IJCNLP-2017 Task 5: Multi-choice Question Answering in Examinations

Shangmin Guo[†], Kang Liu[†‡], Shizhu He[†], Zhuoyu Wei[†], Cao Liu[†‡] and Jun Zhao[†‡]
[†] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, China
[‡] University of Chinese Academy of Sciences, Beijing, 100049, China

Abstract

The IJCNLP-2017 Multi-choice Question Answering(MCQA) task aims at exploring the performance of current Question Answering(QA) techniques via the real-world complex questions collected from Chinese Senior High School Entrance Examination papers and CK12 website[1]. The questions are all 4-way multi-choice questions writing in Chinese and English respectively that cover a wide range of subjects, e.g. Biology, History, Life Science and etc. And, all questions are restrained within the elementary and middle school level. During the whole procedure of this task, 7 teams submitted 323 runs in total. This paper describes the collected data, the format and size of these questions, formal run statistics and results, overview and performance statistics of different methods.

1 Introduction

One critical but challenging problem in natural language understanding (NLU) is to develop a question answering(QA) system which could consistently understand and correctly answer general questions about the world. "Multi-choice Question Answering in Exams"(MCQA) is a typical question answering task that aims to test how accurately the participant QA systems could answer the questions in exams. All questions in this competition come from real examinations. We collected multiple choice questions from several curriculums, such as Biology, History, Life-Science, with a restrain that all questions are limited in the elementary and middle school level. For every question, four answer candidates are provided, where each of them may be a word, a value, a phrase or even a sentence. The participant QA systems are required to select the best one from these four candidates. Fig 1 is an example. To answer these questions, participants could utilize any public toolkits and any resources on the Web, but manually annotation is not permitted.

Peach trees have sweet-smelling blossoms and produce rich fruit. What is the main purpose of the flowers of a peach tree? (Answer is A.)
(A) to attract bees for pollination.
(B) to create flower arrangements.
(C) to protect the tree from disease.
(D) to feed migratory birds.

Figure 1: An example question from English Subset.

As for the knowledge resources, we encourage participants to utilize any resource on Internet, including softwares, toolboxes, and all kinds of corpora. Meanwhile, we also provide a dump of Wikipedia[2] and a collection of related Baidu Baike Corpus[3] under a specific license. These corpora and released questions are all provided in the XML format, which will be explained in section 2.2.

Main characteristics of our task are as follow:

- All the questions are from real word examinations.
- Most of questions require considerable inference ability.
- Some questions require a deep understanding of context.
- Questions from different categories have different characteristics, which makes it harder

[1] http://www.ck12.org/browse/

[2] https://www.wikipedia.org/

[3] https://baike.baidu.com/

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 34–40,
Taipei, Taiwan, November 27 – December 1, 2017. ©2017 AFNLP

for a model to have a good performance on all kinds of questions.

- It concentrates only on the textual content, as questions with figures and tables are all filtered out.

2 Task and Data Description

All questions in MCQA consist of 2 parts, a question and 4 answer candidates, without any figure or table. The participant systems are required to select the only right one from all candidates.

2.1 Languages and Subjects

In order to explore the influence of diversity of questions, we collect questions from seven subjects in two languages, including an English subset and a Chinese subset. The subjects of English subset contain biology, chemistry, physics, earth science and life science. And the subjects of Chinese subset only contain biology and history. The total number of questions is 14,447.

2.2 Format

All questions in our dataset are consisted by the following 7 parts:

1. ID, i.e. the identical number of a specific question;
2. Question, i.e. the question to be answered;
3. Option A, i.e. the content of first answer candidate;
4. Option B, i.e. the content of second answer candidate;
5. Option C, i.e. the content of third answer candidate;
6. Option D, i.e. the content of fourth answer candidate;
7. Correct Answer No., i.e. the number of the correct candidate(0, 1, 2 and 3, which corresponds to four options respectively).

