The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Hierarchical Attention Flow for Multiple-Choice Reading Comprehension

Haichao Zhu,[♮*] Furu Wei,[§] Bing Qin,[♮†] Ting Liu[♮]
[♮]SCIR, Harbin Institute of Technology, China
[§]Microsoft Research, Beijing, China
{hczhu, qinb, tliu}@ir.hit.edu.cn, fuwei@microsoft.com

Abstract

In this paper, we focus on multiple-choice reading comprehension which aims to answer a question given a passage and multiple candidate options. We present the hierarchical attention flow to adequately leverage candidate options to model the interactions among passages, questions and candidate options. We observe that leveraging candidate options to boost evidence gathering from the passages play a vital role in this task, which is ignored in previous works. In addition, we explicitly model the option correlations with attention mechanism to obtain better option representations, which are further fed into a bilinear layer to obtain the ranking score for each option. On a large-scale multiple-choice reading comprehension dataset (i.e. the RACE dataset), the proposed model outperforms two previous neural network baselines on both RACE-M and RACE-H subsets and yields the state-of-the-art overall results.

Introduction

In this paper, we study the task of multiple-choice reading comprehension, in which every question is accompanied with four candidate options and only one is correct. Figure 1 shows an example. Comparing to questions from previous reading comprehension tasks (Hermann et al. 2015; Hill et al. 2016; Onishi et al. 2016; Rajpurkar et al. 2016), multiple-choice questions put no constraints on the answers to be exact match spans of the reference passage. Instead the candidate options are human generated sentences, which may not appear in the passage.

RACE (Lai et al. 2017) and MCTest (Richardson, Burges, and Renshaw 2013) are two representative benchmark datasets generated by human for multiple-choice reading comprehension. Yin, Ebert, and Schütze (2016) use convolutional neural network to build the representation at different hierarchical levels with attention mechanism. Trischler et al. (2016) propose to conduct the match in multiple parallel perspectives with the hierarchical structure, and a training technique for the proposed neural model to converge on MCTest. Lai et al. (2017) adapt two strong neural models GA Reader (Dhingra et al. 2017) and Stanford AR (Chen, Bolton, and Manning 2016) as neural network baselines for the large-scale RACE dataset. Specifically, the models gather and summarize the passage evidence only with the question, then conduct the match between the evidence and candidate options.

Passage: ... In 1993, New York State ordered stores to charge money on beverage containers. Within a year, consumers had returned millions of aluminum cans and glass and plastic bottles. Plenty of companies were eager to accept the aluminum and glass as raw material for new products, but because few could figure out what to do with the plastic, much of it wound end up ...
Question: What regulation was issued by New York State concerning beverage containers?

A. A fee should be charged on used containers for recycling.

B. Throwaways should be collected by the state for recycling.

C. Consumers had to pay for beverage containers and could get their money back on returning them.

D. Beverage companies should be responsible for collecting and reusing discarded plastic soda bottles.
Answer: C

Figure 1: An example multiple-choice reading comprehension question.

Inspired by Yin, Ebert, and Schütze (2016) and Trischler et al. (2016), we present the neural network based hierarchical attention flow, depicted in Figure 2, to adequately leverage candidate options to model the word level and sentence level interactions among passages, questions and candidate options. The attention flow is organized in the following hierarchical order. We use a bi-directional recurrent neural network (BiRNN) to encode passage sentences, question and candidate options separately. Then, the word-level attention layer builds the question-aware passage sentence and the candidate option representation. Next, the sentence context encoder models temporal context between passage sentences with a BiRNN. Afterwards, the sentence-level attention layer gathers evidence from the passage with candidate options in addition to the question, and models option correlations for better option representation. Finally, the bi-

[*]Contribution during internship at Microsoft STC Asia.
[†]Corresponding author.
Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

