Neural Generative Question Answering

Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, Xiaoming Li

https://www.ijcai.org/Proceedings/16/Papers/422.pdf

Overview

1. Introduction
   - background
   - Motivation

2. Task Description
   - Task & dataset

3. GENQA Model
   - Interpreter
   - Enquirer
   - Answerer

4. Experiment and Result

5. Related work

6. Conclusion and future work

1. Introduction

- QA aims at providing correct answers to the questions in natural language

Question & Answer	Triple (*subject*, *predicate*, *object*)
Q: *How tall is Yao Ming?* A: *He is 2.29m and is visible from space.*	(Yao Ming, height, 2.29m)
Q: *Which country was Beethoven from?* A: *He was born in what is now Germany.*	(Ludwig van Beethoven, place of birth, Germany)
Q: *Which club does Messi play for?* A: *Lionel Messi currently plays for FC Barcelona in the Spanish Primera Liga.*	(Lionel Messi, team, FC Barcelon)

- The answer could be retrieval-based or generation-based
- Obviously, generation-based is more likely to handle the **flexibility** and **diversity** of language
  - the answer is generated by a **neural network**
  - no need in building the system using **linguistic knowledge**, e.g., creating a **semantic parser**

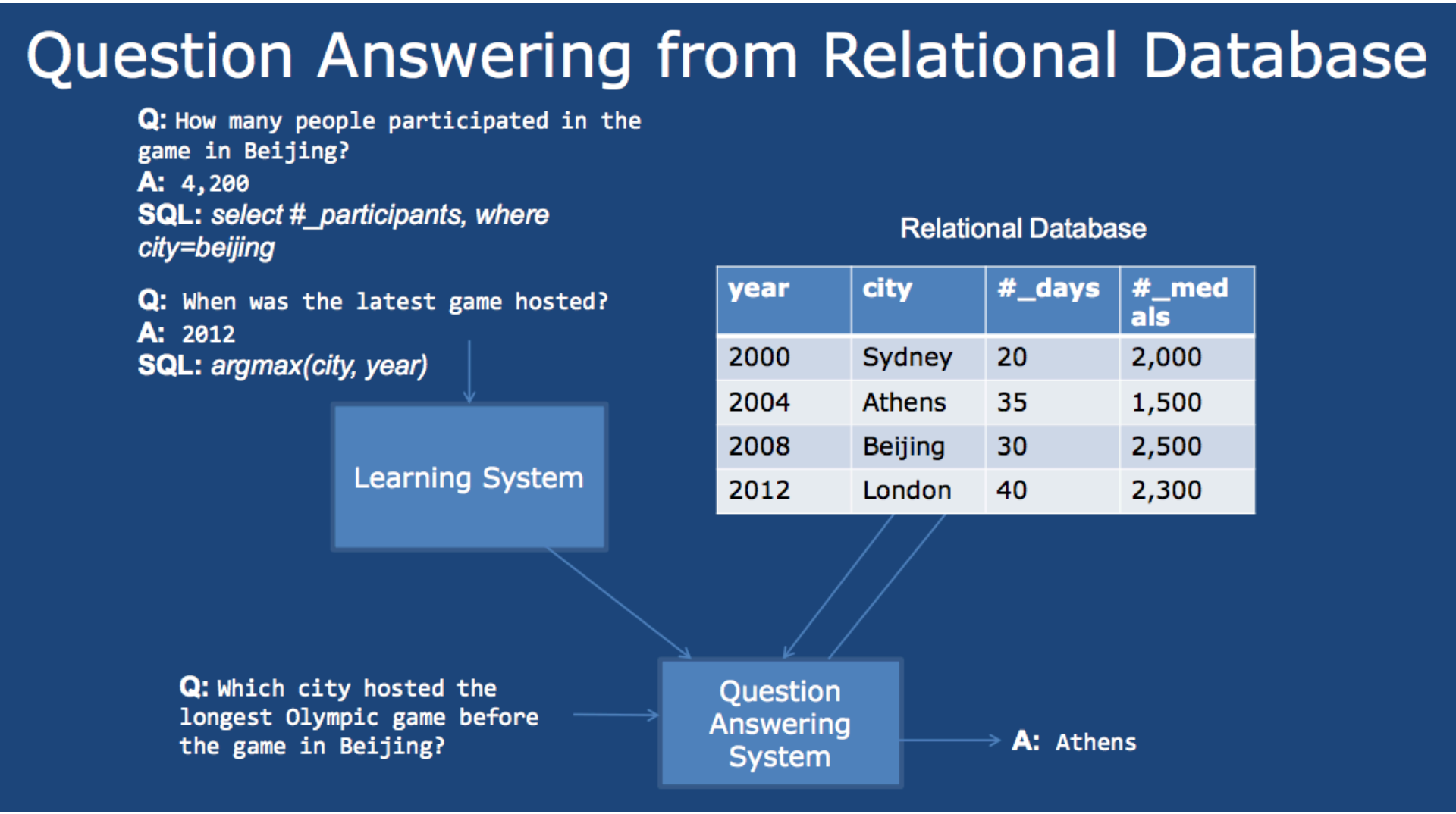
Motivation

- Limitation: It is practically impossible to store all the knowledge in a neural network to achieve a desired precision and coverage in real world QA

