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

Authors: **Tianyu Liu**, Bingzhen Wei, Baobao Chang and Zhifang Sui

Organization: Key Laboratory of Computational Linguistics(ICL), Peking University

Speaker: **Tianyu Liu**

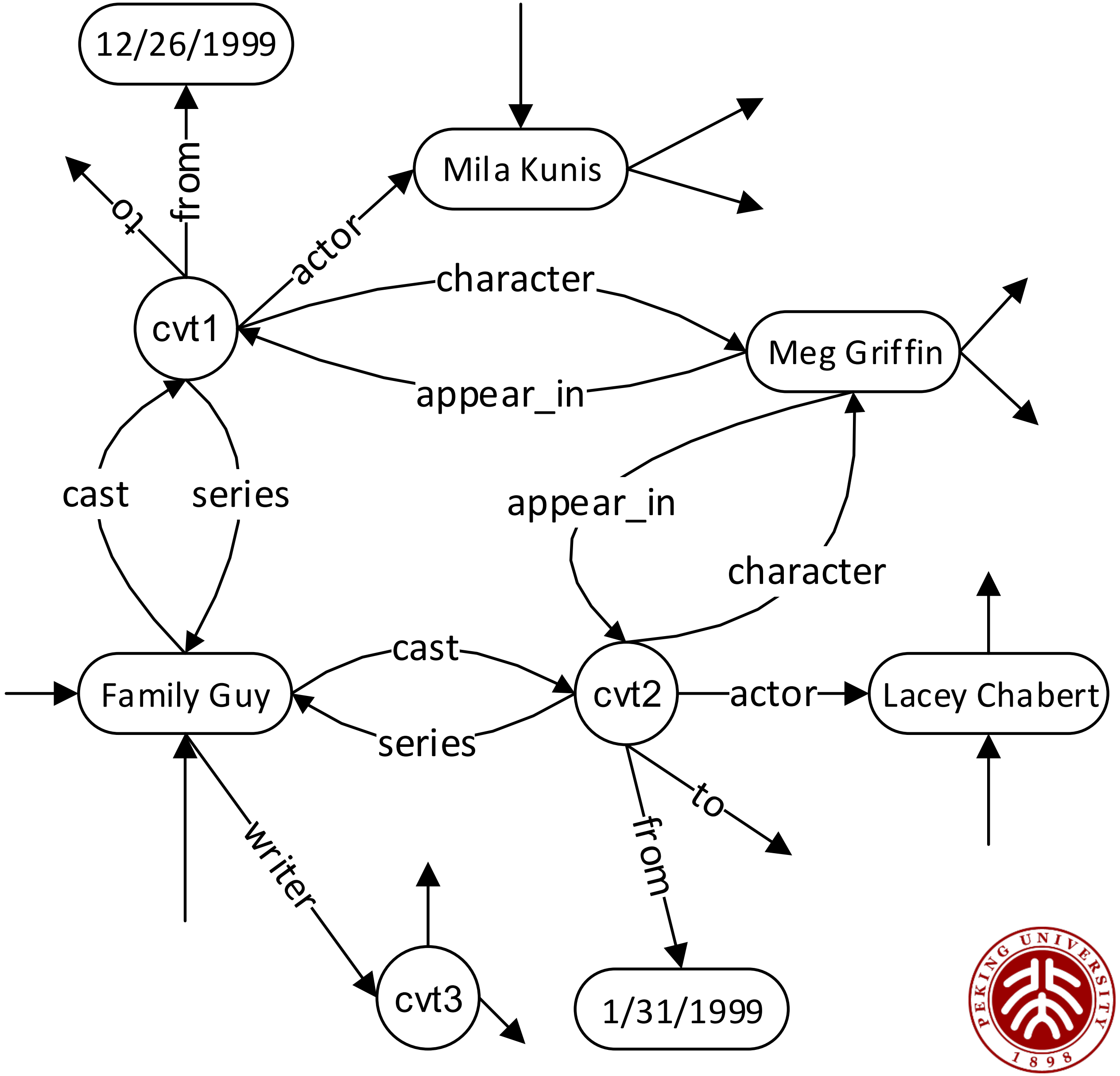
E-mail: tianyu0421@pku.edu.cn

Outline

- Background
  - Graph knowledge base
  - Simple question
- KB-based question generation
  - Template Extraction
  - Template-based Seq2seq model
  - Case study
- Dataset & metrics
- Experiments
- Large-scale Chinese KBQA dataset
  - Triple Selection and Question Filtering
  - Dataset Analysis
- Future Work

Knowledge Base

- Triples of **subj-pred-obj** (h, r, t)
- Knowledge graph
  - Each entity is a node
  - Two related entities linked by a directed edge (predicate)



