Large-scale Simple Question Generation by Template-based Seq2seq Learning

Tianyu Liu, Bingzhen Wei, Baobao Chang, Zhifang Sui

Key Laboratory of Computational Linguistics, Ministry of Education
School of Electronics Engineering and Computer Science, Peking University
No.5 Yiheyuan Road, Haidian District, Beijing, 100871, China
{tianyu0421,weibz,chbb,szf}@pku.edu.cn

Abstract. Numerous machine learning tasks achieved substantial advances with the help of large-scale supervised learning corpora over past decade. However, there's no large-scale question-answer corpora available for Chinese question answering over knowledge bases. In this paper, we present a 28M Chinese Q&A corpora based on the Chinese knowledge base provided by NLPCC2017 KBQA challenge. We propose a novel neural network architecture which combines template-based method and seq2seq learning to generate highly fluent and diverse questions. Both automatic and human evaluation results show that our model achieves outstanding performance (76.8 BLEU and 43.1 ROUGE). We also propose a new statistical metric called DIVERSE to measure the linguistic diversity of generated questions and prove that our model can generate much more diverse questions compared with other baselines.

Keywords: Question Generation, Template-based Seq2seq, Linguistic Diversity

1 Introduction

Question Answering (QA) over knowledge bases (KBs) aims at providing accurate answers formulated in natural language, with factual retrieval and inference in the knowledge bases. One of the major obstacles for training QA systems is the lack of high quality labeled data. The performances of question answering over knowledge bases (KBQA) systems highly depend on the data scale and quality because answer selection in KBQA always involves complex inference over relevant relationships between entities in the knowledge graph, which demands large-scale dataset in the training stage. Furthermore, even more labeled question-answer pairs are required in neural network-based QA systems.

Automatic question generation (QG) ([15], [4], [6]) has become a popular task for solving the data insufficient problems in QA systems. Most KB-based question generation methods have focused on constructing questions by utilizing multiple related facts in the KB. However, the simpler question-answering that involves only one single fact, which is called *Simple Question Answering* ([4]), is still far from solved. Besides, simple QA itself can cover a wide range of practical

Fact #1	全球通史 ||| 装帧 ||| 软装
Fact #2	商务星健身管理软件 ||| 经营范围 ||| 健身俱乐部管理软件
Fact #3	倭叉角羚 ||| 纲 ||| 哺乳纲
Fact #4	焖子 ||| 主要食材 ||| 地瓜淀粉 精瘦肉
Fact #5	真相 ||| 译者 ||| 陈睿 杨通

Fact	Gold	Pure Template	Seq2seq	Tseq2seq
#1	全球通史的装帧是什么样子的?	全球通史这本书共多少页?	全球通史的装帧是什么?	全球通史是怎样装帧的?
#2	商务星健身管理软件的经营范围是什么?	商务星健身管理软件主要做什么生意?	商务星健身管理软件的经营范围是什么?	商务星健身管理软件经营范围包括哪些?
#3	你知道倭叉角羚这种动物是什么纲的吗?	谁能告诉我倭叉角羚属于什么纲?	谁知道***偃叉角羚***是哪个纲的?	倭叉角羚属于什么纲?
#4	我想知道做焖子都需要什么食材?	焖子主要食材有什么?	*仗子*的主要食材是什么?	做焖子需要用什么材料?
#5	我想知道真相这本书是谁翻译的呀?	谁翻译了真相?	真相的译者是谁?	请问真相是谁翻译的?

Fig. 1. Simple questions generated by pure **Template-based Method**, vanilla **seq2seq** framework and **Template-based seq2seq** learning (Tseq2seq). Misleading questions generated by pure template-based method are marked in red. Questions that generates wrong subjects entities of the corresponding facts are marked in green.

usage if the referring KB is well-organized. For instances, all of questions in the training set of NLPCC2017 KBQA challenge can be answered by simple QA reasoning. Hence, as shown in Fig 1, we focus on generating simple questions which are only related to single facts in this paper.