Take a question in Figure 1 for example. Roles of every part are as follow:

1. ID: wb415;
2. Question: ”Peach trees have sweet-smelling blossoms and produce rich fruit. What is the main purpose of the flowers of a peach tree?”;
3. Option A: ”to attract bees for pollination.”;
4. Option B: ”to create flower arrangements.”;
5. Option C: ”to protect the tree from disease.”;
6. Option D: ”to feed migratory birds.”;
7. Correct Answer No.: 0.

It needs to be specified that we exclude the Correct Answer No. in the validation and test set.

2.3 Data Size

The dataset totally contains 14,447 multiple choice questions. In detail, English subset contains 5,367 questions and Chinese subset contains 9,080 questions. We randomly split the dataset into Train, Validation and Test sets. And more detail statistics is showed in Table 1.

	Train	Valid	Test	Total
English Subset				
Biology	281	70	210	561
Chemistry	775	193	581	1549
Physics	299	74	224	597
Earth Science	830	207	622	1659
Life Science	501	125	375	1001
English Total	2686	669	2012	5367
Chinese Subset				
Biology	2266	566	1699	4531
History	2275	568	1706	4549
Chinese Total	4541	1134	3405	9080
Complete Dataset				
Total	7227	1803	5417	14447

Table 1: The statistics of dataset.

2.4 English Subset

We collected all the downloadable quiz from CK12 and only reserved 5367 4-way multi-choice questions with their tags which are also the basis of classifying the questions. For every subject, we randomly separate questions into 3 parts, train set, valid set and test set with 50%, 12.5% and 37.5% questions respectively.

2.5 Chinese Subset

As questions in Senior High School Entrance Examination(SHSEE) differs among different cities, we collected questions in SHSEE from as many cities as we can. After filtering out the questions containing more information than textual content, the answers of left questions were labeled by human. Finally, we got 4,531 questions in Biology and 4,549 questions in History. For every subject, we randomly separate questions into 3 parts, train

English Subset					
Biology	Chemistry	Physics	Earth Science	Life Science	All English
30%	21.24%	25.68%	31.88%	40%	29.45%
Chinese Subset					
Biology		History		All Chinese	
34.81%		54.42%		44.63%	
All Datasets					
39%					

Table 2: The detail performance of the baseline method.

set, valid set and test set with same ratio stated above.

2.6 Evaluation

This challenge employs the accuracy of a method on answering questions in test set as the metric, the accuracy is calculated as follow.

$$\text{Accuracy} = \frac{n_{correct}}{N_{total}} \times 100\%$$

where $n_{correct}$ is the number of correctly answered questions and N_{total} is the total number of all questions.

To automatically evaluate the performance of QA systems, we built a web-site for participants to submit solutions for valid and test data set and get accuracy immediately on the page.

2.7 Baseline

We employ a simple retrieval based method as a baseline, and it is implemented based on Lucene[4] which is an open-source information retrieval software library. We employ the method to build reverse-index on the whole Wikipedia dump[5] for English questions and on the Baidu Baike corpus[6] for Chinese questions.

This method scores pairs of the question and each of its option, the detail steps are shown as follows.

- concatenate a question with an option as the query;
- use Lucene to search relevant documents with the query;
- score relevant documents by the similarity between the query q and the document d, noted as $Sim(q, d)$;
- choose at most three highest scores to calculate the score of the pair of the question and the option as

$$score(q, a) = \frac{1}{n} \sum_{1}^{n} Sim(q, d)$$

where $n \leqslant 3$ and $if n = 0, score(q, a) = 0$;

All questions and options are preprocessed by Stanford CoreNLP[7]. The detail result of the baseline on the validation set is shown in Table 2.

3 Participation

7 teams as shown in Table 3 were participated in the end.

