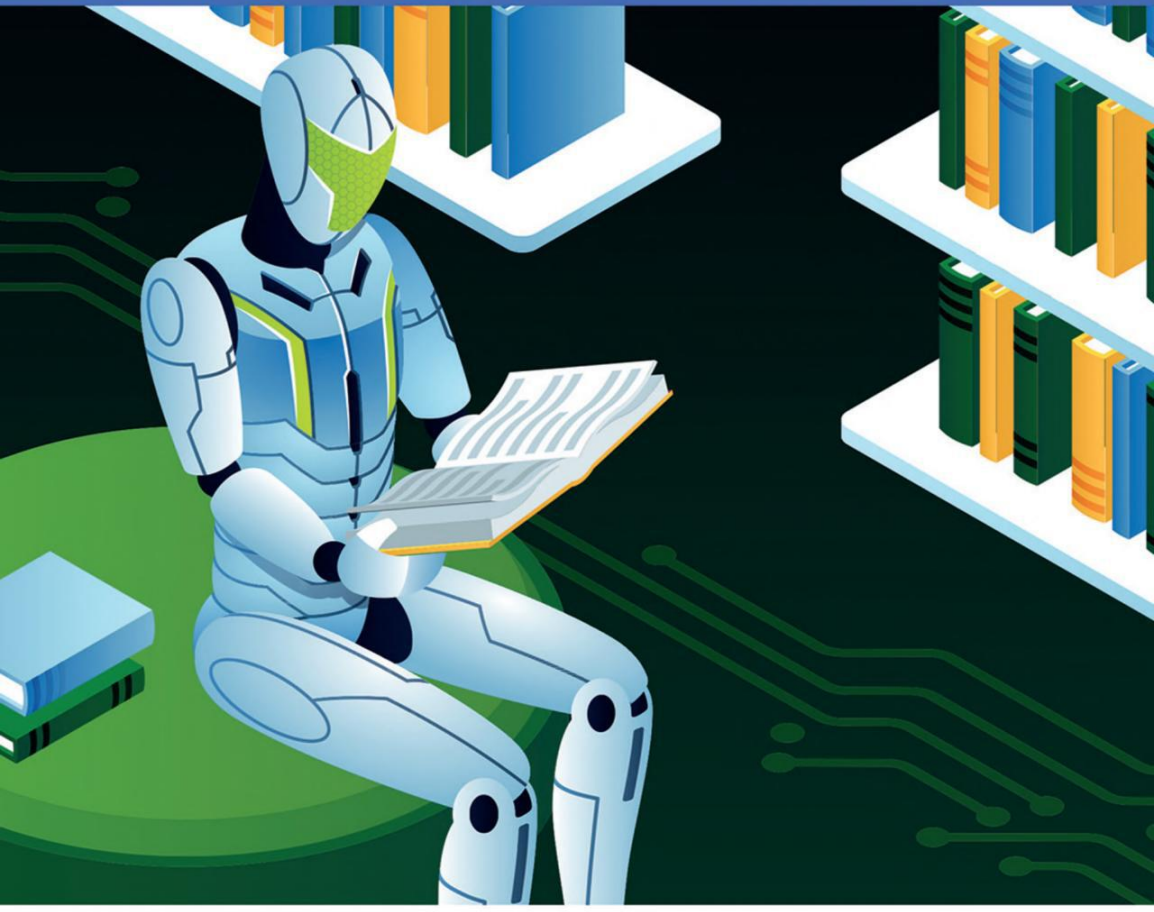
Machine Reading Comprehension

Algorithms and Practice

Chenguang Zhu



ELSEVIER

HZ BOOKS
华章IT

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-323-90118-5

Publisher: Glyn Jones
Editorial Project Manager: Naomi Robertson
Production Project Manager: Punithavathy Govindaradjane
Cover Designer: Mark Rogers

Typeset by MPS Limited, Chennai, India

Contents

This page intentionally left blank

About the author

Dr. Chenguang Zhu is a Principal Research Manager in the Microsoft Corporation. Dr. Zhu obtained his PhD in Computer Science from Stanford University, United States. He is leading efforts in the research and productization of natural language processing in Azure Cognitive AI. He is proficient in artificial intelligence, deep learning, and natural language processing, specializing in machine reading comprehension, text summarization, and dialogue understanding. He has led teams to win the first place in the SQuAD 1.0 Machine Reading Comprehension Competition held by Stanford University, and reach human parity in the CoQA Conversational Reading Comprehension Competition. He has 40 papers published in top AI and NLP conferences, such as ACL, EMNLP, NAACL, and ICLR, with more than 1000 citations.

This page intentionally left blank

Foreword by Xuedong Huang

There are two levels of intelligence. One is perceptual intelligence, which enables the computer to see, hear, and touch. In these areas, artificial intelligence has made many breakthroughs, such as speech recognition, speech synthesis, and computer vision. A higher level is cognitive intelligence, which requires computers to understand and analyze concepts, relationships, and logic. At this level, artificial intelligence is still in its infancy.

As an important medium for human communication and information dissemination, language is at the core of human intelligence. From the Turing Test in the 1950s to the deep learning era today, the understanding and application of natural language has always been a hot research topic. If AI is the crown, speech and language technologies are the jewels on top of the crown. It is fair to say that if computers can fully understand human language, we'll have achieved strong artificial intelligence.

In recent years, machine reading comprehension has become one of the most popular and cutting-edge directions in natural language processing research. It has significant scientific and practical values to enable computer models to read articles, analyze semantics, and answer questions like humans. Machine reading comprehension technology can automate plenty of time-consuming and laborious text analysis work and greatly improve the productivity of many applications, ranging from intelligent customer service to search engines, from automatic essay scoring to intelligent finance.

With the development of deep learning technology, the research of machine reading comprehension has made tremendous progress. In some specific tasks, the answers of computer models are already comparable to the human level. Some media reports have even claimed that the computer is superior to humans in reading comprehension. However, existing models are still far from a genuine and thorough understanding of text. In many cases, these models still rely on simple matching of words and phrases, rather than on a thorough understanding of the syntactic structure and semantics.