Previous rule-based approaches ([13], [14], [5]) mainly relied on hand-crafted rules and heuristics to synthesize artificial QA corpora. The performances of these approaches hinges critically on the well-designed templates or rules. Although template-based methods are capable of generating reasonable questions in most cases, the generated questions still suffer from lack of diversity and fluency due to the limits of pre-defined templates or rules. As shown in Fig 1, the two red-marked questions generated by pure template-based method are misleading ones compared with the target gold questions.

Motivated by recent development of end-to-end neural generation models for machine translation ([12], [2]), image captioning ([10], [17], [7]) and dialogue generation ([16], [18]). We propose a neural generation baseline based on Encoder-decoder architecture. Seq2seq learning with attention mechanism achieves competitive performances in QG. However, vanilla seq2seq model still suffers from generating irrelevant or improper topic entities. Improper topic entities can cause severe damage to the quality of generated questions because factual inference and retrieval over KB are based on the recognized topic words or phrases while answering the questions. As shown in Fig 1, the topic words of the green-marked questions are generated improperly as ’ 偃叉角羚’ and ’ 仗子’ while the true topic words are ’ 倭叉角羚’ and ’ 焖子’.

In this paper, we propose the following three models for generating simple questions based on Chinese KB: 1) Pure template-based method which utilizes the templates extracted from the training set to generate new questions for corresponding facts in the testing set, 2) Vanilla seq2seq architecture for neural question generation, 3) An integration of pure template-based method and vanilla seq2seq structure — Template-based seq2seq learning for generating highly accurate and linguistically diverse questions. The main contributions of our work can be summarized as follows:

- We propose a new Template-based seq2seq neural question generation architecture to tackle the improper-topic-words problem in vanilla seq2seq generation as well as the lack of linguistic diversity and misleading generation problems in pure template-based method.
- Our proposed template-based seq2seq question generation achieve outstanding performance in both human (92.5% in accuracy) and automatic evaluation (76.8 BLEU and 43.1 ROUGE).
- We propose a new statistical metric based on the sentence similarity called **DIVERSE** to measure the linguistic diversity of generated questions and prove that template-based method can generate more diverse questions than vanilla seq2seq structure and pure template-based method.
- To the best of our knowledge, we first proposed a large-scale QA corpus (28M) for Chinese KBQA.

2 Task Definition

In this section, we describe the knowledge base used in this paper and the probabilistic framework for question generation.

2.1 Knowledge bases

A knowledge base (KB) is a highly structured multi-relational database, which consists of entities and corresponding relationships. The relationships in the KBs are directed and always connect exactly two entities. For example, in Freebase ([3]), two entities *Barack Obama* and *Honolulu* are connected by the relation *place of birth* which represents the birthplace of Barack Obama is Honolulu. The two entities with a particular relationship that connects them consist a *factual triple*

	FB2M	**FB5M**	**NLPCC2017**
Entities	2,150,604	4,904,397	6,502,738
Relationships	6,701	7,523	548,225
Facts	14,180,937	22,441,880	43,063,796

Table 1. Statistics of the NLPCC2017 Chinese Knowledge Base used in this paper. The two versions of **Freebase**, **FB2M** and **FB5M** are provided for comparison.

of the KB. The knowledge base we used in this paper is the Chinese knowledge base provided by the NLPCC2017 KBQA challenge[1]. Table 1 lists the number of entities, relations and facts in the NLPCC2017 Chinese Knowledge base.

2.2 Generating Questions from Triples

We intend to generate questions from given factual triples. Given a single factual triple F=(*Subject*, *Relationship*, *Object*), we aims to produce a question which is concerned with the subject and the relationship of the fact and can be properly answered by corresponding object. The question generation procedure can be modeled in a probabilistic framework:

$$P(Q|F) = \prod_{i=1}^{N} P(w_i|w_{<i}, F) \tag{1}$$

Where $Q = (w_1, w_2, \cdots, w_N)$ represents the generated question which consists of tokens $w_1, w_2, \cdots, w_N$. In most cases, the last generated token w_N is '?'.

