Generating Questions for Knowledge Bases via Incorporating Diversified Contexts and Answer-Aware Loss

Cao Liu[1,2], **Kang Liu**[1,2], **Shizhu He**[1,2], **Zaiqing Nie**[3], **Jun Zhao**[1,2]
[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
[3] Alibaba AI Labs, Beijing, 100029, China
{cao.liu, kliu, shizhu.he, jzhao}@nlpr.ia.ac.cn
zaiqing.nzq@alibaba-inc.com

Input	<Statue of Liberty, location/containedby, ***New York City***>		
Output	Matching Predicate	Definite Answer	Question
Q1	×	-	Who **created** the Statue of Liberty?
Q2	√	×	**Where** is Statue of Liberty in?
Q3	√	√	Which ***city*** is Statue of Liberty **located** in?

Figure 1: Examples of KBQG. We aims at generating questions like Q3 which expresses (matches) the given predicate and refers to a definitive answer.

- (1) We leverage diversified contexts and multi-level copy mechanism to alleviate the issue of incorrect predicate expression in traditional methods.

- (2) We propose an answer-aware loss to tackle the issue that conventional methods can not generate questions with definitive answers.

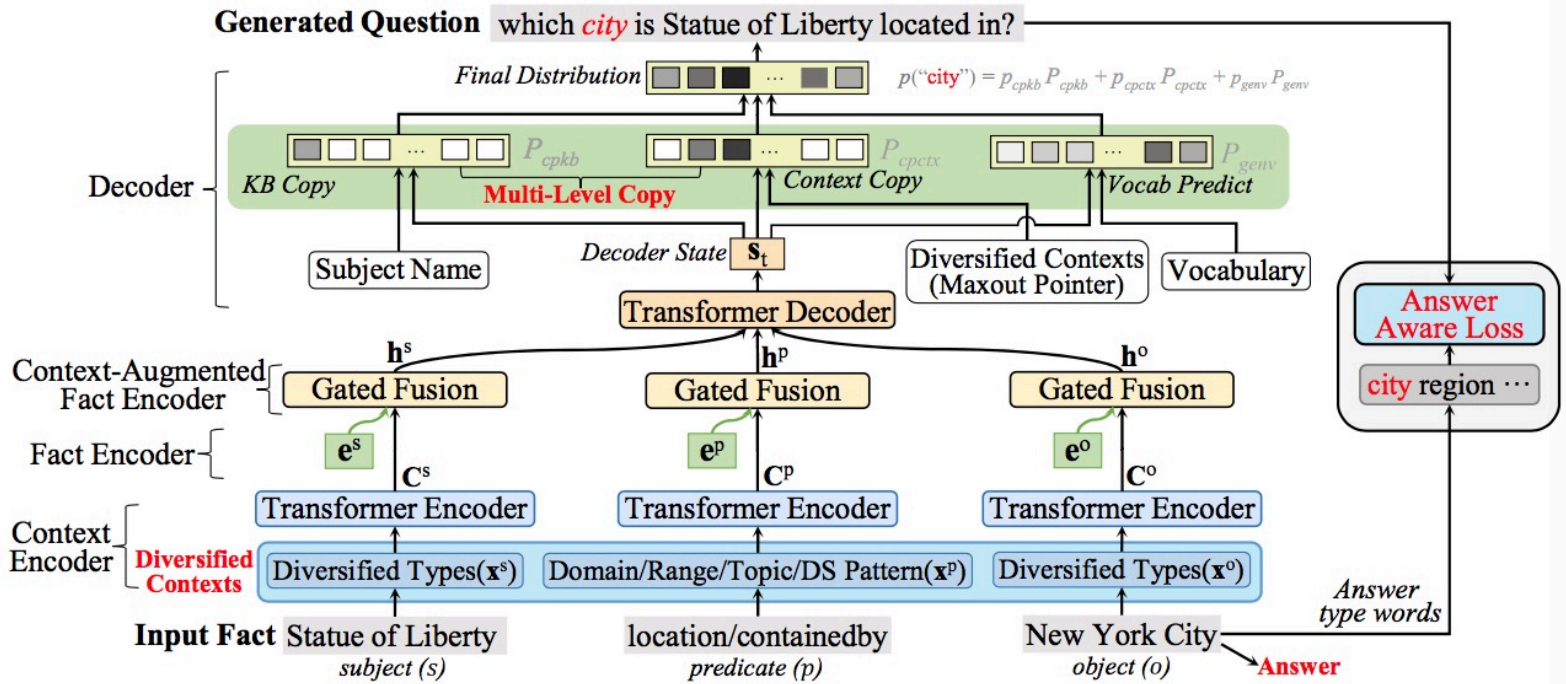

Figure 2: Overall structure of the proposed model for KBQG. A **context encoder** is firstly employed to encode each textual context (Sec. 3.1), where "Diversified Types" represents the subject (object) context, and "DS pattern" denotes the relational pattern from distant supervisions. At the same time, a **fact encoder** transforms the fact into low-dimensional representations (Sec. 3.2). The above two encoders are aggregated by the **context-augmented fact encoder** (Sec. 3.3). Finally, the aggregated representations are fed to the **decoder** (Sec. 3.4), where the decoder leverages multi-level copy mechanism (KB copy and context copy) to generate target question words.

KB-copy

- Previous study found that most questions contain the subject name or its aligns in SimpleQuestion (Petrochuk and Zettlemoyer, 2018). However, the predicate name and object name hardly appear in the question. Therefore, we only copy the subject name in the KB copy,

Answer-Aware Loss

- In order to make generated questions correspond to definitive answers, we propose a novel answer-aware loss.
- By answer-aware loss, we aim at generating an answer type word in the question, which contributes to generating a question word matching the answer type
- We treat object type words as the answer type words because the object is the answer.

$$\mathcal{L}_{ans\_loss} = \min_{a_n, a_n \in A} \min_{y_t, y_t \in Y} H_{a_n, y_t} \qquad (15)$$

$$A = \{a_n\}_{n=1}^{|A|}$$

Collection of Textual Contexts

Zero-Shot Question Generation from Knowledge Graphs for Unseen Predicates and Entity Types

- First, we align each triple in the FB5M KB to sentences in Wikipedia if the subject and the object of this triple co-occur in the same sentence