Team Name	Affliation
YNU-HPCC	Yunnan University
CASIA-NLP	Institute of Automation, Chinese Academy of Sciences
Cone	Dublin City University
G623	Yunnan University
JU_NITM	Jadavpur University
TALN	Universit de Nantes
QA_challenge	Free

Table 3: Active Participating Teams (as of Aug. 31, 2017)

The details of participation of different language subsets are listed in the following Table 4.

4 Submission

In order to avoid the situation that participants submit different permutation of answers to sniff the correct answer labels, we limited the times that a team can submit their solutions. Before the release of test set, a team can submit no more than 5 solutions for valid set in 24 hours. After the release of test set, a team can submit as many as 30 solutions

[4] http://lucene.apache.org/
[5] https://dumps.wikimedia.org/
[6] http://www.nlpr.ia.ac.cn/cip/ijcnlp/baidubaike_corpus.html
[7] https://stanfordnlp.github.io/CoreNLP/

Team Name	Language
YNU-HPCC	Both
CASIA-NLP	Chinese
Cone	English
G623	English
JU_NITM	English
TALN	English
QA_challenge	English

Table 4: Language Selection of Teams (as of Aug. 31, 2017)

for valid set per 24 hours, but no more than 5 solutions for test set in 24 hours. Finally, we got 323 runs in total, in which there are 219 runs for valid set and 104 runs for test set.

5 Results

In our evaluation system, only the best performance of participants were reserved. The detail results of every subset is listed in the following subsections.

5.1 All Questions

There is only one team, "YNU-HPCC", that took the challenge of both English subset and Chinese subset. And, the performance of their system is listed in Table 5.

5.2 English Subset

Totally, there are 5 teams that only took the challenge of English subset and details of their performance are listed in the Table 6.

5.3 Chinese Subset

There are 1 team that only took the challenge of Chinese subset and their performance is listed in the Table 7.

6 Overview of Participant Systems

6.1 YNU-HPCC, An Attetion-based LSTM

YNU-HPCC (Yuan et al., 2017) proposed an attention-based LSTM(AT-LSTM) model for MCQA. According to them, this model can easily capture long contextual information with the help of an attention mechanism. As illustrated in Figure 2, LSTM layer takes the vector representations of question and answers as input and then calculates out the hidden vectors which are the input of attention layer to calculate the weight vector α and weighted hidden representation r. Finally, an softmax layer takes r as input to select the right answer.

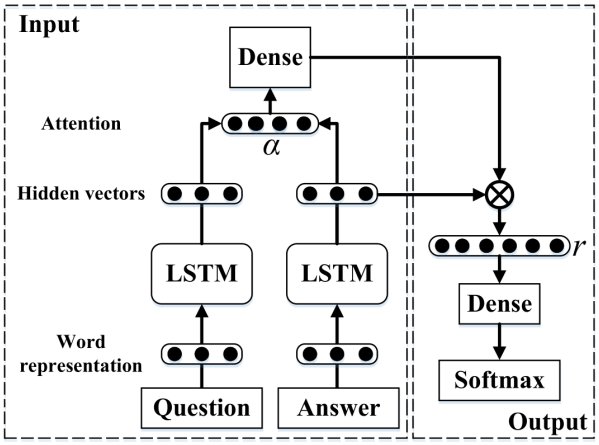

Figure 2: Architecture of AT-LSTM proposed by team YNU-HPCC(Yuan et al., 2017).

6.2 CASIA-NLP, Internet Resources and Localization Method

Based on the phenomenon that many web pages containing answers of the questions in MCQA, CASIA-NLP (Li and Kong, 2017) crawled on Internet and analyzed the content in these pages. When analyzing these pages, they use a localization method to locate the positions of sentences that have same meaning of questions in MCQA by merging a score given by edit distance that evaluates the structural similarity and a cosine score given by a CNN network that evaluates the semantic similarity. Finally, the system can analyze answers to find out the right one. The overview of the system is illustrated in Figure 3 and the CNN network they used is demonstrated in Figure 4.