Simple questions vs. Compositional questions

Simple question

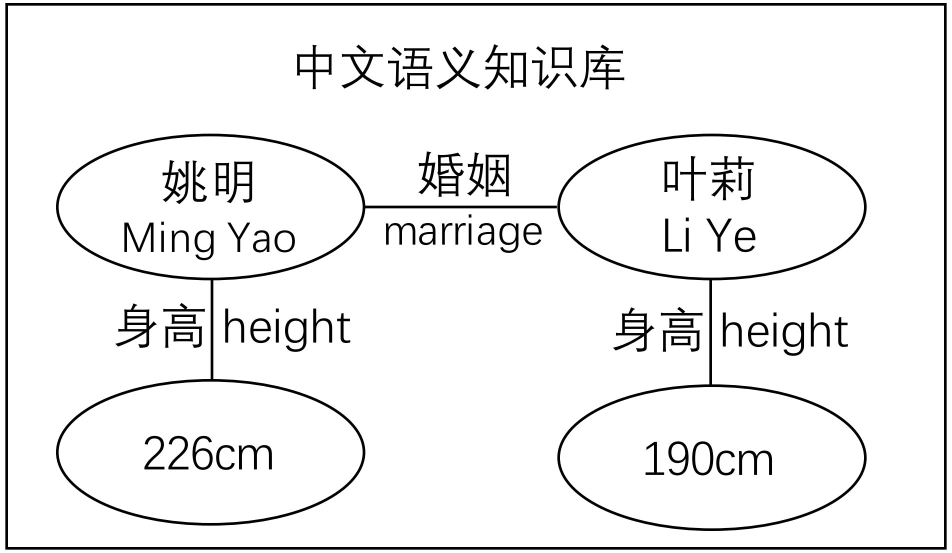
姚明身高多少？ How tall is Ming Yao?

姚明 $\xrightarrow{height}$ 226cm

Compositional question

姚明妻子身高多少？ How tall is Ming Yao’ s wife?