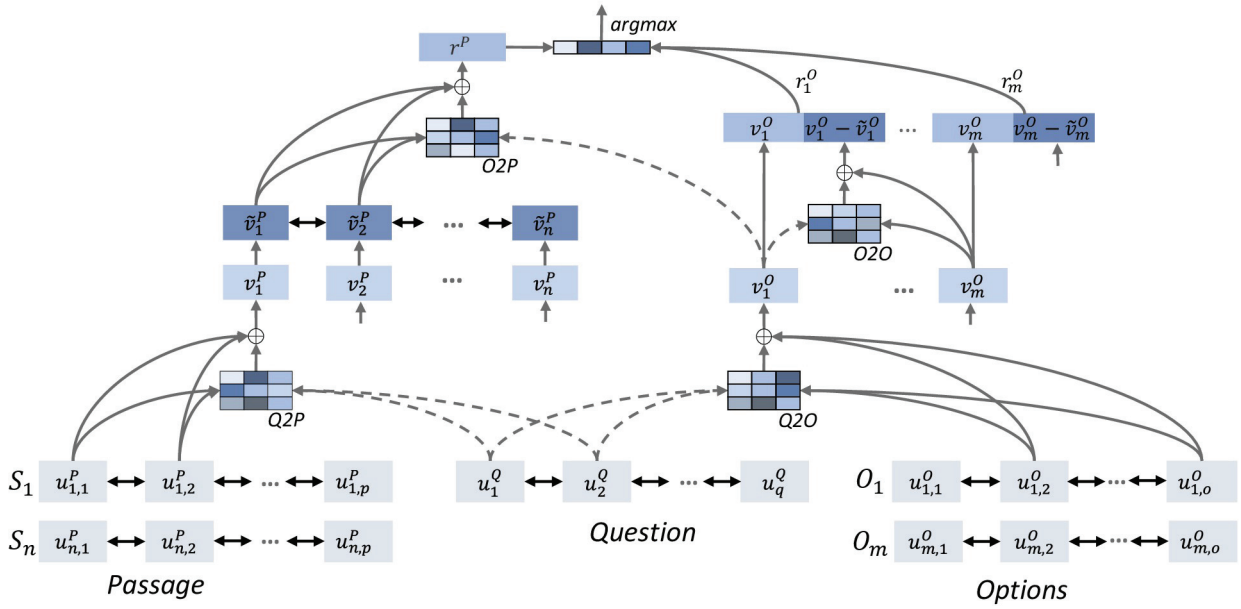

Figure 2: Hierarchical Attention Flow overview. The superscripts P, Q and O correspond to passage, question and option respectively. The text blocks marked with the same color are at the same hierarchical level. $\oplus$ represents the weighted sum operation. Finally, *argmax* function outputs the option with highest score as the answer.

linear layer computes the ranking score for each option. The key contributions of our work are three-fold.

Firstly, we propose to use the candidate options to gather evidence from the passage. In two neural baselines adapted by Lai et al. (2017), only the question is used to gather evidence. While in multiple-choice reading comprehension, the questions are sometimes not informative and clear enough to guide evidence gathering. Especially to the questions with blanks or about general purpose, where the partial evidence gathered by the question alone may mislead the model to wrong predictions. However, the candidate options can provide additional information for clarifying the question's intent. Therefore we leverage the question-aware option representation to boost evidence gathering in our model. In this way, the model utilize the candidate options information in addition to the question information to gather more sufficient evidence to distinguish the answer from the distraction options.

Secondly, to further exploit candidate options, we leverage option correlations, which are ignored by previous works, as additional information to the original independent option representation. Existing models (Richardson, Burges, and Renshaw 2013; Sachan et al. 2015; Narasimhan and Barzilay 2015; Wang et al. 2015; Smith et al. 2015; Trischler et al. 2016; Yin, Ebert, and Schütze 2016; Lai et al. 2017) for multiple-choice reading comprehension score each candidate option independently. We compare options with each other to model option correlations. The correlation is encoded into a vector representation with sentence-level attention, then concatenated to the independent option representation. Thus our proposed model scores each option with considering other options.

Lastly, we conduct extensive experiments on the RACE dataset, in which RACE-M and RACE-H corresponds to middle school and high school difficulty respectively. Our proposed model outperforms previous strong neural network baselines by 1.9% in overall accuracy and achieves state-of-the-art results.

Model

We describe the details of the hierarchical attention flow in the order of the appearance in Figure 2 bottom-up from left to right.

Word Context Encoder

Given a question $\{w_t^Q\}_{t=1}^q$ of q words, a passage $\{w_{i,t}^P\}_{i=1}^n$ of n sentences and candidate options $\{w_{i,t}^O\}_{i=1}^m$ of m candidates. We first map every word w to its respective d-dimensional vector e via an embedding matrix $E \in \mathbb{R}^{|V| \times d}$, where V represents the vocabulary. Then we apply a BiRNN to encode word context from both sides. Here we choose Gated Recurrent Unit (GRU) (Cho et al. 2014) as the recurrent network building block:

$$u_t^Q = BiGRU_Q(u_{t-1}^Q, e_t^Q) \quad (1)$$

$$u_{i,t}^P = BiGRU_P(u_{i,t-1}^P, e_{i,t}^P) \quad (2)$$

$$u_{i,t}^O = BiGRU_O(u_{i,t-1}^O, e_{i,t}^O) \quad (3)$$

Hence, we obtain the context-aware word representation u^Q for question, u_i^P for i-th sentence of passage and u_i^O for i-th candidate option.

