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