姚明 $\xrightarrow{marriage}$ 叶莉 $\xrightarrow{height}$ 190cm



中文语义知识库

姚明
Ming Yao

婚姻
marriage

叶莉
Li Ye

身高 height

身高 height

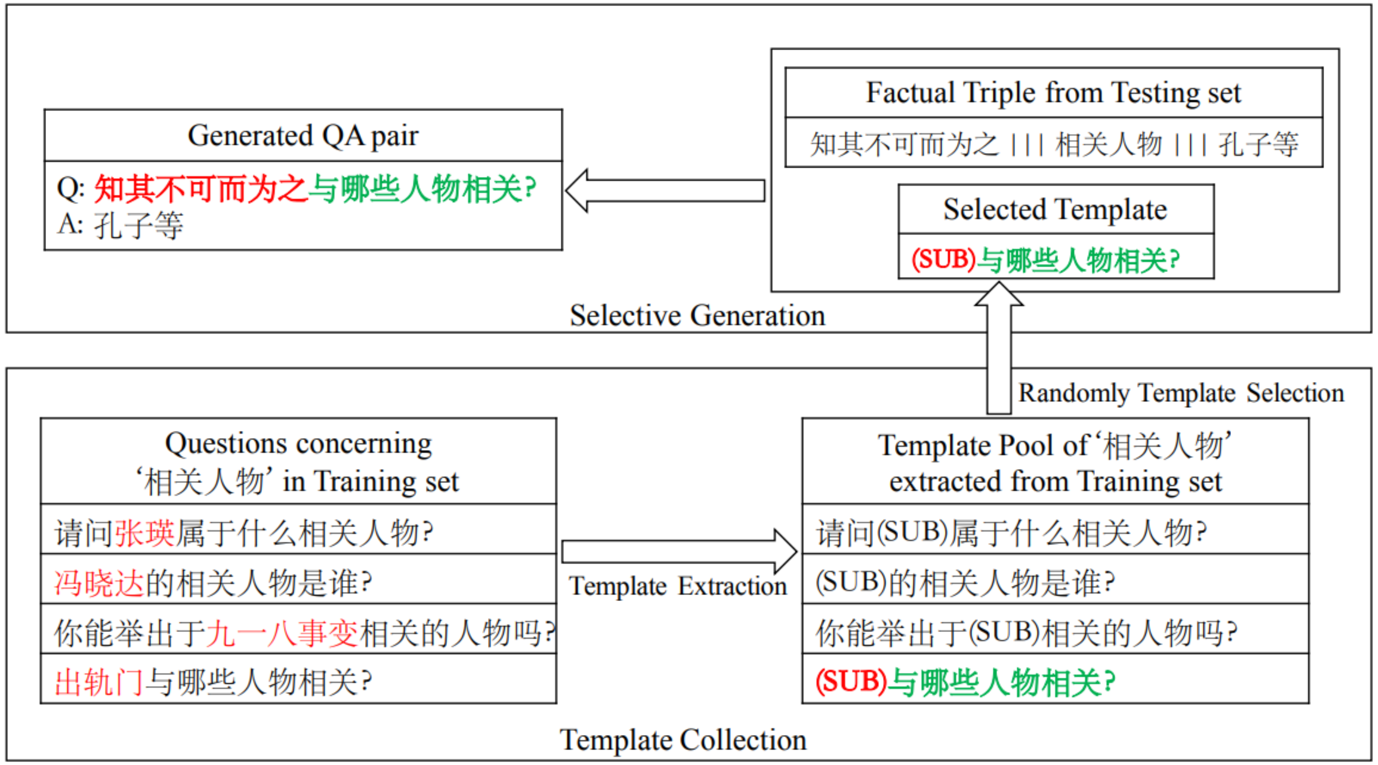
226cm

190cm

Outline

- Background
    - Graph knowledge base
    - Simple question
- KB-based question generation
    - Template Extraction
    - Template-based Seq2seq model
    - Case study
- Dataset & metrics
- Experiments
- Large-scale Chinese KBQA dataset
    - Triple Selection and Question Filtering
    - Dataset Analysis
- Future Work

Pure Template Extraction model



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

Template Collection

For a specific relationship *r*

1. Extract templates by replacing topic entity with a special token *(SUB)* in each question.
2. Collect all the templates concerning *r* to form a template pool.

Selective Generation

Given a factual triple <h, *r*, t>

1. Randomly select a template from *r*’ s template pool.
2. Generate questions by replacing the special token *(SUB)* in the selected pattern.

Template-based seq2seq model

Triple Encoder

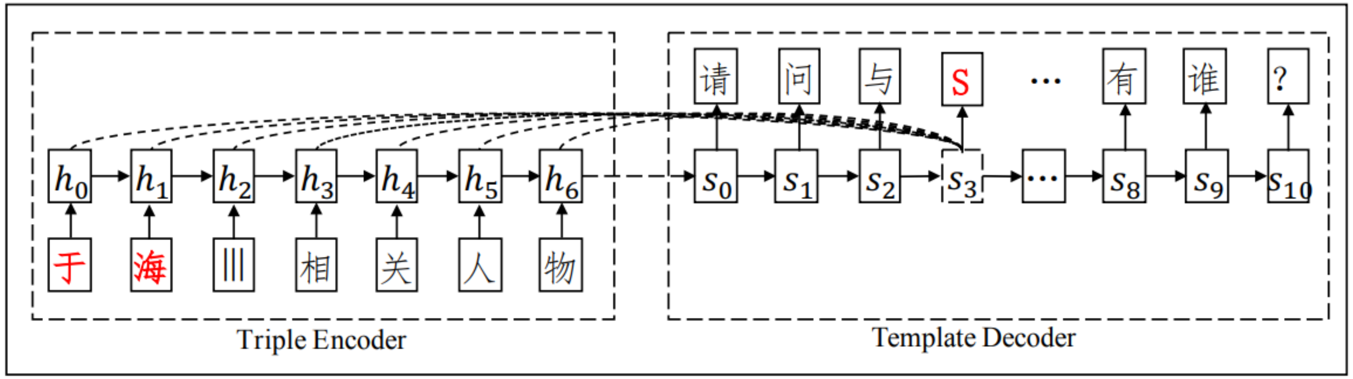
Given a triple fact $F = < t, p, o >$

- Topic entity $t = \{t_1, t_2, \cdots, t_n\}$
- Relation predicate $p = \{p_1, p_2, \cdots p_m\}$
- Object entity o $= \{o_1, o_2, \cdots o_l\}$

➢ Input: $w = [t_1, t_2, \cdots, t_n, SEP, p_1, p_2, \cdots p_m] \in \mathbb{R}^{m+n+1}$

➢ Encoder state: $h_t = LSTM\,(h_{t-1}, w_{t-1})$

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?



Template-based seq2seq model

Template Decoder

Given encoder states $H = \{h_t\}_{t=1}^{L}$ and previous generated tokens $y_{<t}$