In general, there are three key factors to the success of an artificial intelligence system: platform, data, and algorithm. As the computing power and magnitude of data continue to soar, the exploration and improvement of algorithms have become a hotly contested spot for artificial intelligence research.

Currently there are very few books on the market which have a complete introduction to machine reading comprehension. Dr. Chenguang Zhu in our team has worked deeply in this direction for many years and has led the team to achieve top places in a number of international contests. The purpose of his book is to objectively show the field of machine reading comprehension to readers. The book includes a detailed introduction to the latest research results and thoughts on the future directions of machine reading comprehension. I hope this book will inspire readers to work together to achieve human-level machine reading comprehension.

Dr. Xuedong Huang
CTO of Artificial Intelligence, Microsoft,
Redmond, WA, United States

Foreword by Zide Du

Natural language processing aims to solve the problem of understanding and generating natural language. Natural language is the jewel in the crown of artificial intelligence. It is one of the most important abilities of computers, but is also a challenging direction to study. Every human language has its own grammar, but because of the different styles of usage, coupled with factors such as dialects and idioms, the resulting forms of language have a large variation. These variations usually do not interfere much with the communication between humans, but it is very difficult for computers to understand. This is because the current von Neumann computer architecture is good at handling information with clear rules, but is less capable of handling constantly evolving forms of information.

Over the years, researchers have proposed and developed many methods, ranging from rule-based linguistic techniques to models based on statistical machine learning. In recent years, researchers have developed end-to-end deep learning frameworks for natural language processing, including word embeddings, attention mechanisms, encoder–decoder architecture, and the recent, pretraining models. These techniques have greatly improved models' ability to understand text and brought new interesting ideas to natural language processing.

Machine reading comprehension is one of the most popular and cutting-edge research topics in natural language processing. Reading is the basic means for people to obtain information. Without reading there is no understanding, and without understanding one cannot communicate. There are already many chatbots in the market, but people often find them responding off the point. The reason is that the current technology is a black box approach based on text matching. So the chatbot does not really understand what people mean. As we know, humans communicate with context, so that we can easily understand what other people are talking about via referencing. However, it is very difficult to make machines understand the context. In order to solve these problems, researchers have proposed many ways to improve models' ability to understand dialog and articles. Moreover, the release of many reading comprehension datasets have played an important role in promoting the development of technology.

In addition to its research values, machine reading comprehension has many meaningful applications. For example, machine-generated article

summaries can save a lot of time of reading the full text, and the QA system can accurately find answers to user questions from a large number of documents. Machine reading comprehension is also the basis for translation and dialog, which are of great value to computer-assisted services.

Chenguang's book systematically introduces the key technologies and progress in this area as well as existing challenges. I believe that readers will have a clear understanding of the research and application of this field after reading this book.

During high school, Chenguang participated in the National Olympiad in Informatics (NOI) organized by the China Computer Federation (CCF), winning gold medals in the national competition. He was also a candidate for the Chinese team for the International Olympiad in Informatics. As I was the chairman of NOI, I got to know him back then. He later went to Tsinghua University to study computer science, got a PhD at Stanford University, and now works at Microsoft on natural language processing. We rarely see each other, but constantly keep in touch. I think he is a talented young scholar with a rigorous and very sensible style. I am therefore very happy to discuss various issues with him. He asked me to write the foreword for his new book and I am very glad to see his research progress. I also want to express my congratulations to him.

Zide Du
Former Researcher at the Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China;
Secretary General of the Chinese Computer Federation, Beijing, China

Preface

Reading is an important means for humans to acquire knowledge and understand the world. For thousands of years, language has been the carrier of human civilization, containing a wealth of information, experience, and wisdom. Language's nature of high concentration makes reading ability a vital intelligence. The famous science fiction writer, Mr. Liu Cixin, described the efficiency of human language communication in his novel *The Rural Teacher*:

> *You're trying to tell us that a species that has no memory inheritance, communicates with each other in sound waves at an incredible rate of 1 to 10 bits per second, can create a 5B-level civilization?! And this civilization has evolved on its own without any cultivation from external advanced civilization?!*

It is estimated that the average reading speed of humans is about two to three times the speed of speech. Thus, under a rough estimate, even if a person reads for 8 hours a day for 50 years, he or she can only get about 1.5 GB of information out of reading. However, human civilization has far exceeded this magnitude. Therefore reading is a complex process of abstracting text into concepts, ideas, and derived knowledge through understanding.

In today's wave of artificial intelligence, it is very important to enable the computer learn to read. On the one hand, reading ability lies at the core of human intelligence, which is indispensable in the ultimate form of artificial intelligence. On the other hand, with the explosion of text data, computer models can automate the process of text understanding, save a lot of cost and time, and have a wide range of applications in many industries.

Thus in recent years machine reading comprehension (MRC) has been one of the most cutting-edge research topics in natural language processing. The goal of the study is to teach computers to read articles and answer related questions like humans. Lots of AI technologies have been applied to this field, and there have emerged many MRC tasks. I was fortunate enough to be among the first group of researchers in this field, designing and implementing several models to win the first place in the SQuAD competition hosted by Stanford University, and surpass human level performance for the first time in the conversational MRC competition CoQA.

However, as computer models outperform humans on more datasets, many media reports use headlines like *Computer has beaten humans in reading*, which contributes to the so-called claim that AI has replaced humans. As a researcher in this area, I deeply feel that computers are far from humans' ability of reading and understanding. Although current reading comprehension models have achieved huge progress compared with a decade ago, they are above the human level only in specific datasets under various constraints. Studies have shown that the performance of these models will significantly drop when a confusing sentence is appended to the article, while humans find it very easy to judge.