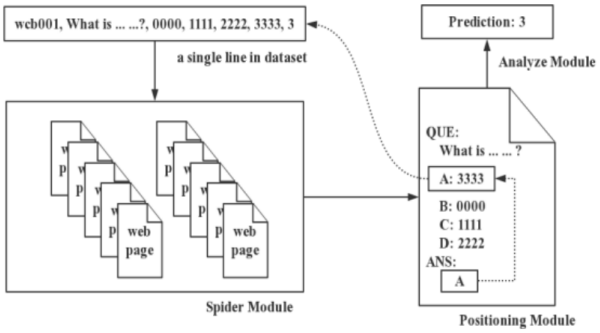

Figure 3: Overview of CAISA-NLP's system (Li and Kong, 2017). Communication between modules is indicated by arrows.

Valid Set			Test Set		
English	Chinese	All	English	Chinese	All
34.5%	46.5%	42.1%	35.5%	46.5%	42.3%

Table 5: Performance of YNU-HPCC (as of Aug. 31, 2017)

Team Name	Valid Set	Test Set
Cone	48.7%	45.6%
G623	42.8%	42.2%
JU_NITM	40.7%	40.6%
TALN	34.7%	30.3%
QA_challenge	21.5%	N/A

Table 6: Performance on English Subset (as of Aug. 31, 2017)

Team Name	Valid Set	Test Set
CASIA-NLP	60.1%	58.1%

Table 7: Performance on English Subset (as of Aug. 31, 2017)

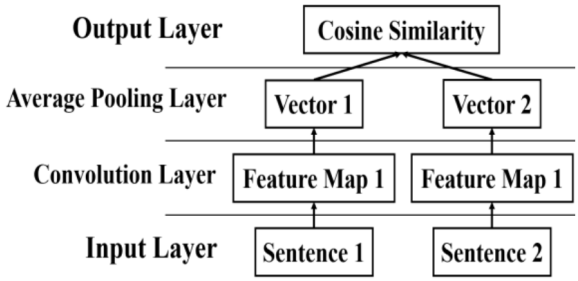

Figure 4: Convolutional architecture used in CASIA-NLP's system (Li and Kong, 2017).

6.3 Cone, Wikipedia and Logistic Regression

The system of Cone (Dzendzik et al., 2017), a team from ADAPT Centre, based on a logistic regression over the string similarities between question, answer, and additional text. Their model is constructed as a four-step pipeline as follow.

1. Preprocessing cleaning of the input data;
2. Data selection relative sentences are extracted from Wikipedia based on key words from question;
3. Feature Vector Concatenation for every question, a feature vector is built as a concatenation of similarities between the answer candidates and sentences obtained in the previous step;
4. Logistic Regression a logistic regression over the feature vector.

The features they employed includes term frequencyinverse document frequency (Tf-IDf) metric, character n-grams (with n ranging from 1 to 4), bag of words,and windows slide (a ratio between answer and substrings of extracted data). While their model is trained in two ways, combining training over all domains and separate model training from each domain, the later one got the best performance.