- ➢ Probability of Generating next token

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

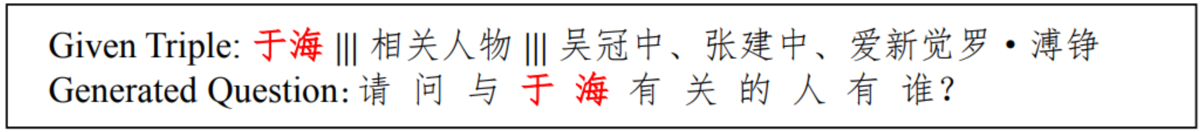
- ➢ Decode states

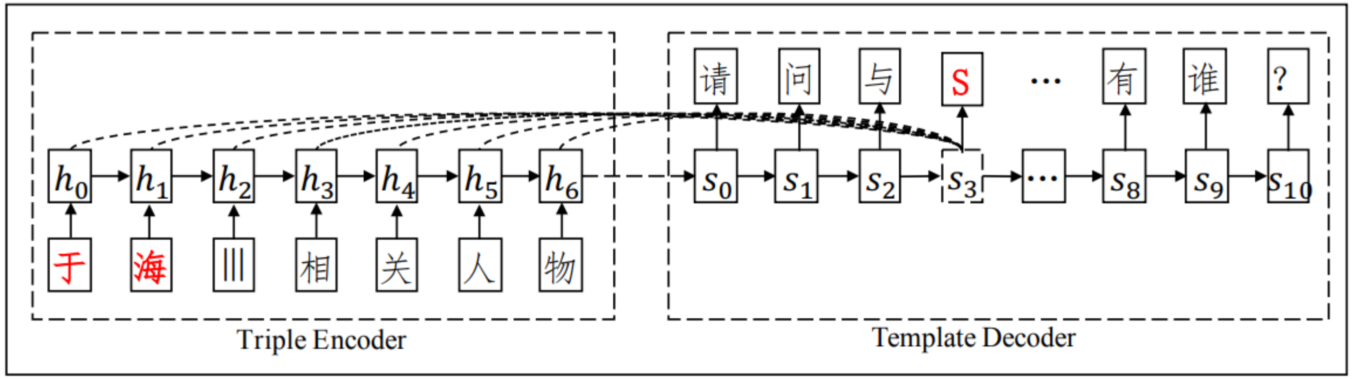
$$s_t = LSTM(s_{t-1}, y_{t-1})$$

- ➢ Attention vector

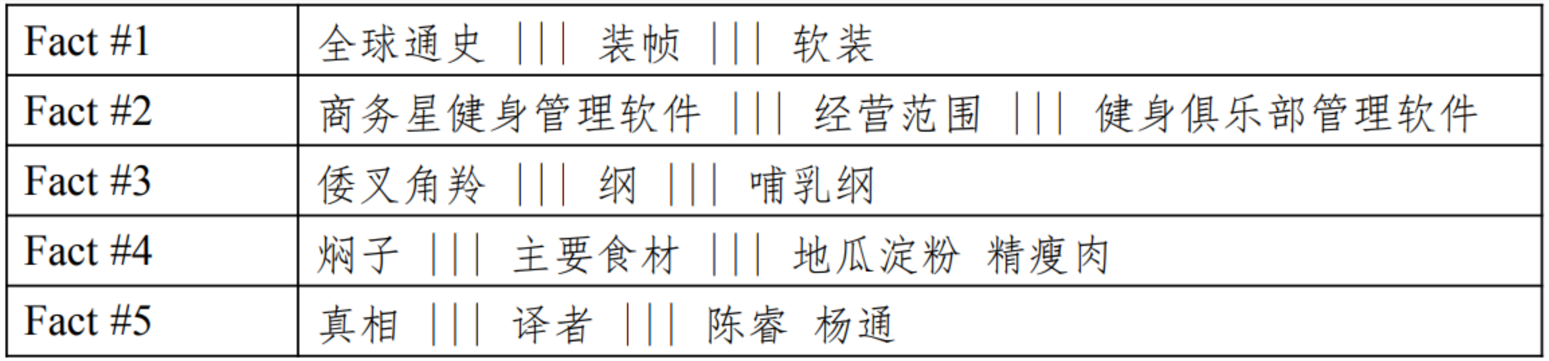
$$a_t = \sum_{i=1}^{L} \alpha_{ti} h_i$$

$$\alpha_{ti} = \frac{e^{g(s_t, h_i)}}{\sum_{j=1}^{N} g(s_t, h_j)} \qquad g(s_t, h_i) = \tanh(W_p h_i) \cdot \tanh(W_q s_t)$$

