mrc

Leave the summary in the end

1 Introduction

Natural language processing(NLP) is an important direction in the field of computer science and artificial intelligence, which studies various theories and technologies that can realize effective communication between human and computer with natural language. Machine reading comprehension(MRC) is to let machine learn to read and understand natural language text, find the answer from the relevant articles of a given question. MRC is one of the most challenging tasks in the field of NLP, which has many application scenarios, such as intelligent Q&A in search engine, intelligent customer service in e-commerce field and so on.

As early as the 1970s, scholars have realized that machine reading technology is the key method to test computer understanding of human language, such as the QUALM system built by Lehnert. Due to the manual coding and the system is very small, it is difficult to be extended to a larger field. Hirschman built the first automatic reading comprehension test system deep read in 1999. The system measures reading comprehension tasks based on stories, and uses bad-of-word model and manual rules for pattern matching, accuracy rate can reach about 40%. Due to the manual rules, however, the generalization ability of the model is poor. Riloff scores the matching degree of questions and candidate sentences in the article by making rules manually, then select the candidate sentence with the highest score as the answer. The traditional MRC technology mostly uses pattern matching method to extract features, even if using machine learning method, it only answers question at the sentence level granularity, and can only extract shallow features from the text. Therefore, the early MRC system has poor performance and it is difficult to put it into practical application, which leads to the slow development of MRC field.

With the rise of deep learning and development of technology in NLP field, in order to make up for the defects of the traditional MRC technology, Hermann, a researcher at deepmind, used neural network model to solve MRC tasks in 2015 and constructed a reading comprehension dataset CNN&Daily Mail which is larger than the previous datasets. They proposed two models : Attentive Reader and Impatient Reader, which based on the neural network and attention mechanism, and the performance of these two models on CNN&Daily Mail is much better than traditional models. This work is regarded as the foundation work in the field of MRC. Since then, more and more scholars have constructed better models based on these two models. For example, Kadlec use dot product to simplify attention operation and achieves better performance. Chen uses bilinear function as activation function to obtain more flexible and effective model. Cui enhances the ability of feature extraction by using attention mechanism between article and question, which attention mechanism is used not only in article, but also in question. Because the models are based on neural network, they are also refered to as neural machine reading comprehension models.

This paper reviews the research tasks, related datasets and models in the field of MRC since 2015.

2 MRC assignment outline

Machine reading comprehension task can be formalized as a supervised learning problem, which the training data is given in the form of triples: (C,Q,A), where C is the context, Q is the question, and A is the answer. The goal of MRC is to learn a mapping relationship f with the context C and question Q as input and answer A as output. MRC has developed from the task of reading a context with the answer is a single word to the task of reading multiple contexts, and the answer needs to be inferred from multiple contexts. Following Chen, according to the different forms of answers, reading comprehension tasks can be classfied into four types: cloze, extractive, multiple choice and descriptive answers.

2.1 classfication

2.1.1 cloze

Given a context C and a sentence Q with a missing word, the task of cloze test is to infer the missing word in Q. Different datasets have different sources of missing words, some from C and some from candidate answer set. The relevant datasets are as follows:

CNN&Daily Mail: Published by Google deepmind and Oxford university in 2015. The dataset consist of 93K articles collected from CNN and 220K articles collected from daily mail. Each article has some general summary sentences and the problem is generated by replacing some entities in these sentences. Meanwhile, anonymous tags are used to represent the replaced entity to prevent the interference of external knowledge.

CBT: Hill et al suggest that the marked answer will reduce the reasoning ability of the dataset needs and it does not conform with real question answering behavior of human beings. They present the CBT(Children Book Test) dataset, which is based on 108 children books with clear narrative structure. Unlike CNN&Daily Mail, the deleted word is not limited to named entities, but also nouns, verbs and prepositions in sentences.

CLOTH: CLOTH dataset collected from cloze test type of English test of Chinese middle school students. Each question is carefully designed by experts to test student's English level. The words in the blank space usually measure student's vocabulary, grammar and reasoning ability, so the CLOTH dataset is more challenging than CNN&Daily Mail and CBT.

2.1.2 multi-choice

The task of multiple choice is to select correct answer from candidate answers set $A = \{A_1, A_2, \cdots, A_n\}$ for a given context C and question Q. The dataset of this task usually comes from the reading comprehension questions in the examination papers. the question and candidate answers set are constructed manually by experts, the number of candidate answers is usually 4. The relevant datasets are as follows:

MCTest: This is the first multiple choice dataset, but because of its small size which only contains 500 story articles, it is difficult to learn by neural network, so MCTest is usually used as verfication set or test set.

