

# Knowledge-enriched, Type-constrained and Grammar-guided Question Generation over Knowledge Bases

Sheng Bi<sup>1</sup> and Xiya Cheng<sup>1</sup> and Yuan-Fang Li<sup>2</sup> and Yongzhen  
Wang<sup>3</sup> and Guilin Qi<sup>1\*</sup>

COLING 2020

# Motivation

- 传统的KBQG方法都是基于编码器-解码器模型实现的，但是这些方法目前仍然存在两个主要的挑战，特别是在较小的子图上
  - 由于子图中只有几个三元组，包含的信息有限，导致生成的问题不够流畅，并且单一、缺乏多样性
  - 由于解码器忽略了答案实体的语义信息，导致语义漂移。

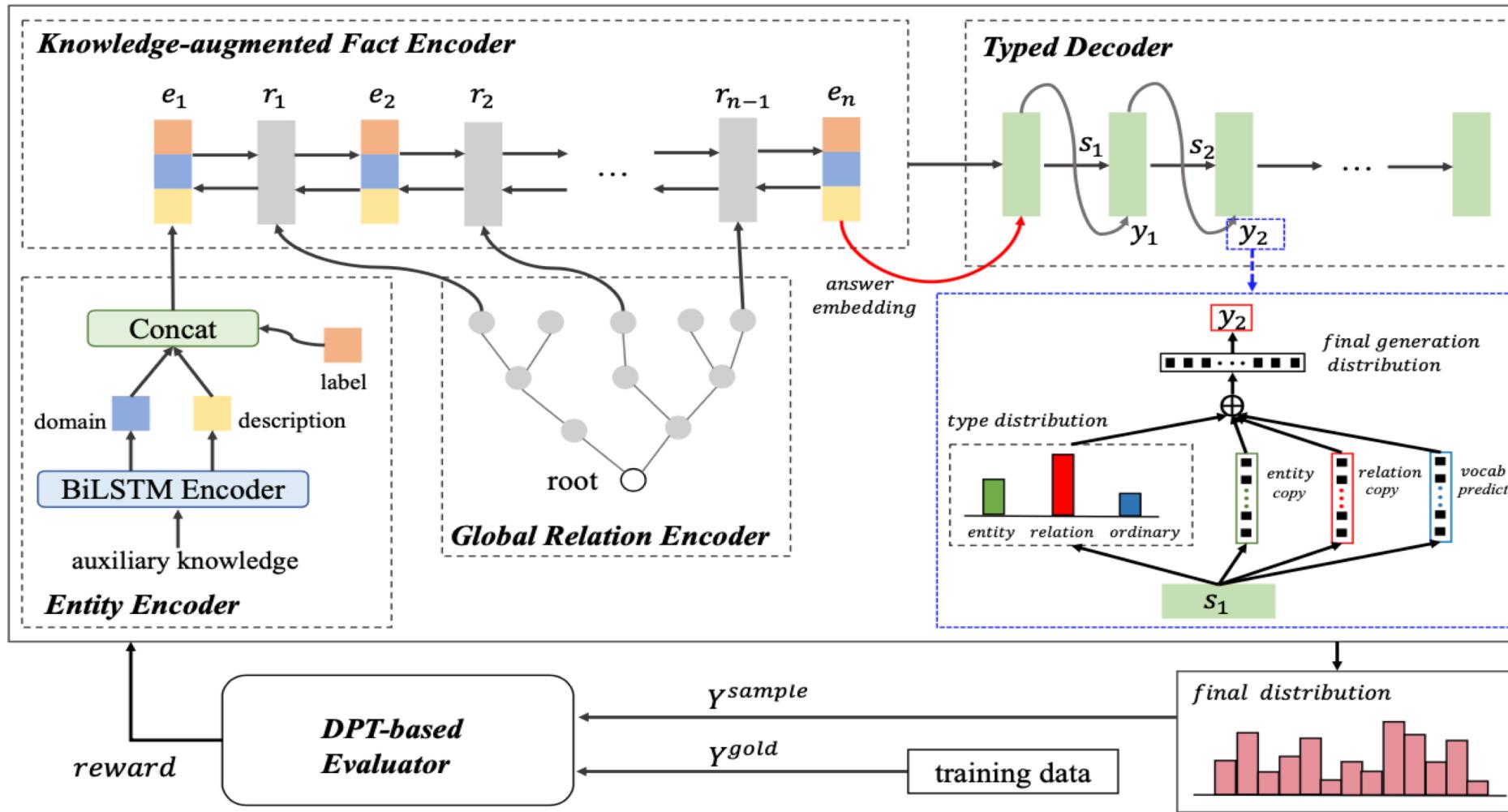
Input	<LeBron James, educated_at, St. Vincent-St. Mary High School><St. Vincent-St. Mary High School, located_in, <i>Ohio</i> >		
Output	Diverse Question	Correct Question	Question
Q1	✗	✓	<b>Where</b> was LeBron James's high school located in?
Q2	✓	✗	<b>When</b> was the high school of LeBron James, <i>the American basketball player</i> , located in?
Q3	✓	✓	<b>Where</b> was the high school of LeBron James, <i>the American basketball player</i> , located in?