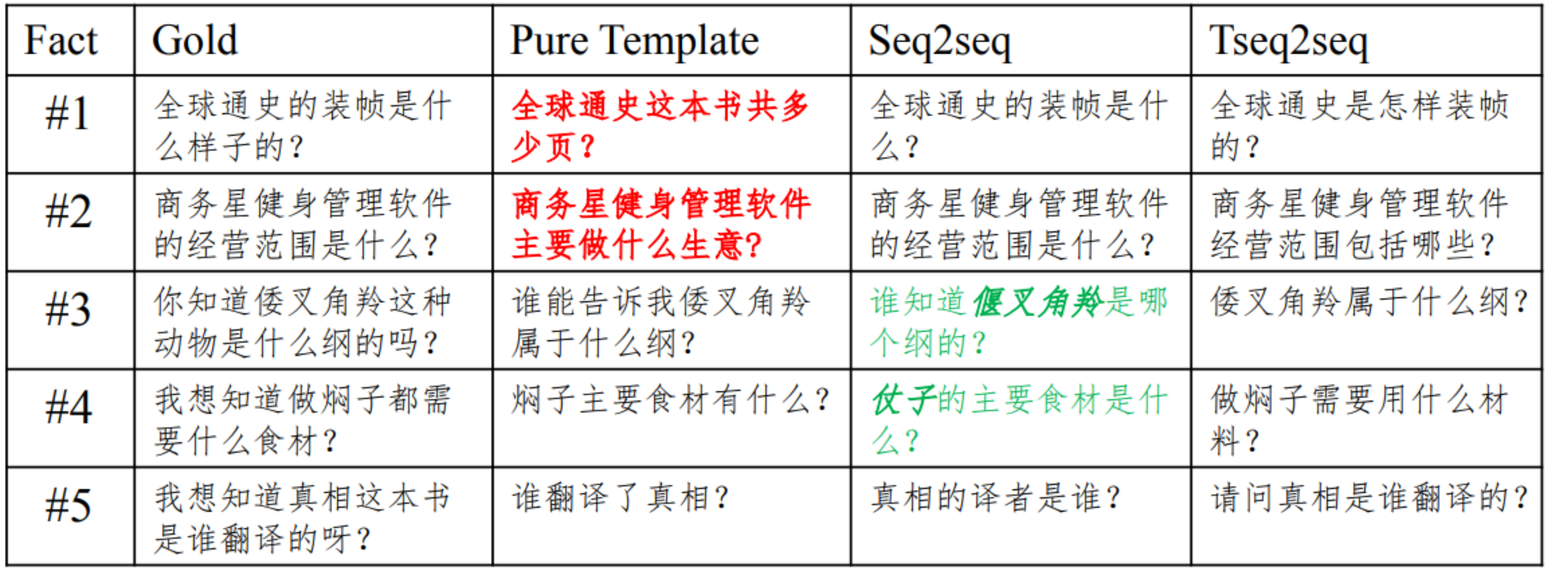

Given Triple: 于海 ||| 相关人物 ||| 吴冠中、张建中、爱新觉罗 • 溥铮
Generated Question: 请问与于海有关的人有谁?



Case Study



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？

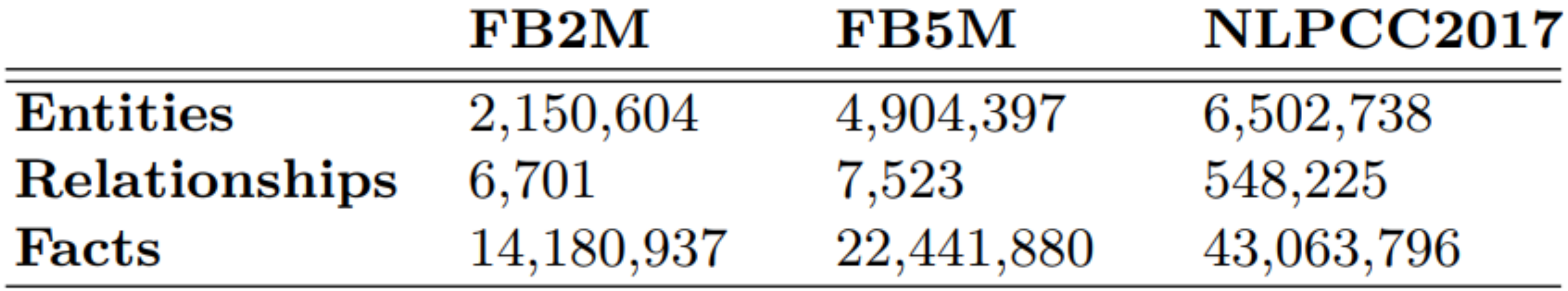
- ➢ Misleading questions generated by pure template-based method are marked in red.
- ➢ Questions that generates wrong subjects entities of the corresponding facts are marked in green

Outline

- Background
  - Graph knowledge base
  - Simple question
- KB-based question generation
  - Template Extraction
  - Template-based Seq2seq model
  - Case study
- Dataset & metrics
- Experiments
- Large-scale Chinese KBQA dataset
  - Triple Selection and Question Filtering
  - Dataset Analysis
- Future Work