3 Pure Template-based Method

Template-based method is an automatic question generation baseline which utilizes templates extracted from training set and then generates questions by filling the particular templates with certain topic entities. Given a question Q and its corresponding factual triple (T, R, O) in which T represents *topic entity*, P represents *relationship*, O represents *object*. As shown in Fig 2, template-based question generation framework consists of two phases: **Template Collection** and **Selective Generation**.

In template collection phase, we extract question templates from training set and produce a template pool P for each relation. Firstly, a question template is produced by replacing the topic entity T in the question with a special token *(SUB)*, and then we regard the question template as an instance of the template pool for corresponding relationship R. As show in Fig 2, each relationship R has a template pool which consists of one or more templates.

In selective generation phase, given a triple $H = (t, r, o)$ from KB and template pool P generated in template collection phase, we randomly select a template Q whose corresponding relationship is r from template pool P. Then, we replace the special token (*SUB*) in the template Q with topic entity t in triple H to get specific generated question.

Template-based method can generate understandable questions for most triples and achieve competitive performance in automatic and human evaluation. However, as described in Section 1 and Fig 1, the questions generated by pure template-based method could be misleading or lack linguistic diversity due to the limits of templates.

[1] `http://tcci.ccf.org.cn/conference/2017/cfp.php`

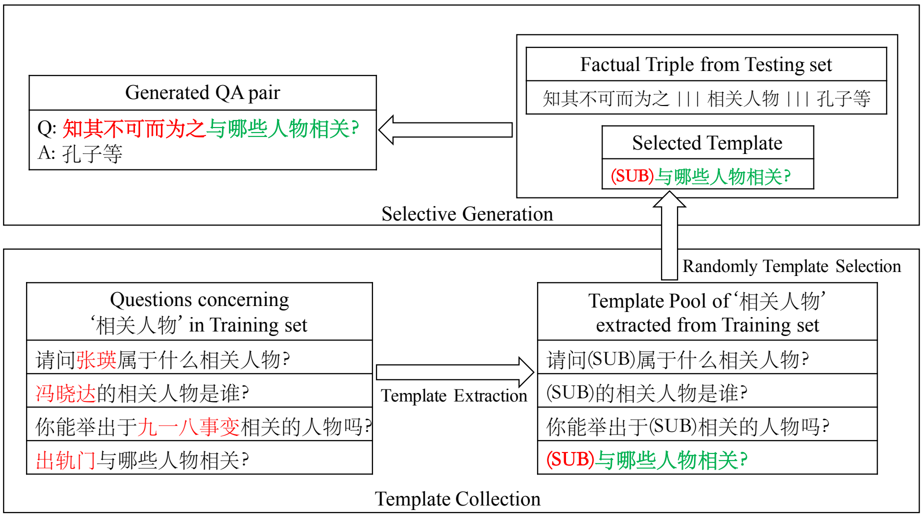
Generated QA pair
Q: 知其不可而为之与哪些人物相关?
A: 孔子等
Factual Triple from Testing set
知其不可而为之 ||| 相关人物 ||| 孔子等
Selected Template
(SUB)与哪些人物相关?
Selective Generation
Randomly Template Selection
Questions concerning ‘相关人物’ in Training set
请问张瑛属于什么相关人物?
冯晓达的相关人物是谁?
你能举出于九一八事变相关的人物吗?
出轨门与哪些人物相关?
Template Extraction
Template Pool of‘相关人物’ extracted from Training set
请问(SUB)属于什么相关人物?
(SUB)的相关人物是谁?
你能举出于(SUB)相关的人物吗?
(SUB)与哪些人物相关?
Template Collection

Fig. 2. Framework for template-based question generation method. The *topic entities* are marked in red while the *selected template* is marked in green. In **Template Collection** module, question templates are extracted from questions in the training set to form a *template pool* for certain relationship. In **Selective Generation** module, the randomly selected templates from *template pool* and the corresponding factual triples are used to generate target QA pairs.

4 Template-based Neural Generation

In this section, we describe the architecture of proposed template-based neural generation by seq2seq learning. Seq2seq model is an effective structure in modeling sequence-to-sequence translation. Our templated-based seq2seq can be viewed as a translation from structured data (factual triples in the KB) to simple questions that can be answered by single triples. The model can be divided into triple encoder and template decoder. The procedure of template-based seq2seq question generation is introduced in the final part of this section.