Attention Flow

In our model, interaction between two components is used to emphasize and organize relevant information accordingly. We adopt the same attention mechanism to every interaction. In this section, we first describe the attention mechanism in detail and then explain various interactions.

Attention Mechanism We adopt the attention mechanism similar to Cui et al. (2016), but we use the bilinear function (Luong, Pham, and Manning 2015) to compute the relevance score instead of the dot product. Given two inputs $X \in \mathbb{R}^{m \times k}$ of m k-dimensional vector and $Y \in \mathbb{R}^{n \times l}$ of n l-dimensional vector, $att(X, Y \mid W_{xy})$ outputs the attention weight vector a, where a_j indicates the attention weight of Y_j according to X.

More specifically, we first compute the matching matrix $A \in \mathbb{R}^{m \times n}$ with the bilinear term $W_{xy} \in \mathbb{R}^{k \times l}$, where $A_{i,j}$ is the relevance score of X_i and Y_j. Then we apply row-wise softmax function to get the attention weight matrix s:

$$A_{i,j} = X_i W_{xy} Y_j \tag{4}$$

$$s_{i,j} = \frac{exp\left(A_{i,j}\right)}{\sum_j^n exp\left(A_{i,j}\right)} \tag{5}$$

where s_i indicates the relevance weight over Y in the view of X_i. Then we combine the view of each vector of X by averaging the attention weight matrix column-wise to get the final attention weight vector a:

$$a_j = \frac{1}{m} \sum_{i=1}^{m} s_{i,j} \tag{6}$$

Question-to-Passage (Q2P) Word-level Attention Words within a sentence are not equally important, the significances may change in tune with the question. To get the vector representation of the passage sentence, we apply word representation of the question u^Q to attend words u_i^P of the *i*-th passage sentence. We utilized the output of every time step of the question BiGRU, instead of the output of the last time step, which is used in (Lai et al. 2017). Then we get the question-aware representation v_i^P of the *i*-th sentence at sentence level:

$$a = att(u^Q, u_i^P \mid W_{qp}) \tag{7}$$

$$v_i^P = \sum_t a_t u_{i,t}^P \tag{8}$$

Question-to-Option (Q2O) Word-level Attention Lai et al. (2017) represent candidate option with the last hidden state of option BiGRU, which produces the question free vector representation. While the meaning of the option is more interpretable when combined with the question. Therefore we incorporate the question information to compose option's word representation into a fixed-size vector with attention mechanism at word level. In a similar way, we get the question-aware representation v_i^O of the *i*-th candidate option:

$$a = att(u^Q, u_i^O \mid W_{qo}) \tag{9}$$

$$v_i^O = \sum_t a_t u_{i,t}^O \tag{10}$$

Sentence Context Encoder The order of sentences within a passage matters just as the order of words within a sentence. But we process the passage sentences in parallel and produce the context independent sentence representation v_i^P. To encode sentence context, similar to word context modeling, we apply another BiGRU on top of v_i^P. Then the sentence context is encoded into $\tilde{v}_i^P$:

$$\tilde{v}_i^P = BiGRU_S(\tilde{v}_{i-1}^P, v_i^P) \tag{11}$$

Option-to-Passage (O2P) Sentence-level Attention In reading comprehension task, the passage generally contains abundant information about the events, places, etc. When a question only concerns about a certain aspect, the irrelevant parts of the passage could be redundancy and noise. To avoid the negative effect of redundant information, Lai et al. (2017) summarize the whole passage into a single vector as evidence with question to passage attention. It is a popular setting in cloze-style reading comprehension models (Hermann et al. 2015; Hill et al. 2016). In our model, we leverage the question-aware candidate option representation v^O to boost evidence gathering from passage sentences with attention mechanism at sentence level. In attention computation procedure, each candidate option assigns higher weights to its corresponding evidence sentences. Then we average the assigned weights of each passage sentence to obtain the final attention weight. Different from previous works, we implicitly incorporate the question information, which is encoded by Q2O word-level attention. Finally, the candidate options, along with the question, summarize the evidence into a fixed-size vector:

$$a = att(v^O, \tilde{v}^P \mid W_{op}) \tag{12}$$

$$r^P = \sum_i a_i \tilde{v}_i^P \tag{13}$$

Option Correlations The option representation v_i^O is generated by Q2O word attention, which is aware of the question. But such representation is independent of other options and no comparison information between options is encoded. To model the option correlations, we compare the candidate options with attention mechanism. In case of the option being compared to itself, we set the diagonal of the attention weight matrix to zero. $s_{i,j}$ indicates the relevance score of the *j*-th option to the *i*-th option, and merging operation is not necessary like before. Inspired by Chen et al. (2017), we model the option correlations with difference $v_i^O - \tilde{v}_i^O$, which then concatenated to the independent option representation as the enhancement:

$$A_{i,j} = v_i^O W_{oo} v_j^O \tag{14}$$

$$s_{i,j} = \frac{\mathbb{1}\left(i \neq j\right) exp\left(A_{i,j}\right)}{\sum_j \mathbb{1}\left(i \neq j\right) exp\left(A_{i,j}\right)} \tag{15}$$

$$\tilde{v}_i^O = \sum_j s_{i,j} v_j^O \tag{16}$$

$$r_i^O = [v_i^O; v_i^O - \tilde{v}_i^O] \tag{17}$$

Answer Prediction

We follow Lai et al. (2017) to compute the matching score s_i of i-th option against the summarized evidence r^P with the bilinear function and the probability p_i of being correct. The model predicts the answer with the argmax function:

$$s_i = r^P W_p r_i^O \tag{18}$$

$$p_i = \frac{exp\,(s_i)}{\sum_j exp\,(s_j)} \tag{19}$$

$$ans = argmax_i\,(p_i) \tag{20}$$

We train the network to minimize the negative log probability of ground truth option.

Experiments

Dataset

Large-scale ReAding Comprehension Dataset From Examinations (RACE) is a multiple-choice reading comprehension dataset. RACE-M and RACE-H correspond to middle school and high school difficulty level. All questions contain four candidate options with only one correct option. RACE contains 27,933 passages and 97,687 questions in total, 5% as development set and 5% as test set. Table 1 shows separation of the dataset. Table 2 is the statistics about the average number of words and sentences of the passage, question and candidate options.

		Train	Dev	Test
RACE-M	#passages	6,409	368	362
	#questions	25,421	1,436	1,436
RACE-H	#passages	18,728	1,021	1,045
	#questions	62,445	3,451	3,498
RACE	#passages	25,137	1,389	1,407
	#questions	87,866	4,887	4,934

Table 1: Dataset separation of training, development and test sets of the RACE dataset.

	RACE-M	RACE-H	RACE
#w/p	249.9	374.9	342.9
#s/p	17.2	19.2	18.7
#w/q	10.1	11.4	11.0
#w/o	4.9	6.8	6.3

Table 2: Data statistics of RACE-M, RACE-H and RACE. #w/p and #s/p represent the average number of words and sentences in the passage. #w/q and #w/o are the average length of the question and option. Training, development and test sets share the similar statistics.

Implementation Details

Following Lai et al. (2017), we combine RACE-M and RACE-H together as training set and development set. We tokenize the passages, questions and options into sentences and words with tokenizer from Natural Language Toolkit[1]. We use Tensorflow[2] to implement our model. To train the model, we adopt stochastic gradient descent with ADAM optimizer (Kingma and Ba 2015), with initial learning rate 0.001. Gradients are clipped in L2-norm to no larger than 10. A mini-batch of 32 samples is used to update the model parameter per step. We keep 50,000 most frequent words in training set as vocabulary and add a special token *UNK* for out-of-vocabulary *(OOV)* words. We initialize word embeddings with 300D pre-trained case-sensitive Glove (Pennington, Socher, and Manning 2014) embeddings[3], which are further updated in training phase. The hidden state size of all GRU network is 128. We apply dropout(Srivastava et al. 2014) to word embeddings and BiGRU's outputs with a drop rate of 0.4.

Results

	RACE-M	RACE-H	RACE
Random†	24.6	25.0	24.9
Sliding Window†	37.3	30.4	32.2
GA Reader (*100D*)†	43.7	44.2	44.1
GA Reader (*300D*)‡	42.4	44.5	43.9
Stanford AR (*100D*)†	44.2	43.0	43.3
Stanford AR (*300D*)‡	44.9	43.7	44.1
Ours (*100D*)	**46.2***	44.1	44.7*
Ours *(300D)*	45.0	**46.4***	**46.0***

Table 3: Accuracy on test set of RACE-M, RACE-H and RACE. † indicates the results from (Lai et al. 2017) which are trained with 100D pre-trained Glove word embeddings, ‡ indicates the results that we get by running the published code (Lai et al. 2017) of GA Reader and Stanford AR with 300D pre-trained Glove word embeddings.