RACE: This is a dataset based on the reading comprehension questions of Chinese middle school, which contains about 28 thousands articles and 100 thousands questions. In addition, RACE covers many fields, such as news, story, advertisement and biography etc. It can better evaluate the reading comprehension ability of machine because of the diversity of its types.

ARC: The most challenging multiple choice dataset is ARC which proposed by Clark. They classfied the questions whose answers can be obtained by retrieval or word co-occurrence as simple sets, and those that

cannot be obtained by the above two methods as challenge sets. The answers in the challenge set will not appear in the context, and need to be inferred and selected by the model combined with external knowledge.

2.1.3 extractive answer

The task of extractive reading comprehension can be regarded as an extension of cloze task. Unlike cloze task, which only requires the answer to be a word in context, extractive task requires the model to extractive a continuous piece of text from the context as the answer, and the length of the answer is not fixed. This kind of reading comprehension task is a popular research direction in the field of MRC because it is more appropriate from the perspective of dataset construction, evaluation metric and application value. The relevant dataset are as follows:

SQuAD: SQuAD is a representative dataset of extractive reading comprehension, which is also one of the most widely used dataset in the field of MRC, it has greatly promoted the development of MRC. The dataset is constructed by crowdsourcing service platform, crowdsourcing workers give questions according to the article in Wikipedia.SQuAD contains 536 articles and more than 10 thousands (Q,A) pairs.

NewsQA: The construction of NewQA is similar to SQuAD. The difference is that the articles in NewQA comes from CNN news, and answers to some questions is null, which makes Rajpurkar et al add 50 thousands unanswerable questions to SQuAD to construct the dataset SQuAD 2.0.

TriviaQA: There is a common feature in the construction of above datasets, which is that the question and answer are constructed after the context is given. However, this will lead to (Q,A) pair implicitly contains context information, which reduces the difficulty of dataset. Joshi et al. collected a large number of <Q,A> pairs from quiz League and then searched the relevant articles from web page or Wikipedia for each question, the retrieved articles are used as the evidence for finding answers. In this way, they built a dataset called TriviaQA, which contains more than 650 thousands (C,Q,A) triples. This way of finding contexts through questions and answers makes the questions and contexts have a large difference in syntax and morphology, which makes the dataset more difficult.

There are also some typical extractive datasets, but these datasets mainly examine the reasoning ability of the model. For example, HotpotQA, which each question corresponds to multiple paragraphs, the answers to the question often needs to deduced step by step in multiple paragraphs.

2.1.4 generate

Extracting a text span from an article makes the answer semantically too stiff. In addition, the answer to the question extracted from the article is more in line with people's reading comprehension form. So researchers began to turn to the free-form reading comprehension task. The answers of this kind of task are free-form, not limited to some words in the article, and the grammar is often more flexible. The relevant dataset are as follows:

MS MARCO: MS MARCO is one of the representatives of free answer reading comprehension data set. The data set is collected by Microsoft through the log of Bing search engine. The paragraphs are from the 10 most relevant query paragraphs returned by Bing search engine, and the answers are extracted from these paragraphs manually.

DuReader: DuReader is a Chinese reading comprehension data set. This construction method is similar to that of MS MARCO. The questions and articles are from Baidu search and Baidu zhidao, and the question types are also include yes no questions.

NarrativeQA: Kovcisky et al think that the difficulty of most data sets in MRC field is too simple, and the answers only focus on context information. Many questions can be answered only by shallow pattern matching.

The data set NarrativeQA released by them avoids this deficiency. NarrativeQA is collected from novels and movie scripts. Some of questions need to understand the whole novel or script to find the answers, which requires the model to have stronger ability of understanding and reasoning.

2.2 evaluation metric

There are different evaluation indexes for different MRC tasks. Both the cloze task and multiple-choice task belong to the objective type, and accuracy can be used to measure the performance of the model.

Extractive task belongs to semi objective type, which usually evaluated by extract match(EM) or f1 score. EM evaluation index can be regarded as an extension of accuracy. In terms of extractive task, EM requires that the predicted answer fragment should be consistent with the standard answer, and the EM value is 1, otherwise it is 0. The calculation of f1 score is a kind of fuzzy matching, which is the harmonic average between precision and recall. Precision refers to the proportion of words in the answer predicted by model are words in standard answer. Recall refers to the proportion of words in standard answer appearing in the predicted answer.