Table 1: An example of KBQG. We aim at generating questions like Q3, which is diverse and correct. Compared with Q3, Q1 has less diversity and Q2 suffers from the semantic drift problem.

# Contribution

- 我们用辅助信息扩充源子图以丰富编码器输入，从而提高生成问题的多样性。
- 我们建议在生成的问题中加入单词类型，并使解码器输出以这些类型为条件，从而缓解语义漂移问题。
- 在强化学习框架中，我们设计了一个基于 DPT 的评估器，以鼓励结构一致性，同时不严格执行子序列匹配。
- 我们对两个基准数据集进行了广泛的实验，结果表明，我们的模型大大优于最先进的方法，并且它可以产生更正确、多样化和流畅的问题。

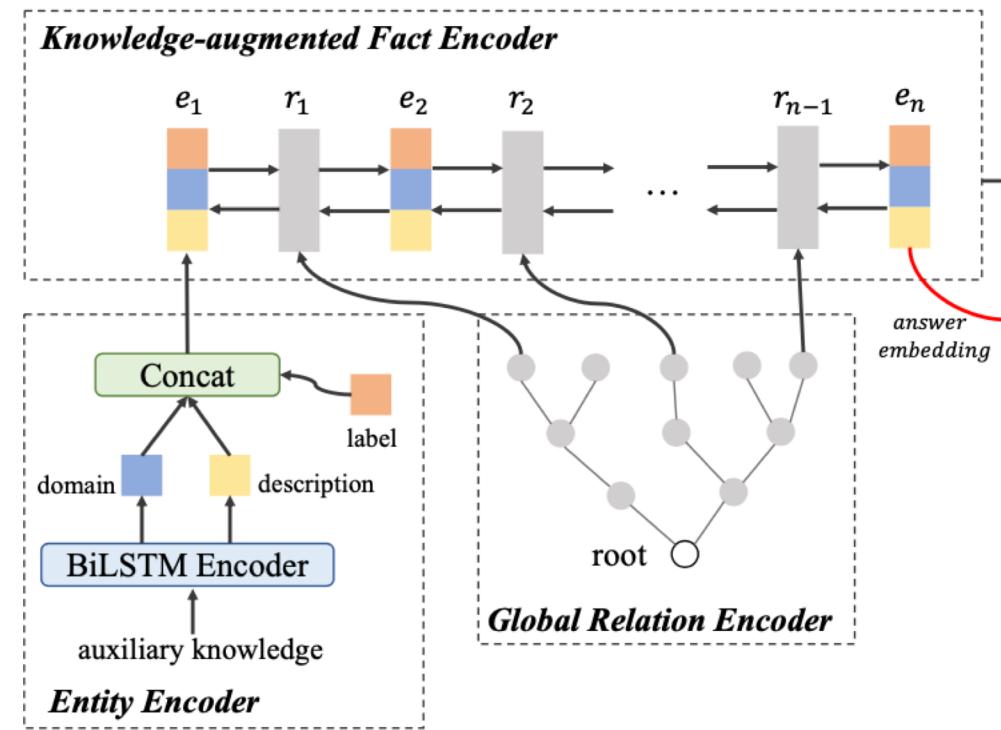
# Overview



# Method

## ➤ Knowledge-augmented Fact Encoder

- 对于实体，引入知识库信息中实体的领域、描述等额外信息进行编码，加上自身的 embedding 信息，三部分拼接。
- 对于关系，发现数据中不同实体之间的关系具有层级关系，借助 Tree-LSTM 对关系进行编码，而不同于以往的工作，并没有深入挖掘不同关系之间的联系。