Dataset and metrics

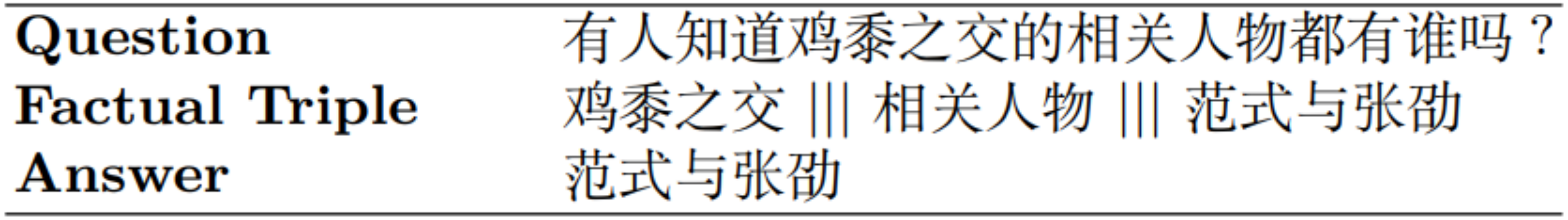
- Knowledge base: Chinese KB from NLPCC2017 KBQA challenge



	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

Statistics of the NLPCC2017 Chinese Knowledge Base

- Train & Test dataset: Training & Testing set from NLPCC2016 KBQA challenge （Train/dev/test: 11687/2922/9870）



Question	有人知道鸡黍之交的相关人物都有谁吗？
Factual Triple	鸡黍之交 ||| 相关人物 ||| 范式与张劭
Answer	范式与张劭

An instance of (Question, Triple, Answer) tuple

Dataset and metrics

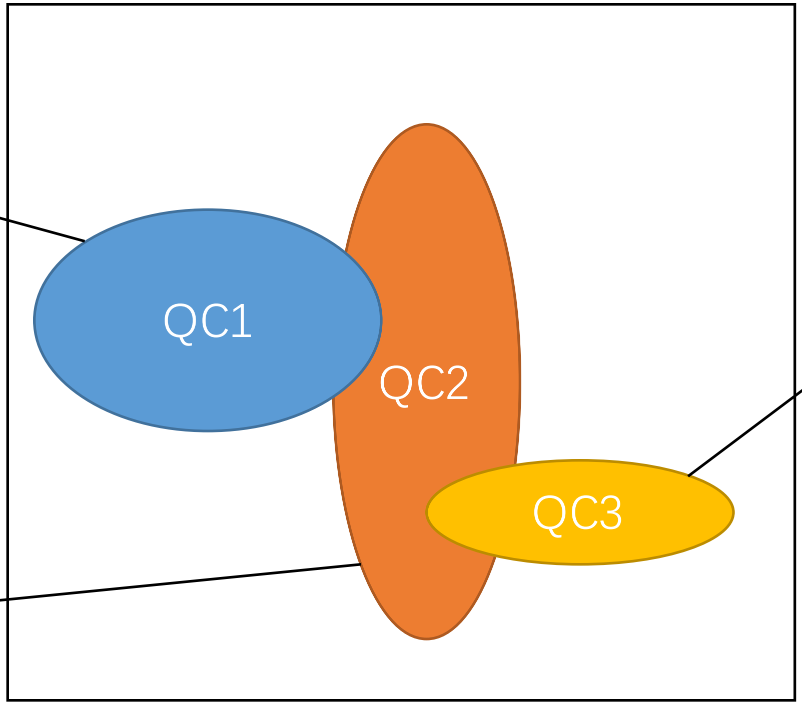
- ➢ Automatic evaluation: Bleu-4 and Rouge-4
- ➢ Human evaluation:
  1) Randomly select 100 generated questions.
  2) ask 2 experts to evaluate whether the question is understandable and answerable (good question or not).
  3) Use the ratio of good questions in the selected 100 questions as human evaluation score

Dataset and metrics

- Diversity evaluation: measure the diversity of generated questions with the same relationship (Question Cluster)

书名
你知道(SUB)的名字吗？
(SUB)的名字是什么？
(SUB)是什么？
知道(SUB)叫什么名字吗？

相关人物
(SUB)的相关人物都有谁？
你知道(SUB)与谁相关么？
有人知道(SUB)的相关人物吗？
(SUB)的相关人物是什么？



A set of question clusters

节目时长
(SUB)这个节目一般多长时间？
(SUB)的时长是多少?
(SUB)有多长？
(SUB)这个节目有多长时间？

Dataset and metrics

- Diversity evaluation: DIVERSE

For a question cluster Q_c $=\{q_1, q_2, \cdots, q_n\}$ and

Corresponding triple cluster $F_c = \{[t_1,\ t_2, \cdots, t_n], R, [o_1,\ o_2, \cdots, o_n]\}$

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

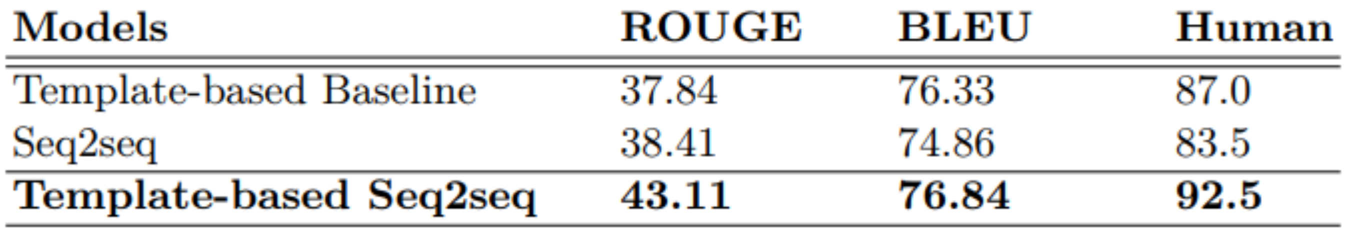
the smaller DIVERSE is, the more linguistically diverse the generated questions are.