In contrast to the boom of machine reading comprehension research, there are so far no books on the market in this field. Most progress is published in the form of academic papers, and little information can be found about the application of MRC in industry. Therefore the purpose of this book is to objectively show the status quo of machine reading comprehension research. Starting from the basic modules, the architecture of models, and to cutting-edge algorithms, the book details the design and implementation of machine reading comprehension models. There are numerous code examples in Python and PyTorch by the author, showcasing the model building process, which has a high practical value. All the code is available at https://github.com/zcgzcgzcg1/MRC_book_en. In addition, the book introduces the landing of machine reading comprehension technology in various industrial applications, such as intelligent customer service and search engines, and points out the challenges and future directions of MRC research.

Although currently machines are still inferior to humans in terms of reading ability, we can leverage the high speed and large storage of computers to overtake humans. As said in the ancient Chinese proverb, *When one learns 300 poems of the Tang Dynasty by heart, he is sure to be able to write poetry*. Nowadays, computers can read a million poems in a split second. So there is every reason to expect a breakthrough in machine reading comprehension. For example, the BERT model combines the merits of massive data and large models to make breakthroughs in many areas of natural language processing including machine reading comprehension. I hope that this book can inspire readers to make computers achieve and exceed the reading ability of humans in the near future.

The book is divided into three parts, with a total of eight chapters.

The first part is Foundation (Chapters 1—3), which introduces the basics of machine reading comprehension and key technologies. These

include the definition of the MRC task, natural language processing techniques used in MRC models, and related deep learning network modules, such as how to represent articles and questions in computers, how to answer multiple-choice questions, and how to generate freestyle answers.

The second part is Architecture (Chapters 4—6), which introduces the basic model architecture and popular models for machine reading comprehension. It also analyzes the state-of-the-art pretrained models, for example, GPT and BERT, that have had a revolutionary impact on machine reading comprehension research.

The third part is Application (Chapter 7: Code Analysis of SDNet Model and Chapter 8: Applications and Future of Machine Reading Comprehension), including the code analysis of the SDNet model which won first place in the CoQA MRC competition in 2018, the process of landing machine reading comprehension technology in various industrial applications, as well as challenges and future directions of MRC research.

The errata information of this book is available in the code link above. If you have any comments, please contact me at zcg.stanford@gmail.com. I look forward to receiving feedback and communicating with dear readers.

Acknowledgment

Many thanks to Dr. Xuedong Huang and Mr. Nanshan Zeng from Microsoft Cognitive Services for their guidance and help.

Many thanks to Mr. Zide Du, Secretary-General of the China Computer Federation, for his long-time encouragement and support.

Thanks to Prof. Maosong Sun of Tsinghua University, Prof. Jiajun Wu of Stanford University, Prof. Meng Jiang of the University of Notre Dame, and Principal Scientist Quoc V. Le of Google for writing recommendations for this book.

Special thanks to my wife Mengyun and daughter. I sacrificed a lot of family time in writing this book. Without their love and support, I could not finish this book.

Thanks to my parents who helped to take care of my daughter to support me to finish writing.

I would like to dedicate this book to my dearest family, as well as to all my friends who love machine reading comprehension!

Chenguang Zhu

This page intentionally left blank

Recommendation

The last several years can be seen as a golden era of natural language processing, especially machine reading comprehension. The rapid progress has the potential to enable many applications that we could only imagine before. Both beginners and experts in machine reading comprehension will enjoy this book because it is comprehensive and it explains difficult concepts in an easy-to-understand manner. I highly recommend it!

Quoc Le

Principal Scientist, Google Brain, Mountain View, CA, United States

We have witnessed rapid progress in natural language processing in the past few years. Dr. Zhu's book is a very good introduction to this field. Through both hands-on code samples and the author's deep understanding and analysis of the area, the book will be highly useful to readers at all levels.

Jiajun Wu

Assistant Professor in the Computer Science Department at Stanford University, Stanford, CA, United States

I love this book. It will help both beginners and long-term researchers. It can be used for teaching, research, and even self-study and experimentation. Thanks to the emergence of this book, I hope you will also start to love artificial intelligence and natural language processing.

Meng Jiang

Assistant Professor in the Department of Computer Science and Engineering, the University of Notre Dame, Notre Dame, IN, United States
Head of Data Mining Towards Decision Making Lab, Notre Dame, IN, United States

Machine reading comprehension is a frontier research topic in natural language processing and artificial intelligence. In recent years, there are many research results and international competitions in this field. However, there are very few books that comprehensively investigate machine reading comprehension. The publication of this book fills this gap at the right time. The author has an educational background at Tsinghua University and Stanford University, with extensive research and engineering experience in the world's leading IT companies. He has led teams to achieve first places in the Stanford Conversational Question Answering Challenge

(CoQA), the Stanford Question Answering competition (SQuAD v1.0), and the AI2 Reasoning Challenge (ARC). Therefore this book has a great combination of both cutting-edge research results and practical applications. At present, the study of machine reading comprehension is rapidly rising, so I believe you will like reading this book.

Maosong Sun
Professor in the Department of Computer Science at Tsinghua University, Beijing, China
Executive Vice President of the Institute of Artificial Intelligence at Tsinghua University, Beijing, China

This book provides an in-depth introduction to the basics of natural language processing, various model architectures, the applications, and challenges of machine reading comprehension, coupled with detailed examples. The author also shares his leading research results on machine reading comprehension and in-depth thinking on the future direction. I recommend this book to students, researchers, and engineers who specialize in natural language processing, especially machine reading comprehension.

Michael Zeng
Partner Research Manager, Head of AI Cognitive Services Group, Microsoft, Redmond, WA, United States

PART I

Foundation

This page intentionally left blank

CHAPTER 1

Introduction to machine reading comprehension

1.1 The machine reading comprehension task

Since the advent of computers, we have dreamed of enabling machines to acquire human-level intelligence. Among the various forms of intelligence, understanding language is essential in human life, including daily communication, description of concepts, and propagation of ideas. As a famous example, the Turing Test proposed by Alan Turing in 1950 employed conversation as an important criterion for artificial intelligence.