  - Why: the neural network is good at representing smooth and shared patterns, i.e., modeling the flexibility and diversity of language, but improper for representing discrete and isolated concepts

- Memory-based neural network models are proposed recently and It is hence a natural choice to connect a neural model for QA with a neural model of knowledge-base on an external memory

Motivation



Slide from http://www.hangli-hl.com/uploads/3/4/4/6/34465961/language_and_intelligence_summit.pdf

2. Task Description

- Input1: a sequence of words as input question
- Input2: knowledge-base, a huge amount of triples (*subject, predicate, object*)
- Output: another sequence of words as output answer

- This work focuses on *simple factoid question* QA, where the question is on *subject* and *predicate* of the triple and the answer is from *object*.

Question & Answer	Triple (*subject*, *predicate*, *object*)
Q: *How tall is Yao Ming?* A: *He is 2.29m and is visible from space.*	(Yao Ming, height, 2.29m)
Q: *Which country was Beethoven from?* A: *He was born in what is now Germany.*	(Ludwig van Beethoven, place of birth, Germany)
Q: *Which club does Messi play for?* A: *Lionel Messi currently plays for FC Barcelona in the Spanish Primera Liga.*	(Lionel Messi, team, FC Barcelon)

2. Task Description – dataset

Table 2: Statistics of the QA data and the knowledge-base.

Community QA	Knowledge-base	
#QA pairs	#entities	#triples
235,171,463	8,935,028	11,020,656

Table 3: Statistics of the training and test dataset for GENQA

Training Data		Test Data	
#QA pairs	#triples	#QA pairs	#triples
696,306	58,019	23,364	1,974