Accuracy is the only metric to evaluate the model performance on RACE. We report the accuracy of our model and several baselines on test set. Our model outperforms previous best neural network baseline, GA Reader (100D), on both RACE-M and RACE-H subset and yields the state-of-the-art overall accuracy.

In Table 3, GA Reader(Dhingra et al. 2017) and Stanford AR (Chen, Bolton, and Manning 2016) are two very strong neural models on cloze-style reading comprehension, and adapted as neural baselines by Lai et al. (2017) along the release of RACE. To make the results more comparable and reduce the impact of parameter count, we train two neural baselines with 300D word embeddings, and with the same code used in previous 100D neural baselines. We also train our model with 100D word embeddings. The result shows that our model outperforms neural baselines under the same word embedding size. Moreover, our model with 100D word embeddings even outperforms 300D neural baselines in overall accuracy.

[1] http://www.nltk.org/

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

[3] https://nlp.stanford.edu/projects/glove/

1	In 1993, New York State ordered stores to charge money on beverage containers.
2	Within a year, consumers had returned millions of aluminum cans and glass and plastic bottles.
3	Plenty of companies were eager to accept the aluminum and glass as raw material for new products, but because few could figure out what to do with the plastic, much of it wound end up buried in landfills.
4	The problem was not limited to New York.
5	Unfortunately, there were too few uses for second-hand plastic.
6	Today, one out of five plastic soda bottles is recycled in the United States.
7	The reason for the change is that now there are dozens of companies across the country buying discarded plastic soda bottles and turning them into fence post, paint brushes, etc.
8	As the New York experience shows, recycling involves more than simply separating valuable materials from the rest of the rubbish.
9	A discard remains a discard until somebody figures out how to give it a second life – and until economic arrangements exist to give that second life value.
10	Without enough markets to take in materials collected for recycling, throwaways actually reduce prices for used materials.
11	Fewer landfill space and rising costs for burying and burning rubbish are forcing local governments to look more closely at recycling.
12	In many areas, the East Coast especially, recycling is already the least expensive waste-management choice.
13	For every ton of waste recycled, a city avoids paying for its disposal, which, in parts of New York, amounts to savings of more than $100 per ton.
14	Recycling also stimulates the local economy by creating jobs and reduces the pollution control and energy costs of industries that make recycled products by giving them a better raw material.

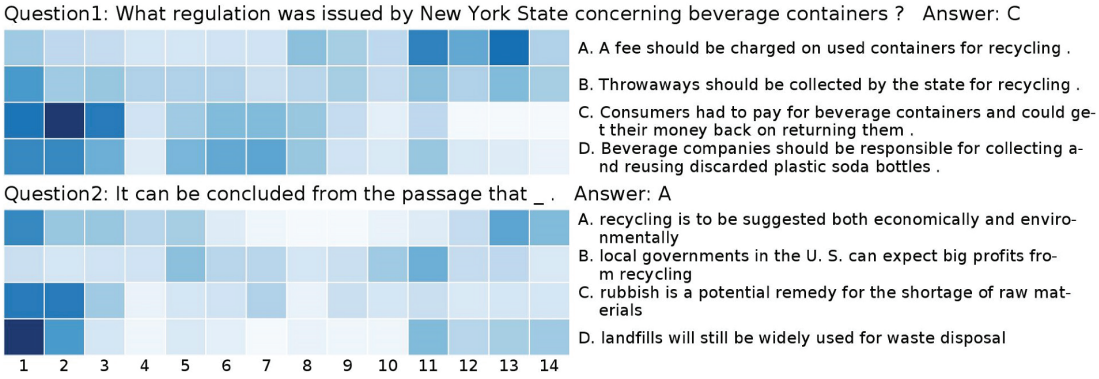

Figure 3: Attention weight matrix visualization of evidence gathering. Reference passage sentences are numbered in order. Darker color indicates the higher weight. Each row represents the attention weight of a option over passage sentences.

Ablation Study

To evaluate how different components contribute to the model's performance, we conduct ablation tests on development set, and results are illustrated in Table 4. We first investigate the influence of sentential temporal relationships by removing sentence context encoder layer. The accuracy is slight influenced and even increases 0.2 point on RACE-M subset. While this is because the evidence summarization over the passage sentences weaken the effect of sentence context encoder layer, which summarizes and encodes context. Then we remove O2P attention and replace options with the question to gather evidence, which is similar to (Lai et al. 2017). The large accuracy drop shows that incorporating candidate options into evidence gathering contributes most to the model's improvement. At last, to verify the effectiveness of option correlations, we directly conduct the match on the independent option representation produced by Q2O word-level attention. The 0.7 point overall accuracy drop reveals that option correlations do strengthen option representation. The above ablation tests results validate the necessity of fully utilizing candidate options for multiple-choice reading comprehension.