4.1 Triple Encoder

Given a factual triple $F = (t, r, o)$, in which topic entity $t = \{T_1, T_2, \cdots, T_m\}$ and relationship $r = \{R_1, R_2, \cdots, R_n\}$, where m, n represent the length of topic entity t and relationship r respectively. One-hot vectors $T_i, R_j \in \mathbb{R}^{|V|}$, where $|V|$ is the size of vocabulary, represent one single token (including a Chinese character, a punctuation and a letter) respectively in the topic entity and corresponding relationship. We get the word embeddings t_i, r_j of T_i, R_j by looking up the embedding matrix $E \in \mathbb{R}^{|V| \times K}$, where K represents the size of word embedding:

$$t_i = E \cdot T_i; r_j = E \cdot R_j \tag{2}$$

We get the representation of the topic entity $\mathbf{t} = \{t_1, t_2, \cdots, t_m\}$ and the relationship $\mathbf{r} = \{r_1, r_2, \cdots, r_n\}$ by concatenating the word embedding of every token in the topic entity and the relationship. To represent the given factual triple F, we insert a special token *SEP* between the topic entity representation $\mathbf{t}$ and the relationship representation $\mathbf{r}$ to separate the two parts in the sequential input of the triple encoder. To be more specific, we exploit a vector $\mathbf{w} \in \mathbb{R}^{m+n+1}$ to represent the factual triple $F = (t, r, o)$ by concatenating the representations of the topic entity, the separation token *SEP* and the relationship:

$$\mathbf{w} = [t_1, t_2, \cdots, t_m, SEP, r_1, r_2, \cdots, r_n] \tag{3}$$

Then, we use the LSTM architecture to encode the factual triple F:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{c}_t \end{pmatrix} = \begin{pmatrix} sigmoid \\ sigmoid \\ sigmoid \\ tanh \end{pmatrix} W_{4n,2n} \begin{pmatrix} w_t \\ h_{t-1} \end{pmatrix} \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \tag{5}$$

$$h_t = o_t \odot tanh(c_t) \tag{6}$$

where h_t is the hidden state at time step t, n is the size of hidden layer, $i_t, f_t, o_t \in \mathbb{R}_n$ are input, forget, output gate of each LSTM unit respectively, $\hat{c}_t$ and c_t are proposed cell value and true cell state at time t, $W_{4n,2n}$ is the model parameter to be learned.

4.2 Template Decoder

To exploit the alignment information between factual triples and generated questions, we use LSTM architecture with attention mechanism as our neural generator. As defined in the equation 1, the generated token y_t at time t in the decoder is predicated based on all the previously generated tokens $y_{<t}$ before y_t and the hidden states $H = \{h_t\}_{t=1}^{L}$ of the triple encoder. To be more specific:

$$P(y_t|H, y_{<t}) = softmax(W_s \odot tanh(W_t[s_t, a_t])) \tag{7}$$

$$s_t = LSTM(y_{t-1}, s_{t-1}) \tag{8}$$

s_t is the t-th hidden state of the decoder calculated by the LSTM unit in which the computational details can be referred in Equation 4, 5 and 11. a_t is the attention vector which is represented by the weighted sum of encoder hidden states.

$$a_t = \sum_{i=1}^{L} \alpha_{t_i} h_i ; \alpha_{t_i} = \frac{e^{g(s_t, h_i)}}{\sum_{j=1}^{N} e^{g(s_t, h_j)}} \tag{9}$$

where $g(s_t, h_i)$ is a relevant score between decoder hidden state s_t and encoder hidden state h_i. There are many different ways to calculate the relevant

Given Triple: 于海 ||| 相关人物 ||| 吴冠中、张建中、爱新觉罗•溥铮
Generated Question: 请 问 与 于 海 有 关 的 人 有 谁?