3. GENQA Model

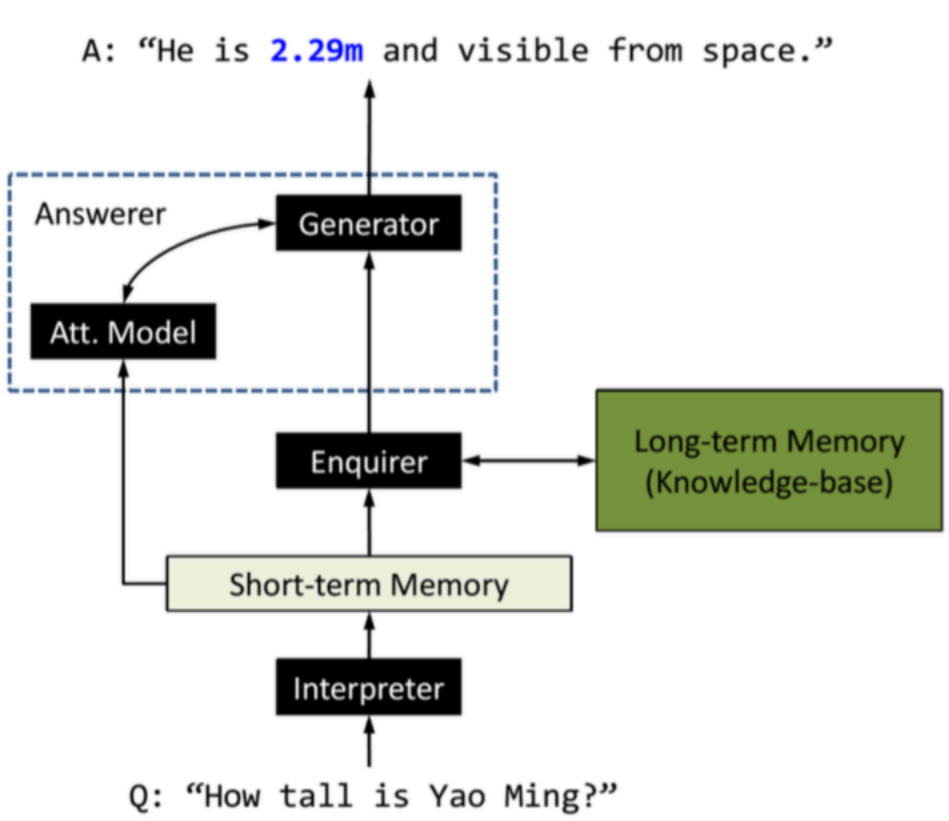


Figure 1: System diagram of GENQA.

- (i) Interpreter,
- (ii) Enquirer,
- (iii) Answerer.

3. GENQA Model – (i) Interpreter

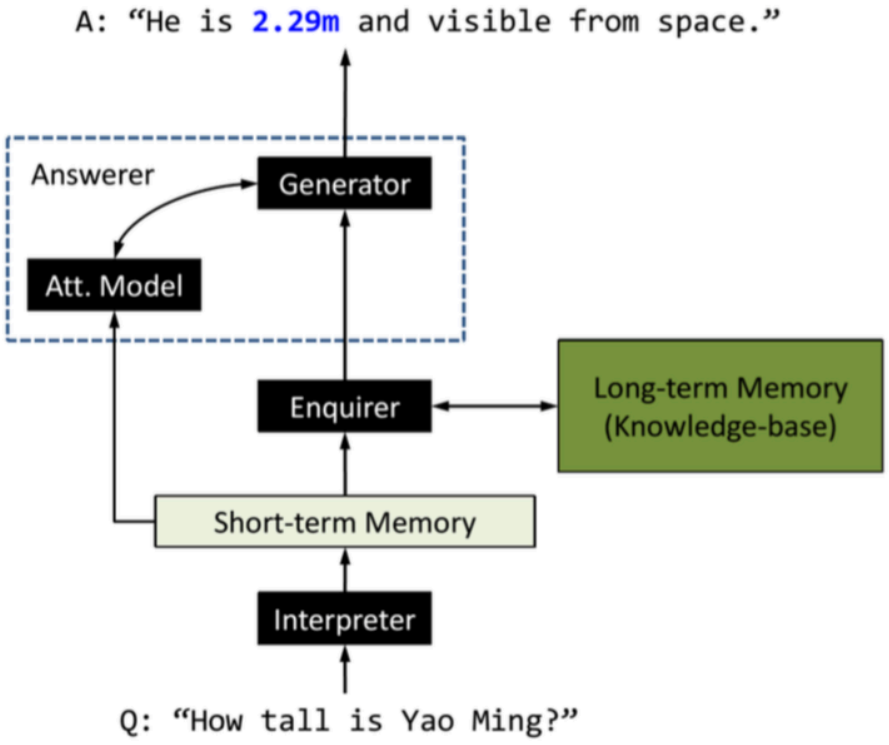


Figure 1: System diagram of GENQA.