	RACE-M/H	RACE
Full Model	45.3/**47.9**	**47.2**
- Sentence Context Encoder	**45.5**/47.6	46.9
- O2P Attention	43.7*/47.2	46.2*
- Option Correlations	44.5/47.4	46.5

Table 4: Layer Ablation on the development set.

Discussion

Evidence Gathering and Option Correlations

To investigate how candidate options boost evidence gathering from the passage, we visualize the attention weight matrix in O2P attention. In Figure 3, the darker color indicates higher weight. The attention weight matrices show that the evidence related to each option scatters in the passage.

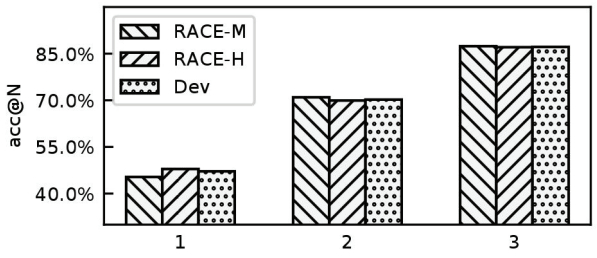

Figure 4: Statistics analysis of the answer ranking position. Accuracy corresponds to the relaxed requirement that as long as the model ranks the answer in top-N.

The merged attention weights over passage sentences summarize necessary information as much as possible. While the question may miss key evidence, especially to the question that does not contain evident indicative words or phrase, just like the second question in Figure 3. Only in combination with the candidate options, the model can get the hint to gather evidence, which emphasize the sentences about "state, money, recycling, landfill, disposal, raw material".

Option correlations are also useful for some cases. We take the first question in Figure 3 as an example. The model without option correlations chooses the wrong option A with a slightly higher score than option C. By incorporating option correlations, the model chooses the correct option C with a extremely high score 0.987.

Top-N Accuracy

To further investigate our model beyond the overall accuracy, we also statistically analyze the ranking score of the answer on development set. Since every question corresponds to only one correct answer, we settle to accept the predication as correct as long as the answer ranked in top-N, and compute the the accuracy (acc@N) accordingly. In Figure 4, acc@2 and acc@3 reach overall accuracies of 70.2% and 87.2% respectively. The lead over the expectation of random guess illustrates the effectiveness of our proposed model from another perspective and reveals the potential improvement can be achieved by reranking.

Difficulty Gap between RACE-M and RACE-H

RACE-M subset is collected from middle school exams, and RACE-H subset corresponds to high school exams. Thus RACE-M is easier in advance. The data statistics also show that the length of the passage, question and candidate options of RACE-M subset are shorter. And RACE-M subset has a smaller vocabulary than RACE-H subset. The difference on the passage length mainly reflects on sentence length, which is narrowed down by the hierarchical structure to minimal difference on sentence count. In terms of the vocabulary, most words in both subsets are covered by the vocabulary. And these affect accuracies on two subsets to be consistent with their prior difficulties.

Related Work

Large-scale Datasets

Large-scale datasets stimulate significant advance in reading comprehension research. According to whether the answers are restricted to an exact match span of the reference passage, we can classify existing datasets into two categories. CNN/Daily Mail (Hermann et al. 2015), Children's Book Test (CBT) (Hill et al. 2016) and Who Did What (WDW) (Onishi et al. 2016) are the automatically generated cloze-style large-scale datasets, in which the answer is a word (often a named entity) of the passage. The answer in SQuAD (Rajpurkar et al. 2016) is a continuous span that is labeled by human to guarantee high quality, instead of a word. RACE (Lai et al. 2017) and MS MARCO (Nguyen et al. 2016) are two large-scale datasets fall into another category that the answers may not appear in the reference passage. This is closer to the setting in human orientated reading comprehension. In addition, RACE is a multiple-choice reading comprehension dataset, where the answer is one of four additional candidate options.