由实体编码器、关系记忆和知识增强事实编码器组成。

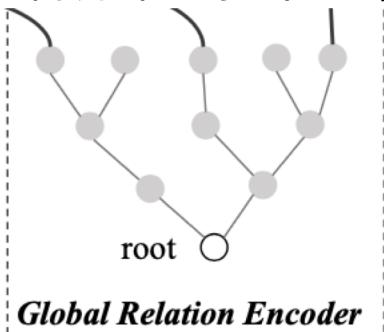
# Method

- Entity Encode

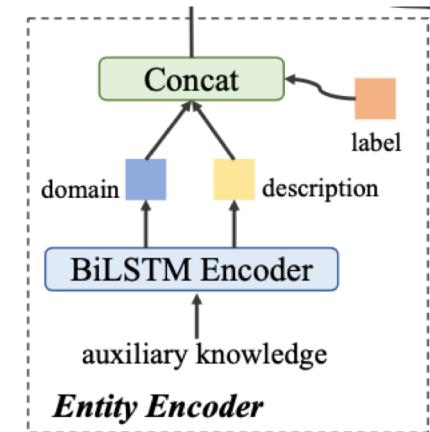
在本文中，我们将每个实体链接到其各自的 Wikidata 页面，并获得相应的辅助知识，包括简要描述和域定义，以丰富源输入。利用标签、描述和域信息来表示每个实体，并且利用一个双层双向LSTM对这些实体进行编码

- Global Relation Encoder

全局关系编码器通过 Tree-LSTM 利用这种分层结构来编码这些关系。关系编码器中的每个 LSTM 单元都能够合并来自多个子单元的信息



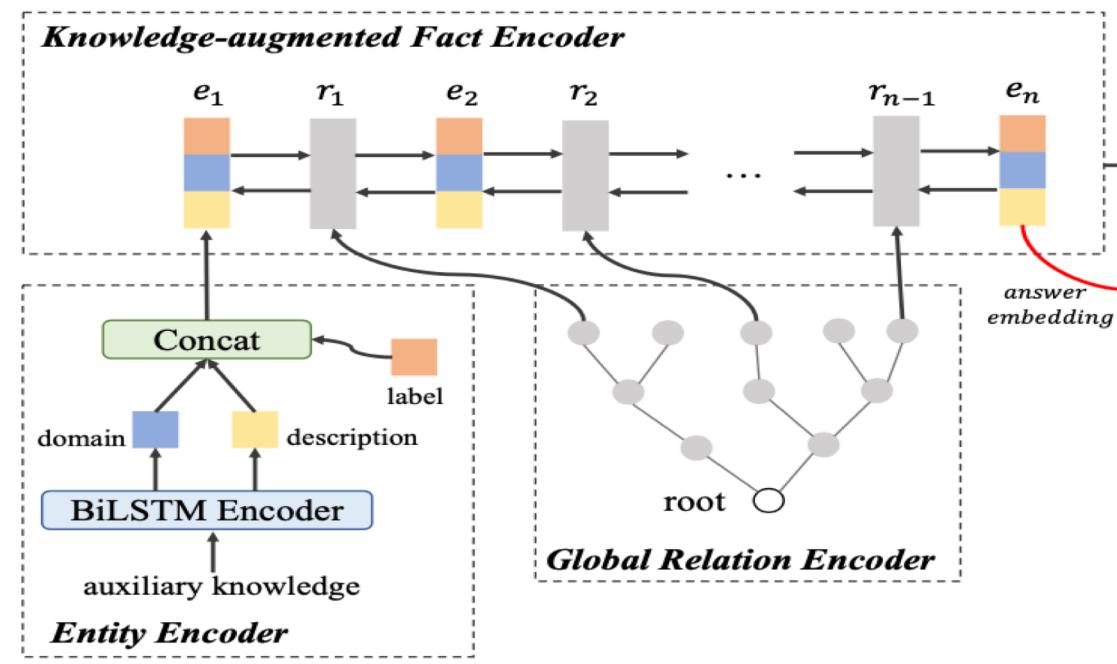
$$\begin{aligned} i_j &= \sigma(W^{(i)}r_j + \sum_{l=1}^N U_l^{(i)}h_{jl} + b^{(i)}), & f_{jk} &= \sigma(W^{(f)}r_j + \sum_{l=1}^N U_{kl}^{(f)}h_{jl} + b^{(f)}), \\ o_j &= \sigma(W^{(o)}r_j + \sum_{l=1}^N U_l^{(o)}h_{jl} + b^{(o)}), & u_j &= \tanh(W^{(u)}r_j + \sum_{l=1}^N U_l^{(u)}h_{jl} + b^{(u)}), \\ c_j &= i_j \odot u_j + \sum_{l=1}^N f_{jl} \odot c_{jl}, & h_j &= o_j \odot \tanh(c_j). \end{aligned}$$