Given the question represented as word sequence $Q = (x_1,...,x_T)$

Word Embedding $E=(e(x_1), ..., e(x_T))$

Bi-directional recurrent neural network

Hidden state $H_Q = (h_1, h_2, ... h_T)$

3. GENQA Model - (ii) Enquirer

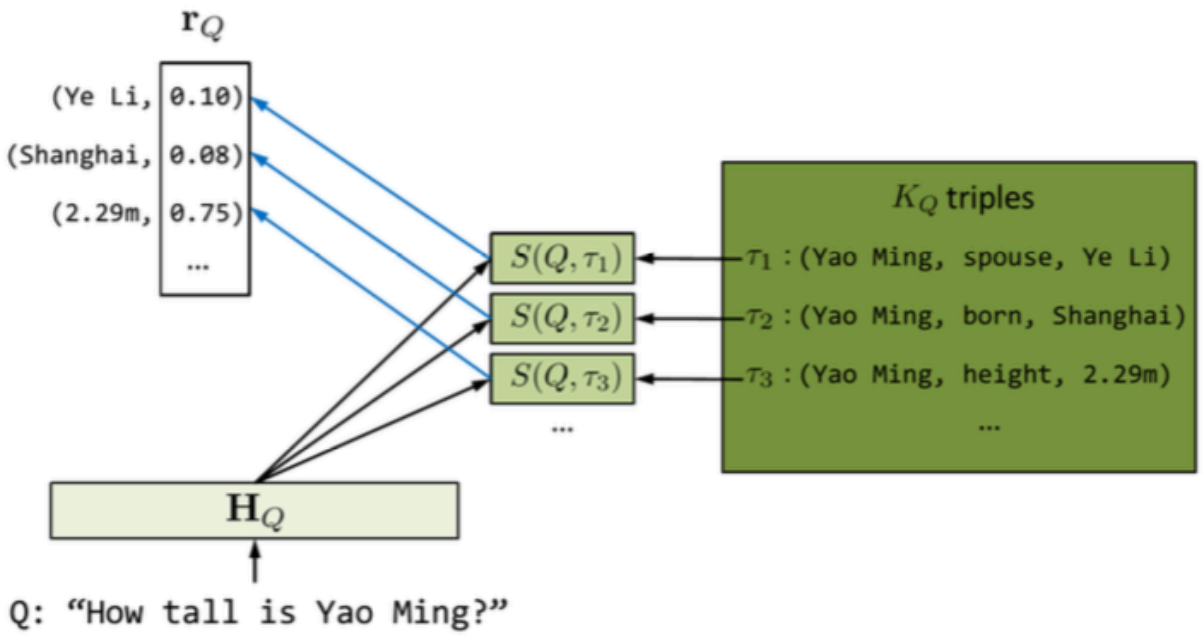


Figure 2: Enquirer of GENQA.

- Enquirer takes H_Q as input to interact with the knowledge- base in the long-term memory, retrieves relevant facts (triples) from the knowledge-base, and summarizes the result in a vector r_Q.

3. GENQA Model - (ii) Enquirer

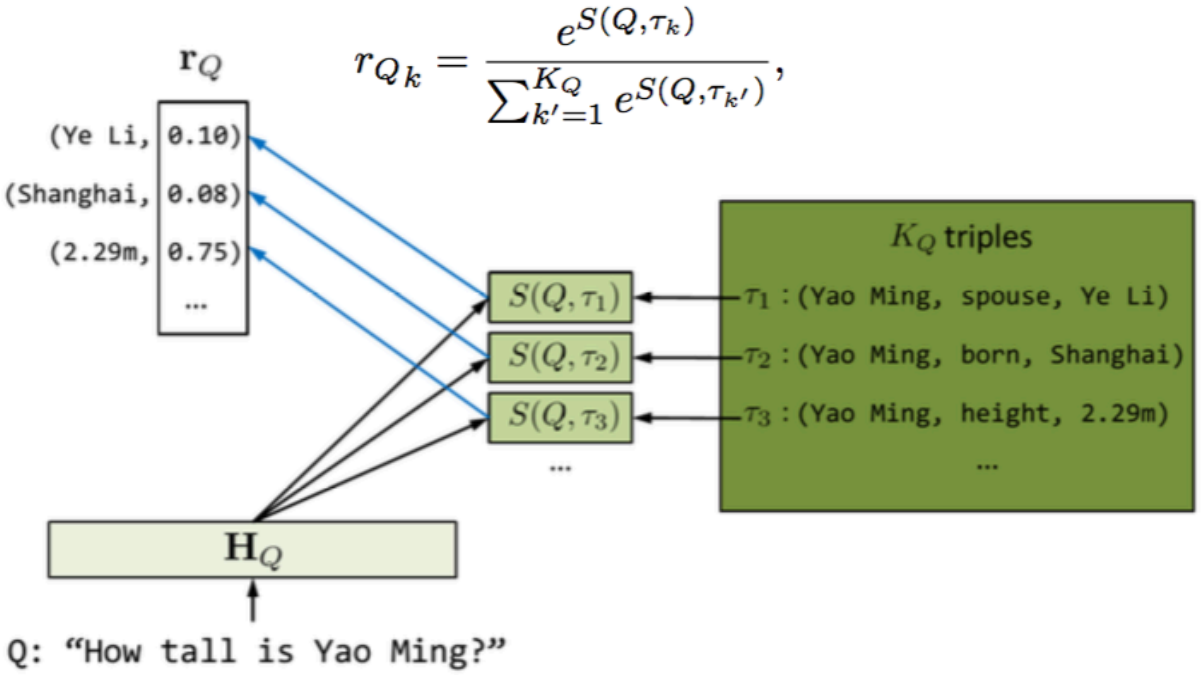


Figure 2: Enquirer of GENQA.