Outline

- Background
  - Graph knowledge base
  - Simple question
- KB-based question generation
  - Template Extraction
  - Template-based Seq2seq model
  - Case study
- Dataset & metrics
- Experiments
- Large-scale Chinese KBQA dataset
  - Triple Selection and Question Filtering
  - Dataset Analysis
- Future Work

Experiment results

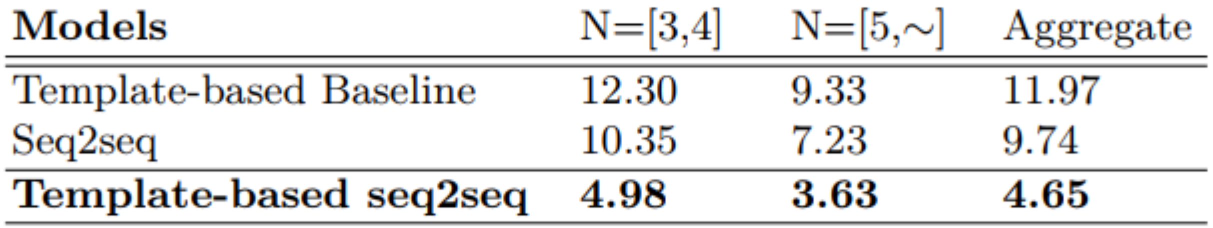
- Automatic & human evaluation



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**

Automatic and human evaluation performance of proposed models.

- Diversity evaluation



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**

Diversity evaluation of proposed models.
N equals the number of facts inside each cluster

Outline

- Background
    - Graph knowledge base
    - Simple question
- KB-based question generation
    - Template Extraction
    - Template-based Seq2seq model
    - Case study
- Dataset & metrics
- Experiments
- Large-scale Chinese KBQA dataset
    - Triple Selection and Question Filtering
    - Dataset Analysis
- Future Work

Triple selection and question filtering

- Triple selection

  for given triple $< h, r, t >$

  1. remove entity description from head entity

     万家灯火~~(林兆华李六乙导演话剧)~~

  2. choose head entities which have more than 5 relationship connections (to ensure the quality of questions)

Head entity with 6 relationship connections
水冷机箱 ||| 别名 ||| 水冷机箱
水冷机箱 ||| 中文名 ||| 水冷机箱
水冷机箱 ||| 缺点 ||| 普遍体积过大，操作不够简单
水冷机箱 ||| 类型 ||| 电脑内部发热部件散热的一种装置
水冷机箱 ||| 功能 ||| 它包括水冷散热系统和防尘机箱
水冷机箱 ||| 英文名 ||| Water-cooled chassis

✓

Head entity with 4 relationship connections
与幸福背道而驰 ||| 别名 ||| 与幸福背道而驰
与幸福背道而驰 ||| 中文名 ||| 与幸福背道而驰
与幸福背道而驰 ||| 作者 ||| 冂虚
与幸福背道而驰 ||| 小说进度 ||| 连载

✖

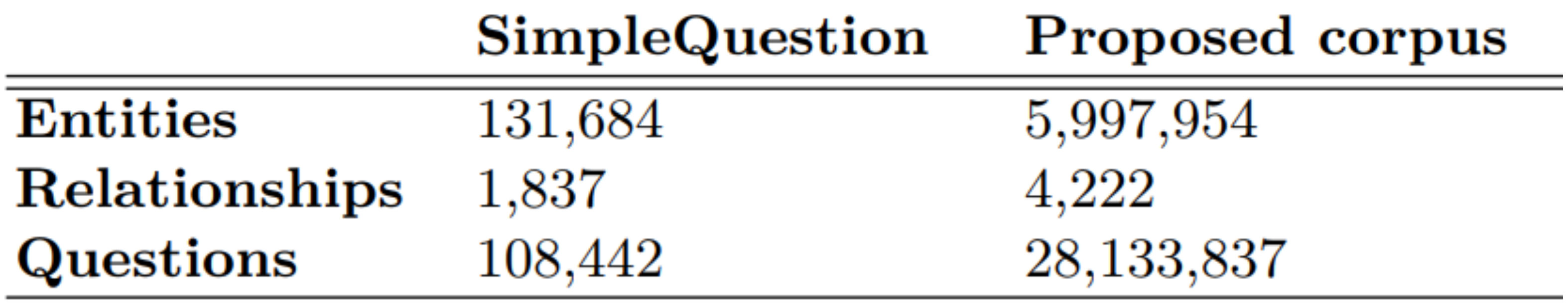
Triple selection and question filtering

- Question filtering
  1. Filter out questions with *UNK* token.
  2. Filter out questions with 2-gram or more repetition.