Multiple-choice Reading Comprehension

Multiple-choice questions are common in language examinations to human. MCTest (Richardson, Burges, and Renshaw 2013) is a multiple-choice reading comprehension dataset of high quality and the difficulty is restricted to 7 years old children. It contains 500 crowdsourcing stories and 2,000 questions, where each question is followed by four candidate answers and only one is correct. But the size is too small to efficiently train a neural network model end-to-end. So the majority of previous work on MCTest are feature-engineering models (Richardson, Burges, and Renshaw 2013; Sachan et al. 2015; Narasimhan and Barzilay 2015; Smith et al. 2015; Wang et al. 2015). These models heavily rely on lexical, syntactic and frame semantic features that extracted by various natural language processing tools. Even though these work lay heavy burden on human, but can achieve good performance on the sparse data. Yin, Ebert, and Schütze (2016) propose a hierarchical attention-based convolutional neural network (HABCNN) to project the passage with attention, and then determine a match or textual entailment. The model outperforms previous neural baselines but still performs far below feature-engineering models. Trischler et al. (2016) propose a parallel-hierarchical neural model, which is similar to HABCNN and achieves the state-of-the-art result, to match the passages, questions and candidate options from the word level to multiple sentence level perspectives. But the model must be trained with the training wheel (Trischler et al. 2016) to converge.

RACE (Lai et al. 2017) is in the same format as MCTest but of much larger size and higher difficulty. It consists of about 28,000 passages and 100,000 questions collected from English exams designed for 12 to 18 years old Chinese students. Lai et al. (2017) build a rule-based baseline (Richardson, Burges, and Renshaw 2013) with sliding window algorithm and adapt the Stanford AR (Chen, Bolton, and Manning 2016) and the GA Reader (Dhingra et al. 2017) to RACE as strong neural baselines. The neural models build

the representation of the passage without the hierarchical structure and summarize the evidence from the passage with the question. The option representation is built free of the passage and the question independently. Then the models conduct a match between the evidence and candidate options with the bilinear function. The neural models trained with vanilla backpropagation outperform the sliding window baseline.

Hierarchical Structure

Processing the passage as a single long sequence is a popular method in reading comprehension models and only a few models incorporate the passage structure. Yin, Ebert, and Schütze (2016) builds and combines representation at the sentence level and the snippet (adjacent sentences) level. But the temporal relationship is not modeled at any hierarchical level. Trischler et al. (2016) incorporate hierarchy to compare passages, questions and candidate options, and model sequential information with a location-based weight. Zhang et al. (2017) incorporate syntactic information to explore better understanding and adaptation, but limited to the question only. Xie and Xing (2017) utilize syntactic information to encode both the question and the passage sentence. They subsequently encode the passage upon sentence representation with RNN. We incorporate the passage hierarchical structure to model interactions and model the temporal context with RNN at word level and sentence level.

Attention Mechanisms in Reading Comprehension Models

Attention mechanism (Bahdanau, Cho, and Bengio 2015) is extremely popular in reading comprehension models (Hermann et al. 2015; Chen, Bolton, and Manning 2016; Kadlec et al. 2016; Dhingra et al. 2017; Sordoni et al. 2016; Shen et al. 2017; Cui et al. 2017; Seo et al. 2017; Wang et al. 2017; Xiong, Zhong, and Socher 2017). And attention is mainly used to model interactions and to predict the answer.

Hermann et al. (2015) and Chen, Bolton, and Manning (2016) use a single question vector to summary passage. Instead of representing the question with single vector, Wang and Jiang (2017), Wang et al. (2017), Cui et al. (2016), Cui et al. (2017), Xiong, Zhong, and Socher (2017) and Seo et al. (2017) utilize every word of the question to interact with passage. In Wang and Jiang (2017) and Wang et al. (2017), question words are aligned with the passage word in every time step of the passage RNN. In Cui et al. (2017) Xiong, Zhong, and Socher (2017)Seo et al. (2017), the attention between question and passage is computed in both direction. Despite the question and passage attention, Wang et al. (2017) propose the self-matching attention to match the passage against itself. Dhingra et al. (2017) propose gated-attention to select relevant part of passage with a single question vector via multiple hops. Sordoni et al. (2016) alternatively compute attention between passage and question. Shen et al. (2017) further propose to determine the iterate steps dynamically with reinforcement learning.

When it comes to the answer prediction, inspired by Vinyals, Fortunato, and Jaitly (2015), Kadlec et al. (2016) directly use attention as the pointer to predict answer for cloze-style reading comprehension. Sordoni et al. (2016), Dhingra et al. (2017) and Cui et al. (2017) subsequently adopt the same method in their prediction layer. Wang and Jiang (2017) use attention to produce the boundary of the answer span. This is an efficient and popular setting of the models on the SQuAD dataset. Xiong, Zhong, and Socher (2017) propose dynamic pointing decoder to produce answer borders iteratively. Lai et al. (2017) use the bilinear function to compute matching score of each option on RACE.