- where S (Q, T_k) denotes the **matching score** between question Q and triple T_k.

Matching score between question Q and triple T_k - *Bilinear Model* or *CNN-based Matching Model*

- Bilinear Model:
  - average of the word embedding vectors in H_Q as the representation of the question
  - average of the embeddings of its *subject* and *predicate* as the representation of the triple

$$\bar{S}(Q,\tau) = \bar{\mathbf{x}}_Q^\top \mathbf{M} \mathbf{u}_\tau$$

- CNN-based Matching Model:
  - question is fed to a convolutional layer followed by a max-pooling layer, and summarized as a fixed-length vector
  - average of the embeddings of its *subject* and *predicate* as the representation of the triple
  - Two above are concatenated as input to a multi-layer perceptron (MLP) to produce their matching score

$$\hat{S}(Q,\tau) = f_{\text{MLP}}([\hat{\mathbf{h}}_Q; \mathbf{u}_\tau])$$

3. GENQA Model - (iii) Answerer

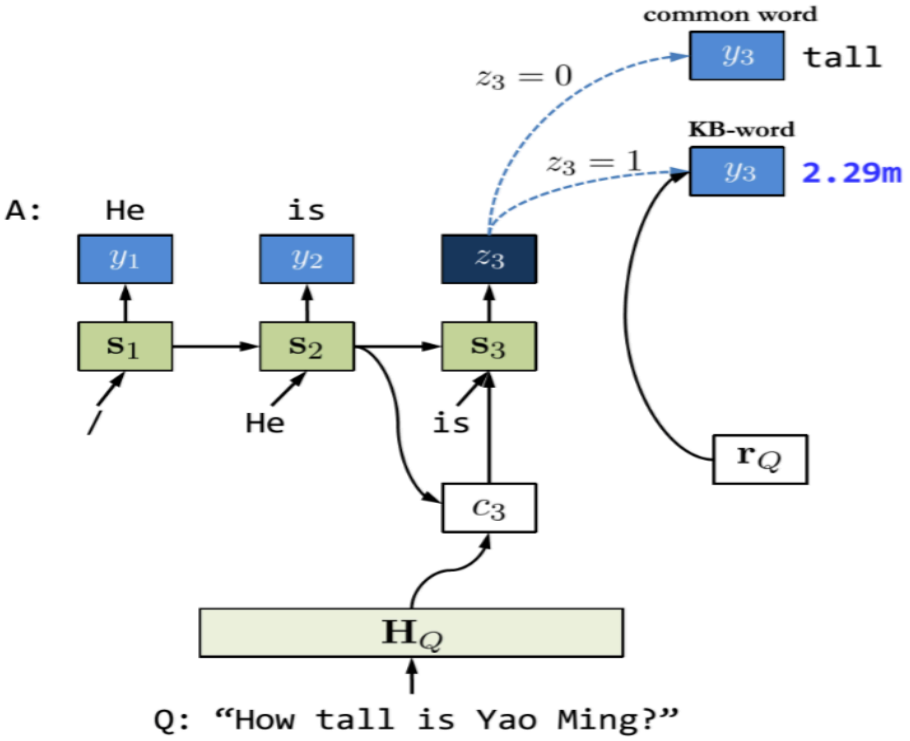


Figure 3: Answerer of GENQA.

- In generating the t-th word y in the answer, the probability is:

$$P(y_t|y_{t-1}) = P(z_t=0|s_t)P(y_t|y_{t-1}, z_t=0) + P(z_t=1|s_t)P(y_t|y_{t-1}, z_t=1)$$

- Where z_t indicates whether the t-th word is generated from a common vocabulary (for $z_t = 0$) or a KB vocabu- lary ($z_t = 1$).

Objective

- Minimizing the negative log-likelihood with regularization on all the parameters

$$\ell(\mathcal{D}, \theta) = -\sum_{i=1}^{N_{\mathcal{D}}} \log(Y^{(i)}|Q^{(i)}, \mathcal{T}_Q^{(i)}) + \lambda\|\theta\|_F^2.$$

- model is trained on machines with GPUs by using stochastic gradient descent with mini-batch.