The matching rate of word level is generally used as the scoring standard for free-form tasks, and common standards are ROUGE-L and BLEU. ROUGE-L is used to calculate the longest common subsequence(LCS) of standard answers and predicted answers. BLEU is originally used to evaluate translation performance. when it is applied to MRC task, it is mainly used to measure the similarity between predicted answers and standard answers. Table1 lists all the data sets introduced in this chapter and the corresponding evaluation methods.

3 Neural machine reading comprehension model

3.1 general architecture

Machine reading comprehension generally needs the following processes:

1. Representing the unstructured data in the text form of paragraphs and questions as a form that can be processed by the computer.
2. In order to enhance the semantic representation of a paragraph or question, it is necessary to enable a word in question or paragraph to pay attention to its contextual information.
3. According to the question, retrieve the part of the paragraph that is most relevant to the question.
4. Summarize the answers from the retrieved article fragments.

It can be seen from the whole process that each step has a clear purpose and corresponds to a certain layer in the neural network. The overall framework of deep learning model for MRC tasks mainly includes the following layers: word embedding layer, feature extraction layer, interaction layer and answer output layer. The word embedding layer corresponds to step 1, which embeds paragraph and question into a low dimensional vector space, and uses each vector to represent each word; The feature extraction layer corresponds to step 2, and its function is to encode the semantic information of paragraph and question, so that each word can pay attention to its context; The interaction layer corresponds to step 3, whose function is to fuse the semantic information of the paragraph with the semantic information of the question, so that the model can learn the most relevant part of paragraph with the question; The answer output layer corresponds to step 4, and the goal is to find the answer to the question from the paragraph.

3.1.1 Word embedding layer

How to express text into a form that can be processed by computers while effectively utilizing the semantic relationship between words has always been one of key issues in NLP field. The early one hot form coding uses a binary vector to represent words, but there is the problem of data sparsity and dimensional disaster with the increase of the number of words. In addition, this form of coding could not represent the semantic relationship between words.

Rumelhart et al. first proposed the concept of distributed representation. Distributed representation is to embed words into a low-dimensional vector space and use a low-dimensional vector to represent a word. This representation, therefore, is also called word embedding. Words with similar semantics have similar distances in vector space. This word representation method solves many problems in one-hot encoding. Bengio et al. first put the idea of deep learning into the language model, and proposed the Neural Network Language Model(NNLM). The mapping matrix at the first level of the model is the word vector learned. Mikolov et al. were inspired by this idea and proposed Word2Vec. Word2Vec uses two models CBOW and Skip-gram to learn the distributed representation of words. CBOW uses the context words of a word to predict the word, and Skip-gram uses the word to predict the words around it. In addition, Glove considered the global statistical information by using the word co-occurrence matrix.

A large number of experiments show that using Word2Vec or GloVe pre-trained word vectors as word features of downstream task texts can significantly improve the performance of the model. In addition to word embedding, there are many fine-grained embedding methods. For example, Seo et al. proposed to combine word embedding and char embedding to alleviate the common OOV(out-of-vocabulary) problem in NLP field. Chen et al. proposed to introduce semantic features of words to enhance embedded representation, such as exact matching features between paragraph and question, part-of-speech features and named entity features of words.

The word vectors trained by Word2Vec and GloVe are static word vectors. After training the model, the expression of words are fixed without considering the context information, so the polysemy question can not be solved. Peters proposed a dynamic context-based word embedding model ELMo, which used two layers of bidirectional BiLSTM with residual connection. The vector of each word is expressed according to its context semantics, which solve the problem of polysemy.

From early one-hot coding to distributed representation and then to context-based word embedding, the emergence of each technology proves that a good text representation method can significantly improve the performance of the model.

3.1.2 coding layer

The purpose of coding layer is to further obtain semantic information at the sentence level on the basis of the word embedding layer. The most commonly used feature extractors in NLP are based on variants of recurrent neural networks(RNNs) such as LSTM and GRU. Vaswani et al. proposed a model Transformer which based on self-attention mechanism. Experiments show that Transformer has better feature extraction ability than RNNs and can speed up training through parallel computation, such as QANet[?]. See section 3.3.1 for details about Transformer.

3.1.3 interaction layer

Interaction layer is the key layer in the whole four-layer network. The purpose of this layer is to integrate the semantic information of paragraph with the semantic information of question so as to have a deep understanding

of paragraph. The most commonly used method in the interaction layer is attention mechanism.