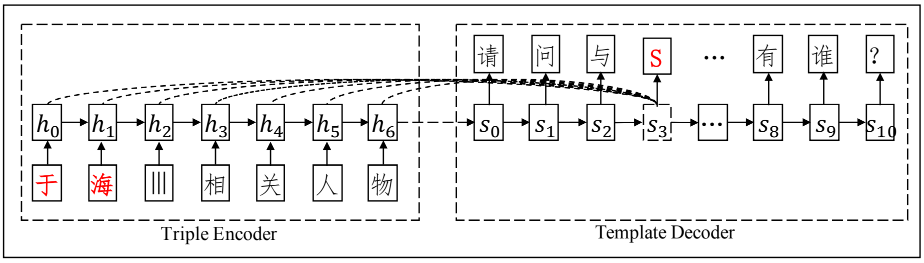

Fig. 3. An example for template-based seq2seq learning. Given an certain factual triple (t, r, o), the topic entity t (marked in red), *SEP* token ('|||') and corresponding relationship r are feed sequentially into the **Triple Encoder** as described in Section 4.1. Then the **Template Decoder** elaborated in Section 4.2 generates a question template in which topic entity is replaced by a specific token *SUB* (red-marked 'S'). Finally, the generated question template is transformed into the complete question by changing the *SUB* token into the topic entity t.

scores, in our paper, we use the following dot product to measure the similarity between s_t and h_i. W_s, W_t, W_p, W_q are all parameters that can be learnt in the model.

$$g(s_t, h_i) = tanh(W_p h_i) \odot tanh(W_q s_t) \tag{10}$$

4.3 Template-based Seq2seq

As explained in Section 1, vanilla seq2seq baseline has a quite severe problem: improper topic words. To alleviate the improper-topic-words problem, we propose template-based seq2seq framework. As shown in Fig 3, after feeding the triple representation **w** in Equation 3 into the triple encoder, we intend to generate a *question template* for that triple rather than the complete question.

In the neural generated *question template*, the topic entity is replaced by a specific token *SUB*. To get the complete question, we change the *SUB* token back to the specific topic entity. In the training procedure, the topic entities of target gold questions are also replaced by *SUB*. We use cross-entropy loss function between the replaced target questions and generated questions while training.

5 Experiments

In this section, we first evaluate the correctness of generated questions in the three proposed models by human and automatic evaluation. After that, we prove that template-based seq2seq model can generate more diverse questions than the other two baselines by the **DIVERSE** metric defined in Section 5.4.

5.1 Dataset and Evaluation Metrics

We conduct experiments on the *Simple Question Answering* dataset provided by NLPCC KBQA Challenge. The question-answering dataset contains 24,479 questions with answers that can be inferred and retrieved from the NLPCC Chinese Knowledge Base described in Section 2.1. As shown in the Table 2, we integrated the corresponding factual triples into the QA pairs by analyzing the given questions and answers and then retrieving related triples by topic entities and target answers from the knowledge base. The training set, validation set and testing set contains 11687/ 2922/ 9870 QA pairs respectively.

We measure the performance of our models by both automatic and human evaluation. To evaluate the correctness of the generated questions, we use **BLEU** (BLEU-4), **ROUGE** (ROUGE-4 F measure) evaluation and human evaluation. For human evaluation, We randomly select 200 instances from the generated questions and manually determine whether a specific question is proper or not. The human evaluation results are the ratio of proper questions. To measure the diversity of generated question, we propose a **DIVERSE** evaluation (described in Section 5.4) which is based on the sentence similarities within a cluster of questions.

<table>
<tr><td><b>Question</b></td><td>有人知道鸡黍之交的相关人物都有谁吗？</td></tr>
<tr><td><b>Factual Triple</b></td><td>鸡黍之交 ||| 相关人物 ||| 范式与张劭</td></tr>
<tr><td><b>Answer</b></td><td>范式与张劭</td></tr>
</table>

Table 2. An instance of (Question, Triple, Answer) tuples used in the experiments.