6.4 G623, A CNN-LSTM Model with Attention Mechanism

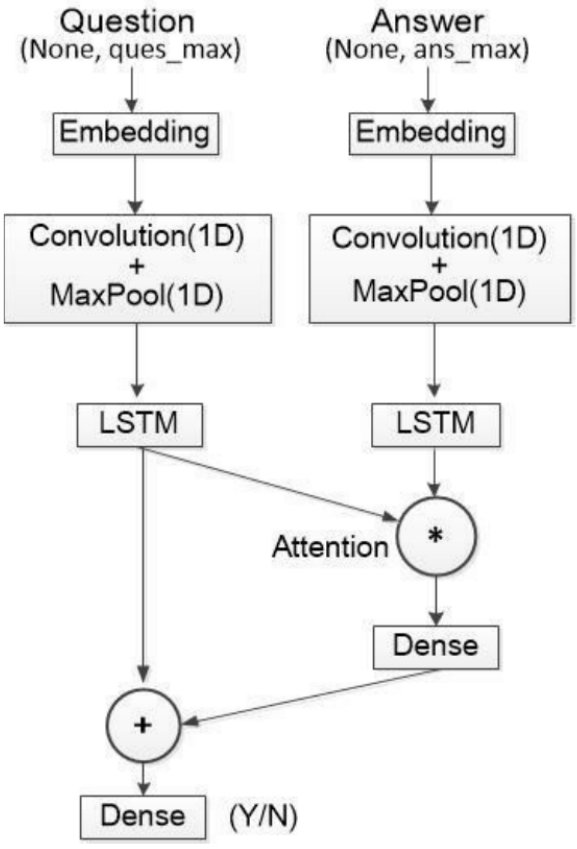

Figure 5: Architecture of the model proposed by G623(Min et al., 2017).

The system of G623 (Min et al., 2017) combined CNN with LSTM network and took into account the attention mechanism. Fistly , question

and answer pairs are fed into a CNN network and produce joint representations of these pairs which are then fed into a LSTM network. The two separate vector representations of question and answer are then calculated to generate the weight vector by dot multiplication. Finally, a softmax layer is applied to classify the join representations with the help of attention weight. The diagram of their system is illustrated in Figure 5.

6.5 JU_NITM, Complex Decision Tree

To handle the questions in MCQA, JU_NITM (Sarkar et al., 2017) built a complex decision tree classifier using word embedding features to predict the right answer. The overview of the whole system is demonstrated in Figure 6. In distributed semantic similarity module, they trained a word embedding dictionary containing 3 million words in 300-dimensional space on GoogleNews. Then, a complex decision tree is used to select the right answer in step2, classification.

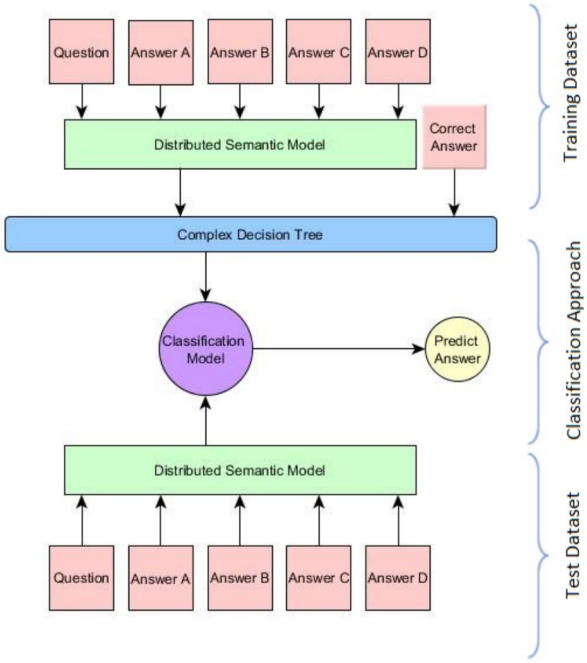

Figure 6: System Framework proposed by JU_NITM(Sarkar et al., 2017).

6.6 TALN, MappSent

Mappsent is proposed in a previous work of TALN, (Hazem et al., 2017). To adapt to the characteristics of MCQA, they retrofitted MappSent model in two different ways(Hazem, 2017). The first approach illustrated in Figure 7 is to follow the same procedure as the question-to-question similarity task, i.e. using anatated pairs of questions and their corresponding answers to build the mapping matrix. The second approach illustrated in Figure 8 tends to keep the strong hypothesis of sentence pairs similarity. They built two mapping matrices, one that represent similar question pairs and ther other one to represent similar answers pairs. For a give test question, the system can extracted the most similar quesiont in the training data and select the candidate with highest similarity score as correct answer.