Machine reading comprehension (MRC) is one of the most important tasks in language understanding. Snow defined reading comprehension in [1] as "the process of extracting and constructing article semantics from written text by interaction." The goal of MRC is to use artificial intelligence technology to enable computers to understand articles like humans.

1.1.1 History of machine reading comprehension

The history of MRC dates back to the 1970s, when researchers started to recognize the significance of text understanding in artificial intelligence. The first reading comprehension system, QUALM [2], was proposed in 1977. The system was built upon hand-coded scripts and focused on pragmatic issues of reading comprehension, including its relation to question answering.

Twenty years later, a reading comprehension dataset consisting of about 120 stories for 3rd–6th grade students was released in 1999, together with a rule-based model Deep Read [3]. The model leveraged low-level linguistic features and could achieve 30%–40% accuracy in question answering.

In 2013 the MRC problem was formally framed as a supervised learning task: given a passage and a related question, the model should give the answer in the required format. With the advent of larger datasets like McTest and ProcessBank, many machine learning models have been proposed. Most of these approaches leveraged linguistic tools such

Machine Reading Comprehension. DOI: https://doi.org/10.1016/B978-0-323-90118-5.00001-1

as dependency parsers and named entity recognition to obtain features and build statistical models to maximize the probability of correctness of the generated answer.

From 2015 the vast majority of MRC algorithms have been built on deep learning and deep neural networks. These approaches utilize their immense model complexity to accurately characterize the semantic space and achieve higher answer accuracy. Moreover, these models typically don't require manually designed features. Instead, a robust and generalizable feature representation can be automatically learned from the data. This greatly reduces the dependence on expert knowledge and downstream linguistic tools. The success of deep learning models is also closely related to the emergence of various large-scale MRC datasets such as SQuAD, RACE, and MS MARCO. In this book, we will primarily focus on the MRC models based on deep learning.

1.1.2 Application of machine reading comprehension

With a plethora of text data generated from various industries, the traditional way of manually processing data has become the bottleneck of many applications due to its slow speed and huge cost. Therefore MRC technology, which can automatically process and analyze text data and extract semantic knowledge from it, is gaining more popularity.

For example, the traditional search engine can only return documents related to user queries, while an MRC model can pinpoint the answers in the document, thereby improving the user experience. MRC can also greatly improve the efficiency in customer service when searching for solutions to users' problems in product documentations. In the field of medical intelligence, a reading comprehension model can analyze the patient's symptoms and automatically consult huge piles of medical records and papers to find possible causes and give a diagnosis. MRC can help revise essays for students and offer suggestions for improvement, enabling students to improve their writing skills anytime, anywhere. Chapter 8, Applications and Future of Machine Reading Comprehension, will cover these applications of MRC in more details.

Thus MRC can help save tremendous manpower and time in scenarios that require automated processing and analysis of a large amount of text. Even if the quality of a reading comprehension model does not completely reach the level of humans, it can save cost by solving a part of the problem space. For instance, in customer service, the computer can

focus on solving the most frequent problems with a high accuracy, while resorting to human agents for the remaining problems. Due to its widespread applications in various domains, MRC has become one of the most popular directions in cutting-edge AI research.

1.2 Natural language processing

MRC is an important direction in natural language processing (NLP). NLP analyzes the patterns and structures of human language, with the goal of designing computer models to understand language and communicate with humans. The history of NLP can be traced back to the birth of artificial intelligence. Over the decades, we have made huge progress in many NLP areas, such as understanding and generation, which has laid a solid foundation for MRC research. Thus in this section we will introduce the status quo of NLP research and its impact on MRC.

1.2.1 The status quo of natural language processing

NLP has evolved over 70 years, with many refined subtasks in the field. Here is an introduction to those important research directions related to MRC:

1. *Information retrieval* studies how to find results related to user queries in a massive number of documents or webpages. The research on information retrieval is relatively mature and widely used in products like search engines. It greatly promotes the dissemination and acquisition of information. When a reading comprehension task involves a large-scale text corpus, information retrieval is usually employed as the first module to extract relevant information.
2. *Question and answering system* establishes a system that automatically answers a user's question. The difference between a QA system and an information retrieval system is that a QA system needs to understand the semantics of complex questions and often support multiple turns of question answering. In a conversational reading comprehension task, the model should analyze the information from both the article and previous rounds of conversation to answer the question.
3. *Text classification* refers to the task of classifying articles, paragraphs, and statements, such as categorizing webpages by content and subject. In MRC, some models build a text classification module to check whether the question is about time, location, or other category of information. This can help improve the accuracy of answers.

4. *Machine translation* studies how to let the computer automatically translate text into other languages. This can be applied to cross-lingual reading comprehension tasks. For example, we can use machine translation to generate training data from popular languages to train a MRC model on low-resource languages.
5. *Text summarization* studies how to summarize an article's salient information in an abridged version. Because text summarization involves a deep analysis of the article semantics, many of its techniques have been applied to MRC, including the encoder–decoder architecture and the pointer–generator network.

1.2.2 Existing issues

Although we have made remarkable achievements in many NLP tasks, there are still many problems that have not yet been well addressed, including the understanding of language structure and semantics. Many of these unsolved problems are also closely related to MRC.

1.2.2.1 The ambiguity of language

One of the characteristics of language is that it can express complex ideas with succinct statements. Thus it is common to have ambiguity in a sentence, that is, there are many reasonable interpretations. Here are some examples.

Example 1: *The fish is ready to eat.*

It can mean that the fish can start eating or the fish can be provided to someone to eat. The ambiguity comes from the different interpretations of the thematic role of the fish: whether it is the agent or patient of the action "eat."

Example 2: *David solved the problem too.*

Without context, it is hard to determine what fact "too" refers to. It can be that someone else solved the problem and David also did it. It can also be that David designed the problem and also solved it.

