**82.24**	**51.61**	20.65	17.82
T5+P	77.97	42.76	22.42	42.27	78.04	43.10	25.53	29.26	78.04	43.48	29.78	17.92
B+P	81.24	48.77	22.97	41.31	81.16	48.74	25.63	28.30	81.03	48.74	29.05	17.06
JointGT+P	81.09	48.28	23.71	41.56	81.10	48.37	26.17	28.41	80.88	48.27	29.85	17.28
Davinci003	71.68	33.62	24.32	42.95	71.75	33.75	26.51	30.50	71.68	33.61	29.21	19.10
ChatGPT	74.54	33.21	**28.88**	**42.96**	74.38	33.14	**31.38**	**30.82**	74.40	33.09	**33.81**	**19.47**
Ours	80.58	49.95	25.17	42.52	80.55	49.94	28.05	29.57	80.31	49.85	31.33	18.27

Table 3: Overall evaluation on WQ (%).

- Injecting paraphrased questions can contribute to diversifying KBQG.
- Our approach surpasses PLMs-based baselines in diversity, which demonstrates the effectiveness of leveraging external natural questions.
- Our approach achieves comparable performance to PLMs-based baselines in relevance and surpasses LLMs-based baselines.

Experiments

- Positive Impact on QA Tasks

  ✓ Setting

  WebQSP is augmented by the questions generated by B+P and our proposed model, which are denoted as "**Augment by B+P**" and "**Augment by Ours**" respectively.

Model	GRAFT-Net		NSM	
	Hits@1	F1	Hits@1	F1
Real	0.677	0.616	0.724	0.663
Augment by B+P	0.676	0.624	0.732	0.673
Augment by Ours	**0.688**	**0.629**	**0.739**	**0.681**

Table 5: QA performance on the augmented QA dataset.

  ✓ Results

  - The generated (question, answer) pairs can be viewed as a method of data augmentation for KBQA.
  - Our model generates questions that significantly outperform those generated by B+P.

Experiments

- Human Evaluation

  ✓ Setting

  We randomly choose 50 instances from the test set of the WQ dataset and then evaluate whether top-3 and top-5 generated questions for each instance using a five-point Likert scale.

Model	Top-3 Questions		Top-5 Questions	
	Diversity	Relevance	Diversity	Relevance
BART	3.45	**4.25**	3.56	**4.18**
B+P	3.67	4.05	3.85	4.02
Ours	**3.98**	3.96	**4.21**	3.89
Pearson	0.935	-	0.949	-

Table 6: Human evaluation results on WQ.

  ✓ Results

  - Our approach produces more diverse questions than other baselines while achieving comparable performance in terms of relevance with the baselines.

  - The value of Pearson correlation coefficient is greater than 0.9, which demonstrates that our devised metric, *Diverse@k,* is highly consistent with human evaluation.

Conclusion

- Contributions

  ✓ We are the first to propose the diversity among top-k generated questions for each instance, ensuring their relevance to the ground truth.

  ✓ We design a novel metric called Diverse@k to measure the diversity.

  ✓ We introduce a dual model framework with two selection strategies that incorporate diverse expressions from external questions into the generation model.

- Future Directions

  ✓ Explore further enhancements in diversifying question generation by addressing both word-level and structure-level diversity.

LREC-COLING 2024

ELRA Language Resources Association

ICCL International Committee on Computational Linguistics

RENMIN UNIVERSITY OF CHINA 中国人民大学

Thank you!

Q&A