4. Comparison Models

- **Neural Responding Machine (NRM)**: [Shang *et al.*, 2015] is a neural network based generative model specially designed for short-text conversation.
  - NRM does not access the knowledge-base during training and test, it actually remembers all the knowledge from the QA pairs in the model.

- **Retrieval-based QA**: the knowledge-base is indexed by an information retrieval system (we use Apache Solr)
  - At the test phase, a question is used as the query and the top-retrieved triple is returned as the answer. (this method cannot generate natural language answers).

- **Embedding-based QA**: as proposed by [Bordes *et al.*, 2014a; 2014b], the model is learnt from the question-triple pairs in the training data. The model learns to map ques-tions and knowledge-base constituents into the same embed- ding space, where the similarity between question and triple is computed as the inner product of two embedding vectors.

4. Results

- Evaluate the performance of the models in terms of
  - 1) accuracy, i.e., the ratio of correctly answered questions,
  - 2) the fluency of answers.

Table 4: Test accuracies

Models	**Test**
Retrieval-based QA	36%
NRM [Shang *et al.*, 2015]	19%
Embedding-based QA [Bordes *et al.*, 2014b]	45%
GENQA	47%
$\text{GENQA}_{\text{CNN}}$	**52%**

- Observation and discussion:
  1. NRM has the lowest accuracy, showing the lack of ability to accurately remember the answers and generalize to questions unseen in the training data.

4. Results for Accuracy

Table 4: Test accuracies

Models	Test
Retrieval-based QA	36%
NRM [Shang *et al.*, 2015]	19%
Embedding-based QA [Bordes *et al.*, 2014b]	45%
GENQA	47%
$\text{GENQA}_{\text{CNN}}$	**52%**

- Observation and discussion:

2. The retrieval-based method achieves a moderate accuracy, but like most string-matching methods it suffers from word mismatch between the question and the triples in the Knowledge base.

3. The embedding-based QA model achieves higher accuracy on test set, thanks to its generalization ability from distributed representations.

4. GENQA and $\text{GENQA}_{\text{CNN}}$ are both better than the competitors, and $\text{GENQA}_{\text{CNN}}$ Performs the best, which may be due to that convolution layer helps to capture salient features in matching.

4. Results for Fluency

- We make some empirical comparisons and find no significant differences between NRM and GENQA in terms of the fluency of answers.

- all the three generation-based models (two of them are ours) yield correct sentences in most of the time.

4. Case Study

ID	Questions	Generated Answers
1	三国演义谁写的? Who wrote the Romance of the Three Kingdoms?	明朝的罗贯中。 Luo Guanzhong in Ming dynasty.
2	项羽是哪国人? Which state does Xiang Yu belong to?	楚国的皇帝。 King of the Chu State.
3	还珠格格多少集? How many episodes does My Fair Princess have?	一共24集。 24 episodes in total.
4	泰戈尔是哪国作家? Which country is the writer Tagore from?	泰戈尔是印度著名的作家 。 Tagore is a famous Indian writer.
5	孙燕姿今年多大啊? How old is Stefanie Sun this year?	32岁，他出生于1978年7月23日。 Thirty-two, he was born on July 23, 1978.
6	《格列佛游记》的作者是谁? Who is the author of Gulliver's Travels?	法国作家斯威夫特。 The French writer, Swift.
7	怪物史莱克4什么时候上映呀? When will Shrek Forever After be released?	上映日期：梦工厂电影发行公司。 Release date: DreamWorks Pictures.
8	凤凰卫视是哪里的电视台? Which place is phoenix television located at?	总部在刘长乐。 The headquarters is at Liu Changle.

Figure 4: Examples of the generated answers by GENQA.

5. Related works

- Please refer to the Section V of the paper https://www.ijcai.org/Proceedings/16/Papers/422.pdf

6. Conclusion and future work

- In this paper the authors have proposed an end-to-end neural network model for generative question answering. The model is built on the encoder-decoder framework for seq2seq learning, while equipped with the **ability to query a knowledge-base**.

- Empirical studies show the proposed model is capable of generating natural and right answers to the questions by referring to the facts in the knowledgebase.

- Future work:
  - iterative question answering: a QA system that can interact with the user to confirm/clarify/answer her questions

  - question answering from complex knowledge base **(Knowledge Graph)**

End