Example 3: *I saw a man on the hill with a telescope.*

The telescope could be in my hand (which I used to see the man) or with the man (I saw the man and his telescope), since both are valid under grammatic rules.

These are just some of the numerous examples of ambiguity in language. Even for humans, it is difficult to judge the true intentions of the speaker. However, if there is enough contextual information, most ambiguity can be eliminated. For example, if a cook says "the fish is ready to eat" before dinner, we know that this dish is ready for dining.

Nevertheless, many NLP models still struggle to understand the semantics of context. By analyzing the results of various models on tasks such as MRC, the researchers find that existing models are largely dependent on keyword or phrase matching, which greatly limits their capability to understand context and handle ambiguity.

1.2.2.2 Common sense and reasoning skills

In many cases, humans can reason from conversations to draw conclusions without explicit explanation. Here is an example dialogue of a customer booking tickets through customer service:

Agent: Hello, how can I help you?

Customer: I'd like to book a flight from San Francisco to New York in early May.

Agent: OK, when do you want to fly?

Customer: Well, I'm going to New York for a conference, which is from the 4th to 7th.

Agent: OK, here's the direct flight information from San Francisco to New York on May 3,...

Here, the customer does not directly answer the agent's question about the departure date. Instead, he gives the start and end date of the conference he will attend. As the flight must arrive in New York before the meeting starts, the agent infers that the departure date is May 3. And if the customer also needs a flight back to San Francisco from New York, the agent should give information about flights departing in the evening of May 7 or on May 8.

Therefore an automatic customer service model needs to infer information like the departure date from previous conversations. This inference requires the model to carry the common sense that the flight must reach its destination before the conference.

In recent years, there have been many efforts in applying common sense and reasoning to NLP. However, it remains an open question on how to equip a model with large-scale common sense and conduct effective reasoning.

1.3 Deep learning

Deep learning is currently one of the hottest areas of research in AI. Models based on deep learning play major roles in image recognition, speech recognition, NLP, and many other applications. The vast majority of MRC models nowadays are based on deep learning as well. Therefore this section will describe the characteristics of deep learning and the successful use cases.

1.3.1 Features of deep learning

Why can deep learning, as a branch of machine learning, stand out from the numerous directions? There are several important reasons as follows.

First, most deep learning models have a large model complexity. Deep learning is based on artificial neural networks (ANN), and one of the characteristics of ANN is that its model size is controllable: even with a fixed input dimension, the number of model parameters can be regulated by adjusting the number of network layers, number of connections, and layer size. As a result, deep learning makes it easy to increase model complexity to make a more efficient use of massive data. At the same time, studies have shown that the accuracy of deep learning models can increase with a larger size of data. As the field of MRC continues to evolve, more and more datasets emerge, making deep learning the most common machine learning architecture in reading comprehension.

Second, deep learning has a powerful **feature learning ability**. In machine learning, the performance of a model largely depends on how it learns a good representation of the data, that is, representation learning. Traditional machine learning models require a predefined procedure of extracting task-specific features. Prior to the advent of deep learning, feature extraction was often manual and required knowledge from domain experts. On the contrary, deep learning relies on neural networks to automatically learn effective feature representations via a nonlinear transformation on the primitive data features, for example, word vectors, picture pixels. In other words, deep learning can effectively obtain salient features that are helpful to the target task, without the need for model designers to possess special domain knowledge. As a result, it greatly increases the efficiency of designing deep learning models for tasks from various applications.

Third, deep learning enables **end-to-end learning**. Previously, many machine learning models proposed multistep solutions in the form of pipelines, such as feature learning → feature categorization → modeling each category → model ensembling. However, since each step can only be independently optimized, it is difficult to simultaneously optimize the whole system to improve its performance. Moreover, if any step within the model is updated, it is likely that all downstream steps have to be adapted as well, which greatly reduces the efficiency. One advantage of deep learning is that it enables end-to-end learning via the featurization ability of neural networks: feed in the raw data as input, and output the required result. This approach can optimize all parameters in an orchestrated manner to boost accuracy. For example, in MRC,

the model takes in the article and question text, and outputs the answer text. This greatly simplifies the optimization and is also easy to use and deploy.

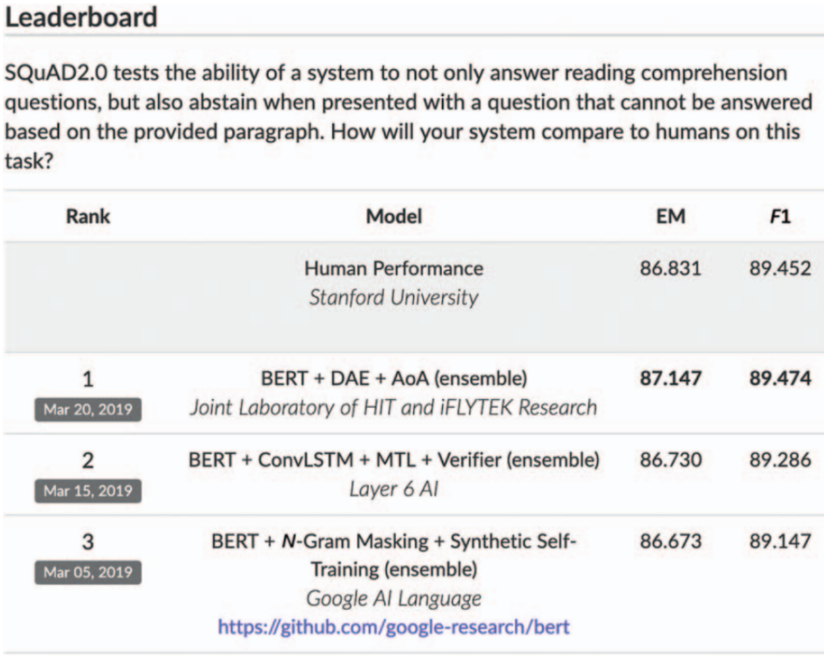
Fourth, the hardware for deep learning, especially **Graphics Processing Unit** (**GPU**), is being continuously upgraded. As deep learning models are usually large, computational efficiency has become a very important factor for the progress of deep learning. Fortunately, the improved design of GPU has greatly accelerated the computation. Compared with CPU, GPU has greater floating-point computing power, faster read−write speed, and better parallelism. The development of GPUs over the last decade follows the Moore's law of early day CPUs, where computing speed and device complexity grow exponentially over time. The GPU industry, represented by companies such as NVIDIA, Intel, and Google, continues to evolve and develop specialized GPUs for deep learning, contributing to the development and application of the entire deep learning field.