# Method

- Knowledge-augmented Fact Encoder

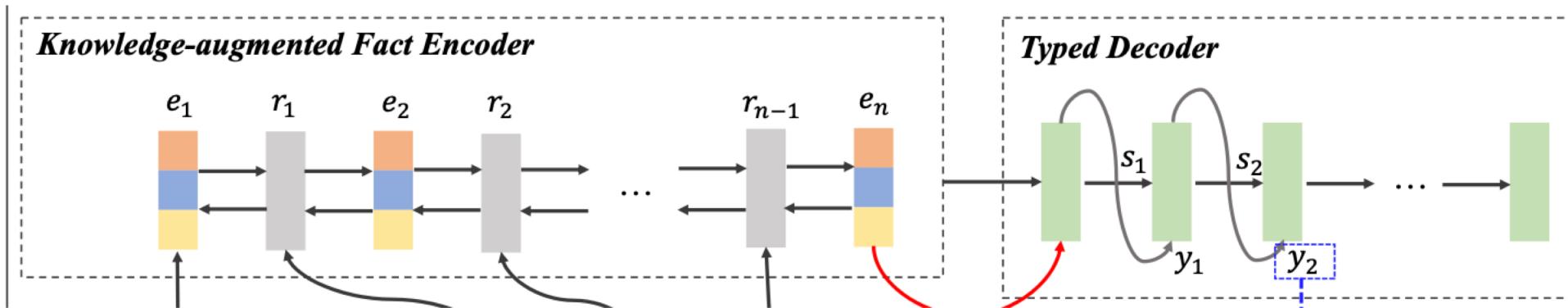
通过所有实体和关系的知识增强嵌入，我们使用具有输入序列 ( $e_1, r_1, e_2, \dots, r_{n-1}, e_n$ ) 的两层双向 LSTM 网络对三元组  $F$  进行编码。



# Method

- Typed Decoder

采用基于 LSTM 的 type deoder 来计算特定类型的单词概率分布，假设每个单词都有一个集合{疑问词、实体词、关系词、普通词}的潜在类型。结合起来，我们采用条件 copy 机制来允许从实体输入或关系输入进行复制。



注：由于生成问题的第一个标记是疑问句，这对于生成问题的语义一致性至关重要，我们首先使用答案嵌入，而不是特殊的序列开始标记，通过明确的答案嵌入，生成的疑问句更加准确，从而缓解了语义漂移问题。