- We replace the positions of the subject and the object mentions with [S] and [O] to the keep track of the information about the direction of the relation

Freebase Relation	Predicate Textual Context
person/place_of_birth	[O] is birthplace of [S]
currency/former_countries	[S] was currency of [O]
airline_accident/operator	[S] was accident for [O]
genre/artists	[S] became a genre of [O]
risk_factor/diseases	[S] increases likelihood of [O]
book/illustrations_by	[S] illustrated by [O]
religious_text/religion	[S] contains principles of [O]
spacecraft/manufacturer	[S] was spacecraft developed by [O]

Table 2: Table showing an example of textual contexts extracted for freebase predicates

- In order to obtain diversified contexts, we additionally employ domain, range and topic of the predicate to improve the coverage of predicate contexts.

- For the subject and object context, we combine the most frequently mentioned entity type (Elsahar et al., 2018) with the type that best describe the entity

- a refined entity type "US state" combines a broad type "administrative region" for the entity "New York" form freebase

Dataset

- We conduct experiments on the SimpleQuestion dataset (Bordes et al., 2015), and there are 75910/10845/21687 question answering pairs (QA-pairs) for training/validation/test.

- outsourcing

Model	BLEU4	ROUGE_L	METEOR
Template	31.36	*	33.12
Serban et al. (2016)	33.32	*	35.38
Elsahar et al. (2018)	36.56	58.09	34.41
Our Model	41.09	68.68	47.75
Our $\text{Model}_{\text{ans\_loss}}$	**41.72**	**69.31**	**48.13**

Table 1: Overall comparisons on the test data, where "ans_loss" represents answer-aware loss.

Model	Pred. Identification
Serban et al. (2016)	53.5
Elsahar et al. (2018)	71.5
Our Model$_{\text{ans_loss}}$	**75.5**

Table 2: Performances on predicate identification.

ID	Model	Question
1	Reference	what is the origin of *Kate Bush ?*
	Serban et al.	where is catherine bush buried ?
	Elsahar et al.	what is the artist of catherine bush ?
	Ours	what is the origin of the artist *Kate Bush* ?
2	Reference	what area contains *River Yare* ?
	Serban et al.	where is the *River Yare* ?
	Elsahar et al.	where is the *River Yare* located ?
	Ours	what **city** is *River Yare* in ?
3	Reference	who composed *Bien O Mal* ?
	Serban et al.	who is the composer of *Bien O Mal* ?
	Elsahar et al.	who is the composer of the song *Bien O Mal* ?
	Ours	who composed *Bien O Mal* ?

Figure 4: Examples of questions by different models.

It can be seen that our generated questions can better express the tar- get predicate such as ID 1

In ID 2, although all questions express the target predicate correctly, only our question refers to a definitive answer since it contains an answer type word “city”