Conclusion and Future Work

In this paper, we present the hierarchical attention flow for multiple-choice reading comprehension. Passage, question and candidate options interact with each other via attention at different hierarchical levels. To fully exploit candidate options, we incorporate options to boost evidence gathering and enhance option representation with correlations, which are not explored in previous works. At last, the proposed model achieves overall sate-of-the-art accuracy on RACE and significantly outperforms two neural network baselines on both RACE-M and RACE-H subsets. We believe syntax and discourse relations can introduce additional structures as complementary information. In future work, we are interested in exploring the passage structure further by incorporating syntactic information or discourse relations for better representation.

Acknowledgments

We thank Dr. Yu Shi and Dr. Yining Chen from Microsoft STC Asia for the discussion on model design and experiments. This work was supported by the National High Technology Development 863 Program of China (No. 2015AA015407), National Natural Science Foundation of China (No. 61632011 and No. 61370164).

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; and Inkpen, D. 2017. Enhanced lstm for natural language inference. In *Proceedings of ACL*, 1657–1668.

Chen, D.; Bolton, J.; and Manning, C. D. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of ACL*, 2358–2367.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of EMNLP*, 1724–1734.

Cui, Y.; Liu, T.; Chen, Z.; Wang, S.; and Hu, G. 2016. Consensus attention-based neural networks for chinese reading comprehension. In *Proceedings of COLING*, 1777–1786.

Cui, Y.; Chen, Z.; Wei, S.; Wang, S.; Liu, T.; and Hu, G. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of ACL*, 593–602.

Dhingra, B.; Liu, H.; Yang, Z.; Cohen, W.; and Salakhutdinov, R. 2017. Gated-attention readers for text comprehension. In *Proceedings of ACL*, 1832–1846.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc. 1693–1701.

Hill, F.; Bordes, A.; Chopra, S.; and Weston, J. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of ICLR*.

Kadlec, R.; Schmid, M.; Bajgar, O.; and Kleindienst, J. 2016. Text understanding with the attention sum reader network. In *Proceedings of ACL*, 908–918.

Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of EMNLP*, 796–805.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, 1412–1421.

Narasimhan, K., and Barzilay, R. 2015. Machine comprehension with discourse relations. In *Proceedings of ACL-IJCNLP*, 1253–1262.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Onishi, T.; Wang, H.; Bansal, M.; Gimpel, K.; and McAllester, D. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of EMNLP*, 2230–2235.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, 1532–1543.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, 2383–2392.

Richardson, M.; Burges, C. J.; and Renshaw, E. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of EMNLP*, 193–203.

Sachan, M.; Dubey, K.; Xing, E.; and Richardson, M. 2015. Learning answer-entailing structures for machine comprehension. In *Proceedings of ACL-IJCNLP*, 239–249.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of ICLR*.

Shen, Y.; Huang, P.-S.; Gao, J.; and Chen, W. 2017. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1047–1055.

Smith, E.; Greco, N.; Bosnjak, M.; and Vlachos, A. 2015. A strong lexical matching method for the machine comprehension test. In *Proceedings of EMNLP*, 1693–1698.

Sordoni, A.; Bachman, P.; Trischler, A.; and Bengio, Y. 2016. Iterative alternating neural attention for machine reading. *arXiv preprint arXiv:1606.02245*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15:1929–1958.

Trischler, A.; Ye, Z.; Yuan, X.; He, J.; and Bachman, P. 2016. A parallel-hierarchical model for machine comprehension on sparse data. In *Proceedings of ACL*, 432–441.

Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer networks. In *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc. 2692–2700.

Wang, S., and Jiang, J. 2017. Machine comprehension using match-lstm and answer pointer. In *Proceedings of ICLR*.

Wang, H.; Bansal, M.; Gimpel, K.; and McAllester, D. 2015. Machine comprehension with syntax, frames, and semantics. In *Proceedings of ACL-IJCNLP*, 700–706.

Wang, W.; Yang, N.; Wei, F.; Chang, B.; and Zhou, M. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of ACL*, 189–198.

Xie, P., and Xing, E. 2017. A constituent-centric neural architecture for reading comprehension. In *Proceedings of ACL*, 1405–1414.

Xiong, C.; Zhong, V.; and Socher, R. 2017. Dynamic coattention networks for question answering. In *Proceedings of ICLR*.

Yin, W.; Ebert, S.; and Schütze, H. 2016. Attention-based convolutional neural network for machine comprehension. In *Proceedings of the Workshop on Human-Computer Question Answering*, 15–21.

Zhang, J.; Zhu, X.; Chen, Q.; Dai, L.; and Jiang, H. 2017. Exploring question understanding and adaptation in neural-network-based question answering. *arXiv preprint arXiv:1703.04617*.