     谁知道~~知道~~再造幽冥进度怎么样了？
  3. Filter out questions whose length are longer than 50.

Dataset Analysis - quantitative analysis

- Statistics of proposed dataset



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

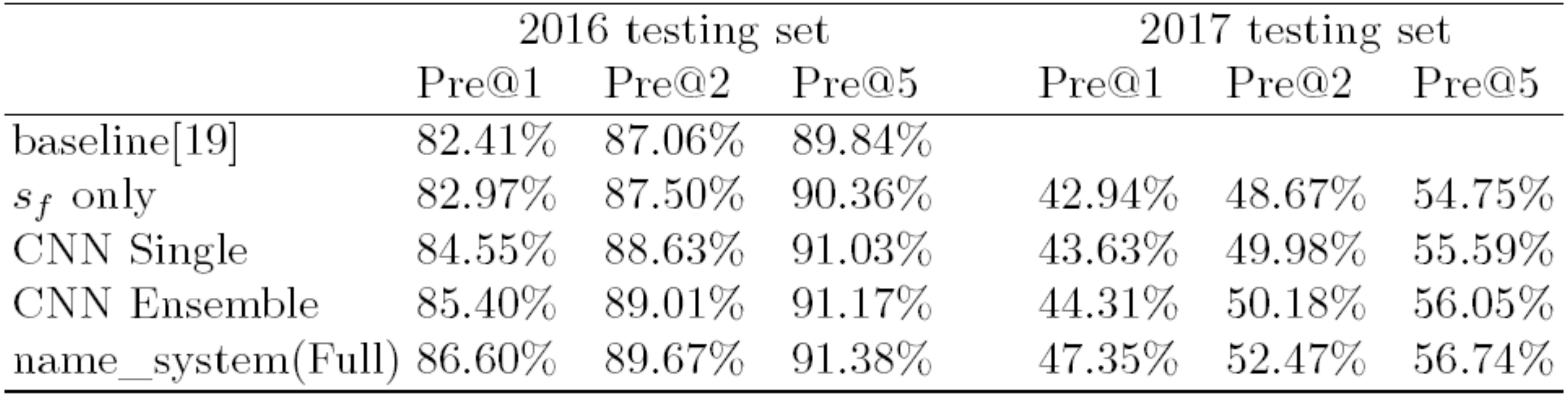
Dataset Analysis - quality analysis

- Performance of different models on the proposed dataset

we randomly select 21065 instances (question-answer pairs) from the proposed dataset and test the performance of three competitive models on the selected dataset.

Model	Dataset	Precision(%)	Recall(%)	F1(%)
(Lai et al.) KBQA Challenge winner	2016test	86.60	86.60	86.60
	2017test	47.23	47.23	47.23
	Our	**89.07**	**89.07**	**89.07**
System 1 In the challenge	2016test	76.55	76.55	76.55
	2017test	36.51	36.51	36.51
	Our	**78.25**	**78.25**	**78.25**
System 2 In the challenge	2016test	74.38	74.38	74.38
	2017test	31.46	31.46	31.46
	Our	**75.21**	**75.21**	**75.21**

Dataset Analysis - quality analysis



	2016 testing set			2017 testing set		
	Pre@1	Pre@2	Pre@5	Pre@1	Pre@2	Pre@5
baseline[19]	82.41%	87.06%	89.84%			
s_f only	82.97%	87.50%	90.36%	42.94%	48.67%	54.75%
CNN Single	84.55%	88.63%	91.03%	43.63%	49.98%	55.59%
CNN Ensemble	85.40%	89.01%	91.17%	44.31%	50.18%	56.05%
name_system(Full)	86.60%	89.67%	91.38%	47.35%	52.47%	56.74%

	Pre@1	Pre@2	Pre@5
Our	**89.77**	**90.15**	**90.45**

Future work

- **Compositional Question Generation** based on relational path

  Given a relational path e.g. name->marriage->height. Generating compositional questions like ‘how tall is <name>’ s wife/husband?’

- Question Generation based on **machine comprehension**

  Given an article or several paragraphs. Try to generate meaningful questions according to the context.

Learning to Ask: Neural Question Generation for Reading Comprehension Du et al.
Identifying Where to Focus in Reading Comprehension for Neural Question Generation Du et al.

Thanks for your listening!