Fifth, the emergence and prosperity of deep learning frameworks and community immensely help prompt the booming of deep learning. With the advent of frameworks, such as TensorFlow, PyTorch, and Keras, neural networks can be automatically optimized and the most commonly used network modules are predefined, making deep learning development much simpler. Meanwhile, deep learning communities quickly thrive. Every time a new research result appears, there will be developers who immediately implement, validate, and open source models, making the popularization of new technologies to be at an unprecedented level. Academic paper repositories (e.g., arXiv) and open-source code platforms (e.g., GitHub) greatly facilitate the communication between researchers and developers, and considerably lower the threshold for participation in deep learning research. For example, within a few months of the publication and open source of the breakthrough Bidirectional Encoder Representations from Transformers (BERT) model in 2018 (more details in Chapter 6, Pretrained Language Model), models utilizing BERT had taken top places in MRC competitions such as SQuAD and CoQA (Fig. 1.1).

1.3.2 Achievements of deep learning

Since the advent of deep learning, it has achieved many remarkable results in various fields such as speech, vision, and NLP.

In 2009 the father of deep learning, Geoffrey Hinton, worked with Microsoft Research to significantly improve the accuracy of speech recognition systems through the Deep Belief Network, which was quickly



Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	*F1*
	Human Performance *Stanford University*	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + *N*-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147

Figure 1.1 The top three models in the machine reading comprehension competition SQuAD 2.0 are all based on BERT.

reproduced by IBM, Google, and HKUST. This is also one of the earliest success stories of deep learning. Seven years later, Microsoft further used a large-scale deep learning network to reduce the word error rate of speech recognition to 5.9%. This is the first time ever a computer model achieved the same performance as a professional stenographer.

In 2012 the deep learning model AlexNet achieved 84.6% in Top-5 accuracy in the large-scale image recognition contest ILSVRC2012, outperforming the second place by over 10%.

In 2016 Stanford University introduced the MRC dataset SQuAD, which includes 500 articles and over 100,000 questions. Just 2 years later, Google's pretrained deep learning model BERT reached an accuracy of 87.4% in exact match and 93.2% in *F*1 score, surpassing the human performance (82.3% in exact match and 91.2% in *F*1 score), which impressed the whole industry.

In 2018 Microsoft developed a deep learning translation system which for the first time achieved the same level of translation quality and accuracy as a human translator on the Chinese–English News dataset.

These achievements manifest the power of deep learning from different aspects, and also lay a solid foundation for its landing in the industry. However, we also observe that deep learning has some unresolved issues. For example, many deep learning models are still a "black box," making it impossible to explain how the model produces output for a particular input instance, and very hard to correct specific errors. In addition, most deep learning models lack the ability of reasoning, induction, and common sense. There are many ongoing researches to solve these issues. Hopefully, in the near future, deep learning will solve these problems and enable computers to have the same level of intelligence as humans.

1.4 Evaluation of machine reading comprehension

MRC is similar to the human reading comprehension task. Therefore it needs to be evaluated by the model's ability to understand the content of the article. Unlike mathematical problems, reading comprehension requires specific evaluation metrics for semantic understanding. It is well-known that the assessment of reading comprehension for humans is usually in the form of question and answer, in which the reader is asked to answer questions related to the article. Therefore the evaluation of MRC models can take the same form: the model answers relevant questions of the article and is evaluated by the answer quality. In this section, we will describe commonly used methods to evaluate a MRC model.

1.4.1 Answer forms

Most MRC tasks are assessed by the quality of their answers to given questions related to articles. The evaluation criteria depends on the form of answer. Here are some common answer forms:

- Multiple choice, that is, the model needs to select the correct answer from a number of options.
- Extractive, that is, the answer is bound to be a segment of text within the article, so the model needs to mark the correct starting and ending position of the answer in the article.
- Freestyle, that is, there is no limitation on the answer's text, which allows the model to freely generate answers.
- Cloze test, that is, certain keywords are removed from the article and the model needs to fill in the blanks with correct words or phrases.

In addition, some datasets design "unanswerable" questions, that is, a question that may not have a suitable answer given the article. In this case, the model should output "unanswerable" as the answer.

In the above forms, multiple choices and cloze tests can be objectively evaluated by directly comparing with the ground truth. Thus the accuracy can be used as the evaluation criterion. Extractive answers are of a semi-objective type. We can compare the model's output with the correct answer, and give a score of 1 when they are exactly the same, and 0 otherwise. This metric is called **Exact Match**. However, this will treat partially correct answers as wrong ones. For example, if the correct answer is "eight o'clock" and the model's output is "It's eight o'clock," the exact match score will be 0, although the model's output is very close to the correct answer. Therefore for extractive answers, the $F1$ metric is also often used, which is the harmonic mean of the precision and recall:

$$F_1 = 2 \times \frac{Precision \times recall}{Precision + recall}$$

Precision refers to the ratio of words in the model's output that also appear in the correct answer, and recall is the ratio of words in the correct answer that appear in the model's output. Thus the $F1$ metric can give a partial score when the model's output is partially correct. Table 1.1 shows an example of computing the Exact Match and $F1$ scores.