Attention mechanism can be regarded as a mapping process between a query vector and a set of key-value pairs(in NLP tasks, key equals value). The function f is used to measure the similarity between query and key to generate a weight vector, and then the weight vector is normalized(usually using softmax function). The value will be weighted and summed by the normalized weight vector, and the result is the attention of query to key-value pairs. The formula is as follows:

MRC model has two directions for attention operation: from question to context(Q2C) and from context to question(C2Q). Taking Q2C as an example, its attention refers to treating question as Q and context as K,V. Definiting $C = [c_1, c_2, \cdots, c_n] \in R^{n \times d}$, which represents the semantic representation of context, where n is the number of words in the context, d is the vector dimension, $Q \in R^d$ represents the semantic representation of the whole question. The attention calculation steps of Q2C are as follows:

The calculated attention is also called context-aware question representation. The attention operation of C2Q is in a similar way. The difference is that at this time the context is regarded as Q and the question is regarded as K,V, the attention calculated at this time is called question-aware context representation. In addition, attention mechanism in MRC field can be divided into one-hop and multi-hop forms. One-hop refers to the weight vector is obtained by doing one interactive calculation between context and question. Multi-hop can be seen as a stack of one-hop, which aims at realizing multi-step reasoning. For complex problems, the answer may not be found in one sentence, so the reasoning needs to be carried out in multiple sentences. In each step of reasoning, the object of attention will be changed.

3.1.4 output layer

The answer prediction layer is the last layer of the whole model architecture. As mentioned in the previous chapter 2, MRC tasks can be divided into cloze type, multi-choice type, extraction type and free-answer type according to the answer form, so the design of this layer needs to consider the answer form. Since the method of multi-choice tasks can be summed up as cloze tasks(each choice can be regarded as a candidate for cloze positions), and cloze tasks are special cases of extractive tasks, this section focuses on the design of the output layer for extractive and free-answer tasks.

For extractive reading comprehension tasks, a continuous span text is extracted from the article as the answer.The pointer network model evolved from seq2seq model, and directly outputs the prediction results according to the calculated attention weight distribution, thus solving the problem that the output originates from the input.

Inspired by pointer networks, Wang et al. proposed two output models based on pointer networks. The first is a sequential model, which predicts each position of the answer in a sequential form. The second is a boundary model. Unlike the sequential model which serially predict each position of answer, only the start position and end position of the answer are predicted in boundary model since the answer to predicted is a continuous piece of text. The method is simpler than the first one, and its experimental results also show that its more efficient. Therefore, the output layer of extractive MRC model is always designed to predict only the start position and end position. The loss function of output layer can be written as $L(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log P^S_{y_i^s} + \log P^E_{y_i^e}$, where θ is the model parameter, N represents the number of examples, y_i^s and y_i^e indicate the start position and end position of the answer in the context of the i-th example respectively.

For free-answer tasks, the form of the answer is no longer a continuous text in the article, but a text conforming to grammatical norms needs to be generated according to the context and question. This kind of

task requires higher ability of answer generation module. A typical architecture for handling generation tasks is the seq2seq model, which takes the context and question as the input of encoder, and decoder generates answers based on words in the vocabulary. There is also a classical architecture for processing generation tasks, which is the pointer generator network proposed by See et al. The model combines the generation mechanism of seq2seq network and the copy mechanism of pointer network, so that the predicted words can be generated from the vocabulary or copied from the original text. The experimental results of DrQA show that the performance of PGNet in free-answer reading comprehension tasks is better than that of the traditional seq2seq model.

3.2 Classical MRC model

Among these four layers, the word embedding layer and coding layer is not unique to MRC model, and other tasks of NLP also include these two layers. What truly embodies the characteristics of each MRC model are the interaction layer and the answer output layer, especially the interaction layer. Therefore, this section focuses on the design of attention mechanism in the interaction layer of each model. In view of the complexity of the attention mechanism used in the interaction layer of most model at present, it is difficult to completely distinguish each model according to the form introduced in Section 3. In the light of the thinking of Liu et al. this paper divides each model according to the direction and structure of attention mechanism.

3.2.1 uni-direction

Unidirectional attention is usually to calculate the attention of the problem to the context(Q2C), and the calculated value represents the weight of the corresponding words in the context, so as to highlight the part of the context that is most silimar to the question. Hermann et al was the first to use neural network and integrate attention mechanism to solve MRC tasks. They proposed two different models of unidirectional attention mechanism, AttentiveReader and ImpatientReader, which both calculate the attention in the direction from question to context, and the calculation method of attention adopts feedforward neural network(Equation .). On this basis, Chen et al. use bilinear term(formula 3) to replace the original feedforward neural network, and Kadlec et al. use inner product(formula 5) as attention calculation method.