# Method

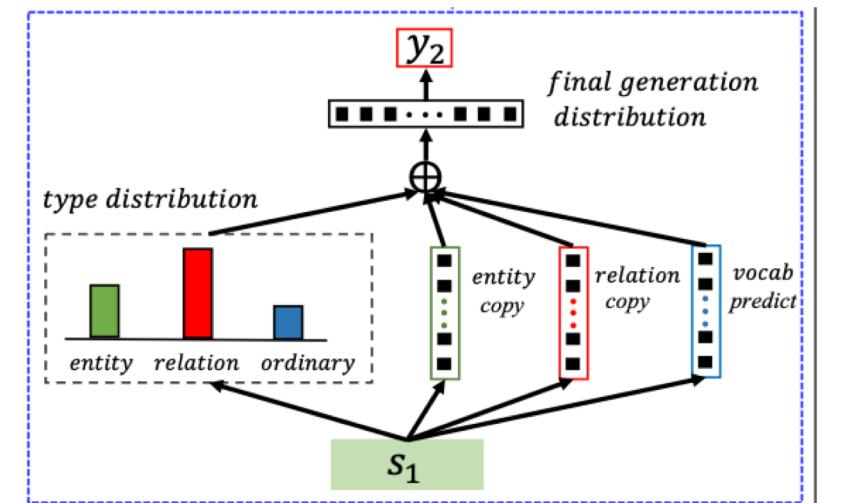
- Typed Decoder

拷贝机制：与传统解码器不同，我们的类型解码器计算特定类型的生成分布。生成疑问词后，接下来的解码步骤中词类只包括{实体词、关系词、普通词}。我们首先估计算单词类型上的类型分布，然后根据单词类型决定复制或生成单词

$$P_E(\omega) = \sum_{i:\omega_i=\omega} \alpha_{t,i}^e, \quad P_R(\omega) = \sum_{i:\omega_i=\omega} \alpha_{t,i}^r.$$

$$P(y_t|y_{<t}, E, R, K) = \sum_{g_i \in \{g_e, g_r, g_o\}} P(y_t|\tau_{y_t} = g_i, y_{<t}, E, R, K) \cdot P(\tau_{y_t} = g_i|y_{<t}, E, R, K).$$

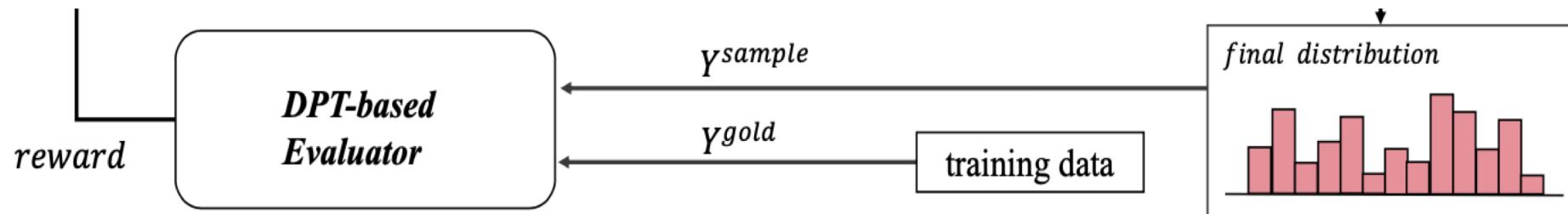
$$P(y_t|\tau_{y_t} = g_e, y_{<t}, F, K) = P_E, P(y_t|\tau_{y_t} = g_r, y_{<t}, F, K) = P_R, P(y_t|\tau_{y_t} = g_o, y_{<t}, F, K) = P_V.$$



# Method

- Evaluator

使用强化学习，不同于以往大部分文本生成工作会使用Bleu、rouge作为reward，这些方法会使模型生成在n-gram与标准相同的问句，但并不能有效提升生成问题的多样性。本文使用DPT[22]计算生成问题和标准问题的句法依存树的相似性，来帮助模型生成与标准问句相似的问句，确保生成问句在句法正确的同时具有多样性。



# Experiment

- Datasets

试验部分选择了数据集SimpleQuestion和PathQuestion。

- Data Proecess

为了获得辅助信息，本文对数据集进行了处理，将输入子图中的实体和关系链接到开放知识库Wikidata中，获得相应的实体description和domain信息，以及关系的层级信息。