5.2 Experiment Setup

We use character-by-character input in the seq2seq learning, the triple inputs for the encoder and the question inputs for the decoder are all split into characters, including Chinese characters, letters and punctuations. All the word embeddings of these characters are initialized with the 400-dimensional vectors without pre-training. We build the vocabulary dictionary, which contains 3652 unique tokens, by including all the characters in the training set. Long Short Term Memory Unit (LSTM)[9] is chosen for our models and the hidden size of LSTM unit is set to 200. All the variables used in the network are initialized by Xavier initializer [8]. We use Adam optimizer [11] for optimization with a first momentum coefficient of 0.9 and a second momentum coefficient of 0.999. The initial learning rate is 5e-4 and the batch size in the training stage is 32 determined by a grid search over combinations of initial learning rate [1e-4, 5e-4, 1e-3, 5e-3] and batch sizes [16, 32, 64, 128]. We get the best configuration of parameters based on performance on the validation set, and only evaluate that specific configuration on the testing set. All the implementations of our models are based on the TensorFlow [1] framework.

5.3 Quality Analysis

In order to analyze the correctness and quality of generated questions by our models. We use both automatic and human evaluation to analyze the performance of proposed template-based neural generation model as well as template-based method and vanilla seq2seq generation model. Table 3 shows the performances of three models that we proposed. We have following observations:

(1) **Pure Template-based Method** achieves competitive results in both automatic and human evaluation, especially in the BLEU metric. We assume the underlying reasons for this condition are the homogeneity between training and testing dataset and the limited size of testing set. The templates extracted from training set suit the factual triples in the testing set well because of similar sentence patterns in both training and testing set. Furthermore, the size of testing set are comparatively small so that the several extracted templates are able to cover most of questions in the testing set.

(2) **Vanilla Seq2seq** model achieves a little better performance in ROUGE metric than template-based method because of the strong ability of the encoder-decoder architecture in language generation. However, its results for both ROUGE and human evaluation can't rival those of template-based method because vanilla seq2seq generation might produce wrong topic entities, especially for long and complex entities, which greatly hurt the quality of generated questions.

(3) **Template-based seq2seq** model gets the best performance among all three proposed models. Template-based seq2seq model combines the advantages of template-based baseline and seq2seq learning. Compared with vanilla seq2seq model, template-based seq2seq model deal with the improper-topic-entity problem by incorporating template mechanism so that it outperforms the vanilla seq2seq model by approximately 5 ROUGE and 2 BLEU. Additionally, template-based model outperform the pure template-based method by 0.5 BLEU on the condition that training set and testing set are highly homogeneous.

Models	**ROUGE**	**BLEU**	**Human**
Template-based Baseline	37.84	76.33	87.0
Seq2seq	38.41	74.86	83.5
Template-based Seq2seq	**43.11**	**76.84**	**92.5**

Table 3. Automatic and human evaluation performance of proposed models.

5.4 Study on Diversity

In the question generation task, the capability of generating linguistically diverse questions is another key point apart from the semantic correctness of the questions. So we make a comparison over the linguistic diversity of all the generated questions among the three proposed models.

To measure the linguistic diversity of generated questions, we propose a new statistical metric called **DIVERSE** which is based on the TF-IDF similarities of n generated questions $Q = (q_1, q_2, \cdots, q_n)$ that share the same relationship

Models	N=[3,4]	N=[5,∼]	Aggregate
Template-based Baseline	12.30	9.33	11.97
Seq2seq	10.35	7.23	9.74
Template-based seq2seq	**4.98**	**3.63**	**4.65**

Table 4. Results of **DIVERSE** in different configurations. As explained in Sec 5.4, the smaller DIVERSE is, the more linguistically diverse the generated questions are.