3.2.2 bidirectional attention

Unidirectional attention can extract limited interactive information, while bidirectional attention can complement the two directions and provide more comprehensive interactive information. Xiong et al. proposed Dynamic Co-attention Network(DCN) model, which adopts a cooperation attention mechanism in interaction layer, that is, synchronously calculates the attention of C2Q and Q2C, and finally fuses the attention of the two directions as the output of the interaction layer. The bidirectional attention flow(bidaf) model proposed by Seo et al. also calculates the attention of two directions(C2Q and Q2C). Different from previous model, BiDAF merges the paragraph semantic representation output by the coding layer and the question-aware paragraph representation calculated by interaction layer to flow to the following layer. In this way, to a certain extent, the problem of information loss caused by the premature generalization of paragraph semantics is avoided.

3.2.3 self-attention

Q2C attention can be considered as reading the context with the question, while C2Q attention can be considered as reading the question with the context. These two kinds of attention belong to interactive attention mechanism. However, the excessive dependence on prior information of interactive attention, especially for Q2C attention,

may lead the model to only pay attention to the information with high correlation to the question in the article and ignore the semantic information emphasized by the model itself. The self-attention mechanism enables each word in the article to pay attention to all the others word, making the model achieve a deeper understanding of the semantics of the article. Many models add self-attention mechanism on the basis of interac tive attention. For example, RNet and BiDAF++, etc.

3.2.4 hop

One-hop structure means that the interaction between the context and the question is calculated only once. Either the whole question is compressed into a vector, and then the attention is calculated with the context, such as AttentiveReader, AS Reader, etc. or the representation of the question and the context is calculated in parallel, such as DCN, BiDAF etc.

The one-hop structure can not achieve the effect of multi-step reasoning. Multi-hop structure can be regarded as a stack of one-hop structures. The purpose is to deepen the understanding of context and question by calculating the interaction between context and question many times. Each interaction calculation will change the objects of attention appropriately, so as to achieve the purpose of multi-step reasoning. There are several ways to realize multi-step reasoning:

1. The first method is to calculate the interaction between the paragraph and question of the current time step based on the question-aware paragraph representation which calculated by the previous time step, such as ImpatientReader. This is to calculate attention in a sequential way, which is similar to human reading way, that is, we constantly interact between question and paragraph in the process of reading.

2. The second method is to use RNNs, which is based on the previous hidden state to update the next hidden state. For example, the Match-LSTM model proposed by Wang et al. reads the paragraph sequentially while using LSTM to store the question-aware paragraph representation of each time step. Specifically, the attention between the word at the current time of the paragraph and the question is calculated, and the obtained question-aware vector representation and the vector representation of current word are combined as the input of Match-LSTM. Similar models such as RNet and IA Reader, etc.

3. The third method is to achieve the purpose of multi-step reasoning by stacking multiple levels of computation attention, such as GA Reader model and Reinforced Mnemonic Reader etc.

3.3 MRC based on pretrain

Pre-training model has received great attention in NLP field in recent years. The pre-training method originates from the concept of transfer learning. Firstly, the model is pre-trained on other related tasks so that the model can learn some knowledge, then further optimization is made on the target tasks to realize the transfer of knowledge learned by the model. For NLP domain, the pre-training process is to learn the general knowledge representation on a large number of text data. The comparison between the structure based on pre-training and the traditional structure is shown in figure 2.

3.3.1 Pre-training model

OpenAI proposes a generative pre-training(GPT) model, which uses the decoder of transformer as feature extractor. The objective function of pre-training stage is the objective function of unidirectional language model, see formula 1.

GPT belongs to autoregressive language model. The disadvantage of autoregressive language model is that it cannot predict a word by using the context information of the word at the same time due to the nature of autoregressive. Devlin et al. proposed a very powerful pre-training model BERT. The difference with GPT lies in the pre-training method adopts denoising autoencoder, which randomly masks some words and obtains the probability distribution of masking positions at the output layer, and lets the model predict the masked word according to the context of the masked position. This mechanism also called masked language model(MLM) or bidirectional language model. In addition to MLM task, BERT also uses Next Sentence Prediction(NSP) task to make the model perform better in downstream tasks such text entailment, question answering, which need to judge the semantic relationship between the two sentences. It can be seen that the pre-training process of BERT is essentially a multi-task learning process, and the semantic expression ability of BERT is improved through MLM and NSP tasks. The performance of BERT on the most classic machine reading comprehension dataset SQuAD has surpassed that of human beings. BERT opened the era of pre-training models in NLP domain. Since then, many more powerful pre-training models have been proposed one after another.

$$\pi = x_i s \tag{1}$$