# Experiment

- Baseline

Datasets	SimpleQuestion						PathQuestion						
	Metrics	B-4	ME.	R-L	Nat.	Div.	Cor.	B-4	ME.	R-L	Nat.	Div.	Cor.
RNN-based		19.98	28.43	46.02	2.3	1.2	2.0	25.78	33.17	50.78	2.5	1.6	2.2
Zero-shot		22.71	30.39	51.07	2.6	1.8	2.3	29.44	38.12	56.94	2.9	1.9	2.6
Multi-hop		25.98	34.14	56.03	2.8	1.9	2.4	34.14	41.77	62.12	3.1	2.2	2.8
Ans-aware		28.19	36.98	59.17	3.3	2.3	2.8	37.44	43.12	64.78	3.4	2.5	3.1
BiGraph2Seq		31.12	39.23	62.14	3.5	2.4	3.1	39.88	46.65	67.15	3.6	2.7	3.2
KTG $\oplus$ BLEU		34.89	42.55	65.54	3.9	3.0	3.4	42.09	49.77	69.98	4.1	3.3	3.7
KTG $\oplus$ ROUGE		34.68	42.04	64.89	3.8	2.9	3.4	41.67	49.12	69.24	4.0	3.2	3.6
KTG $\oplus$ QSS		35.04	43.12	65.99	3.9	3.0	3.4	42.85	50.36	70.44	4.1	3.3	3.7
<b>KTG</b>		<b>38.05</b>	<b>46.37</b>	<b>68.13</b>	<b>4.1</b>	<b>3.2</b>	<b>3.8</b>	<b>45.58</b>	<b>52.31</b>	<b>73.21</b>	<b>4.3</b>	<b>3.5</b>	<b>4.0</b>

# Experiment

- Ablation Test

Datasets	SimpleQuestion						PathQuestion						
	Metrics	B-4	ME.	R-L	Nat.	Div.	Cor.	B-4	ME.	R-L	Nat.	Div.	Cor.
<b>KTG</b>		<b>38.05</b>	<b>46.37</b>	<b>68.13</b>	<b>4.1</b>	<b>3.2</b>	<b>3.8</b>	<b>45.58</b>	<b>52.31</b>	<b>73.21</b>	<b>4.3</b>	<b>3.5</b>	<b>4.0</b>
w/o knowledge		28.01	36.97	59.79	3.4	2.3	3.7	39.24	45.63	66.38	3.5	2.6	3.9
w/o type		29.27	39.19	63.58	3.7	2.7	3.4	40.78	47.99	68.74	4.1	3.1	3.5
w/o RL		28.21	38.68	62.97	3.6	2.6	3.3	40.37	47.42	67.99	3.7	3.2	3.4

# Experiment

- Case Study

Model	$F = \{(laura\_devon, spouse, brian\_kelly\_hell), (brian\_kelly\_hell, institution, university\_of\_michigan)\}$
Ans-aware	who laura_devon's spouse is?
BiGraph2Seq	what is institution laura_devon's spouse?
KTG $\oplus$ QSS w/o knowledge	where does the American actress laura_devon's husband work for? where does the husband of laura_devon work for?
w/o type	what does the husband of American actress laura_devon work for?
w/o RL	where the American actress laura_devon husband work for?
<b>KTG</b>	<b>where does the husband of laura_devon, an American actress, work for?</b>
<b>Gold</b>	<b>where does the husband of American actress laura_devon work for?</b>

# Conclusion

- 解决了两个关键挑战：知识库（KBQG）问题生成任务的源输入不足和语义漂移问题。我们使用辅助知识丰富编码器输入，包括实体描述和谓词域，以提高问题的多样性。我们采用具有条件复制机制的类型化解码器来进一步提高生成问题的语义一致性。
- 通过强化学习进一步优化模型性能，并设计了一种基于语法相似性但不基于 n-gram 重叠的新奖励函数。这种奖励确保生成语法和语义上有效的问题，同时允许更多的多样性和流畅性。
- 两个基准数据集的实验结果表明，模型在所有自动和人工评估指标上都比最先进的模型实现了显著改进。

# Discussion

这篇文章设计方法一定程度上解决了问题生成存在的疑问词难以准确生成、问题提及实体无法准确生成、多样性较低等问题，都是KBQG领域大家比较关注的研究点。

这篇文章是为了提升生成问题的多样性和语义准确性，同时生成的高质量问题也可以用来训练QA模型，提升效果