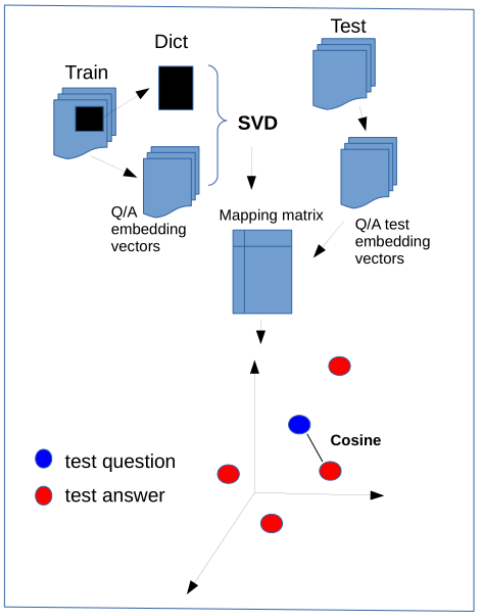

Figure 7: Fist adaptation of MappSent(Hazem, 2017).

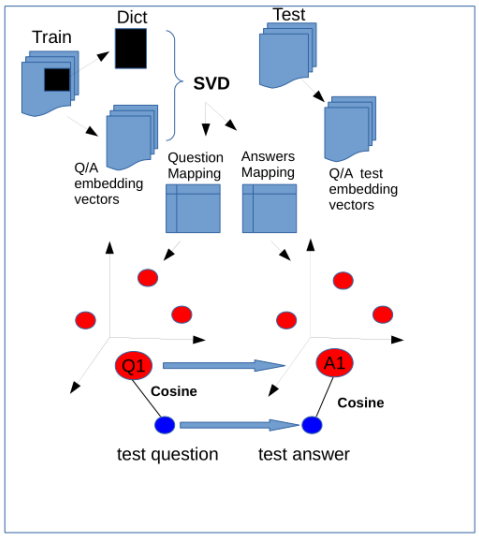

Figure 8: Second adaptation of MappSent(Hazem, 2017).

7 Conclusions

We described the overview of the Multi-choice Question Answering task. The goal is exploring the performance of current Question Answering(QA) techniques via the real-world complex questions collected from Chinese Senior High School Entrance Examination(SHSEE) papers and CK12 website. We collected 14,447 questions covering 2 language in 7 different subjects. 7 teams submitted 323 runs in total. We describe the collected data, the format and size of these questions, formal run statistics and results, overview and performance statistics of different methods in this paper.

Acknowledgments

Our thanks to participants. This task organization was supported by the Natural Science Foundation of China (No.61533018) and the National Basic Research Program of China (No. 2014CB340503). And this research work was also supported by Google through focused research awards program.

References

Daria Dzendzik, Alberto Poncelas, Carl Vogel, and Qun Liu. 2017. A similarity-based logistic regression approach to multi-choice question answering in an examinations shared task. In *IJCNLP-2017, Shared Task 5*.

Amir Hazem. 2017. A textual similarity approach applied to multi-choice question answering in examinations shared task. In *IJCNLP-2017, Shared Task 5*.

Amir Hazem, Basma el amel Boussaha, and Niclas Hernandez. 2017. Mappsent: a textual mapping ap-proach for question-toquestion similarity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2017*.

Changliang Li and Cunliang Kong. 2017. Answer localization for multi-choice question answering in exams. In *IJCNLP-2017, Shared Task 5*.

Wang Min, Qingxun Liu, Peng Ding, Yongbin Li, and Xiaobing Zhou. 2017. A cnn-lstm model with attention for multi-choice question answering in examinations. In *IJCNLP-2017, Shared Task 5*.

Sandip Sarkar, Dipankar Das, and Partha Pakray. 2017. Ju nitm: A classification approach for answer selection in multi-choice question answering system. In *IJCNLP-2017, Shared Task 5*.

Hang Yuan, You Zhang, Jin Wang, and Xuejie Zhang. 2017. Using an attention-based lstm for multi-choice question answering in exams. In *IJCNLP-2017, Shared Task 5*.