Freestyle answer is the most flexible form. The ideal metric should give full credit when the model's output has exactly the same meaning as the correct answer, and partial credit otherwise. However, it is a complex and unsolved problem to automatically judge whether two statements express the same meaning. On the other hand, human evaluation is very time-consuming and labor-intensive, while suffering from a high variance. Thus most widely used metrics for freestyle answers are based on the ratio of matched words/phrases between the model's output and the correct answer. These metrics include ROUGE, BLEU, and METEOR. We introduce the ROUGE metric in the next section.

Table 1.1 Exact match and $F1$ metrics.

Correct answer	Model's output	Exact match	Precision	Recall	$F1$
20 miles	is 20 miles	0	$2/3 = 0.66$	$1/1 = 1$	0.8
20 miles	20 miles	1	1	1	1

1.4.2 Recall-oriented understudy for gisting evaluation: metric for evaluating freestyle answers

ROUGE (**Recall-Oriented Understudy for Gisting Evaluation**) is a set of text similarity metrics based on recall [4]. It is used to measure the proportion of words and phrases in the correct answer that appear in the model's output. Because one question may have different expressions for the correct answer, ROUGE allows a set of reference answers for the same question.

ROUGE includes metrics such as ROUGE-N, ROUGE-S, and ROUGE-L. ROUGE-N measures the recall for N-grams, which is computed as follows:

$$\text{ROUGE} - N(M) = \frac{\sum_{A \in references} \sum_{s \in N-\text{grams in } A} \min\{count_s(A), count_s(M)\}}{\sum_{A \in references} \sum_{s \in N-\text{grams in } A} count_s(A)},$$

where M is the model's output, an N-gram is a phrase that consists of N neighboring words in the text, and $count_s(A)$ is the number of appearances of the N-gram s in A.

ROUGE-S is similar to ROUGE-2 ($N = 2$), but does not require the two words to be neighboring. Instead, the two words can be at most Skip words apart, where Skip is a parameter. For example, in "I like to run at night," if Skip = 2, "I like," "I to" and "I run" are all 2-grams in ROUGE-S.

ROUGE-L computes the longest common subsequence (LCS) between the model's output and the reference answer. The subsequence does not need to be contiguous in the original sequence. For instance, the LCS between "I want to have lunch" and "I forget to bring lunch to school" is "I to lunch" with a length of $L = 3$. Then, the ROUGE-L score is defined to be F_{LCS}, computed as follows:

$$R_{LCS} = \frac{LCS(correct\ answer, model's\ output)}{Length\ of\ correct\ answer}$$

$$P_{LCS} = \frac{LCS(correct\ answer, model's\ output)}{Length\ of\ model's\ output}$$

$$F_{LCS} = \frac{(1 + \beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}}$$

where β is a parameter. Table 1.2 shows an example of ROUGE-N, ROUGE-S, and ROUGE-L.

Table 1.2 Different ROUGE metrics.

Correct answer Model's output	I like this school I also like that school		
ROUGE-1 ROUGE-2 ROUGE-S Skip = 1	3/4 = 0.75 (I, like, school) 0/3 = 0 2/5 = 0.4 (I like, like school)		
ROUGE-L $\beta = 1$	Longest common subsequence (LCS): I like school		
	$R_{LCS} = 3/4 = 0.75$	$P_{LCS} = 3/5 = 0.6$	$F_{LCS} = 0.67$

There is a certain level of correlation between the ROUGE metric with human evaluation. However, there also exist discrepancies as it only measures lexical overlapping. Thus in addition to automatic metrics like ROUGE, freestyle answers are often manually evaluated on their correctness and naturalness.

1.5 Machine reading comprehension datasets

There are many public datasets in various NLP areas. Through the evaluation on these datasets, one can test the quality of models and compare the pros and cons. As a result, these datasets have greatly contributed to the development of related researches.

In MRC, there are also many datasets and competitions. Depending on the form of articles, we categorize these datasets into three types: single-paragraph, multiparagraph, and corpus. For single-paragraph and multiparagraph articles, the model can directly look for the answer within the article. For corpus-based articles, an information retrieval module is required. This module looks for the most relevant paragraphs or statements in the corpus based on the question, and then the model gets the answer within the retrieved results. In the following, we will describe the three types of MRC datasets with examples.

1.5.1 Single-paragraph datasets

A single-paragraph MRC dataset requires the model to answer questions about a given paragraph. During this process, the model does not need to refer to any external information. Thus this kind of dataset inspects the core reading comprehension ability of a model. As the construction of a single-paragraph dataset is relatively simple, it is the most common type in MRC.

1.5.1.1 RACE

RACE [5] is a large-scale English MRC dataset introduced by CMU in 2017. The dataset comes from the English tests for Chinese students. RACE contains 28,000 articles and nearly 100,000 multiple-choice questions. The model needs to select the correct answer from the options. The RACE dataset is divided into RACE-M, which is for middle school students, and RACE-H, which is for high school students. It is worth mentioning that in the process of answer collection, RACE uses an optical character recognition system to identify the answers from publicly available images.

1.5.1.2 NewsQA

NewsQA [6] is a news reading comprehension dataset from Maluuba in 2016, with more than 12,000 convolutional neural network (CNN) news articles and nearly 120,000 manually edited questions with extractive answers. One of the key goals of the NewsQA dataset is to access the model's reasoning and induction abilities, that is, to get the final answer based on information from different places in the article. Also, it provides "unanswerable" questions.

1.5.1.3 CNN/DailyMail

CNN/DailyMail [7] is a reading comprehension dataset by DeepMind in 2015, where the articles come from CNN and DailyMail. This dataset contains approximately 1.4 million instances, each containing an article, a question, and an answer. The CNN/DailyMail dataset adopts a cloze-style design. In order to let models focus on semantic understanding, the entities in the articles such as persons and places are replaced by ids. Therefore the model only needs to select the correct entity ids from the article to fill in placeholders in the question.