in the factual triples $Fs = ([S_1, S_2, \cdots, S_n], R, [O_1, O_2, \cdots, O_n])$. We call these triples *Fs triple clusters*. The TF-IDF question similarity $Tfidf_{sim}$ is a statistical measurement base on the frequencies of words within the questions generated by the *triple clusters*. We use *gensim*[2] to calculate the TF-IDF similarities between questions. To determine the aggregated similarity among all the questions in the cluster, the DIVERSE metric is defined as the average TF-I DF similarity of all possible permutations in the *Cartesian Product* of (Q, Q):

$$\textbf{DIVERSE} = \frac{1}{C_n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 1(i \neq j) \times Tfidf_{sim}(q_i, q_j) \tag{11}$$

where C_n^2 is a combination number, $1(x)$ is a conditional expression whose value is 1 if boolean expression x is True, otherwise 0. DIVERSE of certain question cluster generated by a particular *triple cluster* reflects the aggregated similarity of questions within that cluster. Since generated questions within the same cluster share the same relationship, we believe those clusters in which the aggregated sentence similarities are smaller are more linguistically diverse. In other words, **the smaller DIVERSE is, the more linguistically diverse the generated questions are.**

Table 4 shows the DIVERSE in different experimental configurations. To avoid the negative influence of relationships which are included only in one or two triples. We chose 505 relationships from the testing set which are included in more than two triples. The number of relationships which are included in exactly 3, 4, 5, 6 and more than 6 facts are 6/ 400/ 4/ 81 and 14 respectively. $N = [3, 4]$ means the DIVERSE of relationships which are included in 3 and 4 facts while $N = [5, \sim]$ means the DIVERSE of relationships which are included in more than 4 facts. As demonstrated in Table 4, we can see that template-based seq2seq have smallest DIVERSE in all configurations which means template-based neural generation model can create more diverse questions than the other two proposed models.

5.5 Proposed KBQA Corpus

To make full use of the proposed template-based neural generation model, we create a large-scale Chinese simple question-answering dataset [3]. Firstly, we retrieve all the triples whose relationships are contained in the training set of

[2] `http://radimrehurek.com/gensim/`

[3] We will release the dataset in the future

	SimpleQuestion	**Proposed corpus**
Entities	131,684	5,997,954
Relationships	1,837	4,222
Questions	108,442	28,133,837

Table 5. Statistics of proposed QA corpus for Chinese KBQA. We compare our propose corpus with the famous SimpleQuestion [4] Dataset.

NLPCC2017 KBQA Challenge dataset and then generate raw questions according to corresponding triples under the best configuration of proposed model. After that, we utilize the following filtering approaches to reprocess the raw questions to get the filtered questions: 1) filtering out the questions which contain *UNK* token. 2) To make sure the neural generated questions are readable and understandable, we also filter questions which doesn't end with '?' or whose length is longer than 50 to avoid meaningless repetitive questions.

The details of the proposed corpus are listed in Table 5. From the table, we can find out that our dataset has much larger scale than the famous SimpleQuestion dataset ([4]). We hope the given dataset might be useful while conducting further researches in the field of Chinese KBQA.

6 Conclusion

We propose a Template-based Seq2seq Neural Generation model for generating simple Chinese questions for question-answering over knowledge bases. The proposed template-based seq2seq model achieves outstanding performance in both human and automatic evaluations. Furthermore, we propose a new statistical metric **DIVERSE** to measure the linguistic diversity of the generated questions and prove that the template-based seq2seq model can also generate more diverse questions than vanilla seq2seq model and pure template-based method. We also utilize the template-based seq2seq model and several filtering approaches to create a large-scale dataset for Chinese KBQA.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work is supported by the National Key Basic Research Program of China (No. 2014CB340504) and the National Natural Science Foundation of China (No.61375074, 61273318). The contact authors are Zhifang Sui and Baobao Chang.

References

1. Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. 2016.

2. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
3. Kurt D Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. pages 1247–1250, 2008.
4. Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*, 2015.
5. Antoine Bordes, Jason Weston, and Nicolas Usunier. Open question answering with weakly supervised embedding models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 165–180. Springer, 2014.
6. Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*, 2017.
7. Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.
8. Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, pages 249–256, 2010.
9. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735, 1997.
10. Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.
11. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.
12. Minh-Thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*, 2014.
13. Ruslan Mitkov and Le An Ha. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, pages 17–22. Association for Computational Linguistics, 2003.
14. Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. The first question generation shared task evaluation challenge. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 251–257. Association for Computational Linguistics, 2010.
15. Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807*, 2016.
16. Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.
17. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
18. Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*, 2015.