1.5.1.4 SQuAD

SQuAD [8] is the most influential and popular MRC contest, launched by Stanford University in 2016. The SQuAD dataset comes from 536 Wikipedia articles with more than 100,000 questions and extractive answers. SQuAD v2.0, which was launched in 2018, includes a large number of "unanswerable" questions which makes a total of 150,000 questions. The SQuAD dataset receives a lot of attention from both academia and industry for its massive size and high quality. As of December 2019, there are 294 submissions to SQuAD from teams around the world.

On October 5, 2018 Google's BERT model scored above the human level in SQuAD v1.1 for the first time, making headlines in the field of MRC.

1.5.1.5 CoQA

CoQA [9] is a multiturn conversational MRC competition introduced by Stanford University in 2018. Its distinguishing feature is the inclusion of context in QA dialogues, that is, multiple turns of questions and answers for each article. One needs to understand both the article and previous rounds of questions and answers to answer each question. This requires the model to have the ability to understand the context. There are more than 8000 articles and more than 120,000 questions in this dataset, with an average of 15 rounds of questions and answers per article. In addition, the test set of CoQA includes questions from domains unseen in the training set (Reddit forum and scientific questions) to test the generalization capability of models. CoQA contains extractive, "yes/no," "cannot answer," and a small number of freestyle answers. In March 2019 Microsoft's MMFT model achieved an $F1$ score of 89.4%, surpassing the human level of 88.8% for the first time, once again proving the effectiveness of MRC models.

1.5.2 Multiparagraph datasets

A multiparagraph MRC dataset requires the model to read multiple paragraphs and answer related questions. The correct answer may be in one paragraph, so the model needs to compute the correlation between the question and each paragraph; or the answer is obtained by collecting clues from multiple paragraphs, so the model must conduct multistep reasoning.

1.5.2.1 MS MARCO

MS MARCO [10] is a large MRC dataset launched by Microsoft in 2016, containing more than 1 million questions and more than 8 million articles. The questions in this dataset come from queries submitted by real users, while the relevant paragraphs are from Bing's search results for the query. MS MARCO adopts freestyle answers, and it has three tasks:

- determine whether the answer can be obtained from the given paragraph;
- generate the answer text; and
- sort multiple given paragraphs by their relevance to the question.

1.5.2.2 DuReader

DuReader [11] is a Chinese MRC dataset launched by Baidu in 2017. DuReader uses data from user queries and retrieved documents from Baidu search engine. In DuReader, the articles are full text from the webpages, instead of extracted paragraphs like in MS MARCO, making the task more challenging. In addition, due to the different standpoints of various articles, DuReader provides several candidate answers to some questions, which better aligns with real scenarios. DuReader contains 200,000 questions and 1 million documents, with both freestyle and yes/no types of answers.

1.5.2.3 QAngaroo

QAngaroo [12] is a multidocument reasoning MRC dataset introduced by University College London in 2017. It consists of two subsets: WikiHop from Wikipedia and MedHop from the abstract of the medical paper archive PubMed. The biggest feature of QAngaroo is that the answer cannot be drawn from a single paragraph. One must collect clues scattered across multiple paragraphs. As a result, QAngaroo requires the model to analyze multiple paragraphs and use multihop reasoning to get the answer. The dataset contains more than 50,000 questions and related documents, with multiple-choice answers.

1.5.2.4 HotpotQA

HotpotQA [13] is a multiparagraph reasoning MRC dataset introduced by Carnegie Mellon University, Stanford University, the University of Montreal, and Google in 2018. Similar to QAngaroo, HotpotQA requires the model to search for clues in multiple paragraphs and use multistep reasoning to get the answer. HotpotQA contains 110,000 questions and related Wikipedia paragraphs with extractive answers.

1.5.3 Corpus-based datasets

Corpus-based MRC datasets typically provide a large text corpus. The model should first find relevant paragraphs/articles in the corpus given the question, and then analyze the retrieved results to obtain the answer. Among the three types of MRC datasets, corpus-based dataset is closest to real applications such as QA in online search. Because corpus-based dataset does not limit the source of answers, it is also known as **Open Domain Machine Reading Comprehension**.

1.5.3.1 AI2 reasoning challenge

AI2 reasoning challenge (ARC) [14] is a corpus-based MRC dataset on sciences, launched by the Allen Institute of Artificial Intelligence in 2018. The questions in ARC come from 7800 scientific questions for US students in grades 3−9. All answers are in the form of multiple choices. ARC provides a large corpus of scientific text with 14 million sentences, which are the retrieved results by a search engine on scientific queries. Models are allowed to use both the corpus and external information to answer questions.

1.6 How to make an machine reading comprehension dataset

The previous section describes popular datasets and competitions for MRC. To ensure the quality of data, the generation of datasets and the acquisition of answers are often manually processed and verified. One common approach is **crowdsourcing**, that is, labelers are hired to generate and annotate data.

1.6.1 Generation of articles and questions

The three core concepts in any machine reading understanding dataset are articles, questions, and answers. Because the answer can be reasoned from the article based on the question, we often focus on how to collect articles and questions. It is not very common that both articles and questions can be automatically generated (e.g., a cloze dataset by randomly deleting keywords, or a QA dataset by converting reading comprehension tests for students). In most cases, only the articles or the questions are available for dataset makers. Thus one needs to employ labelers or use algorithms to get the other part of information, either generating questions from articles or generating articles from questions.

1.6.1.1 Generating questions from articles

MRC datasets often leverage publicly available corpora as article sources, such as Wikipedia and news reports. However, most of these articles are without related questions. Therefore the dataset makers need to employ labelers to generate questions relevant to the articles. These labelers can decide the question language, but must ensure the quality and difficulty, as well as the relevance to the article. For example, the SQuAD dataset uses paragraphs from Wikipedia as articles and leverages the crowdsourcing platform Amazon Mechanical Turk

to hire labelers to generate up to five questions per article. To ensure label quality, each labeler should have a minimum of 97% approval rate for their historical labeling tasks with at least 1000 labels. Labelers are exposed to detailed instructions with examples of high-quality and low-quality questions. For instance, they must use their own language to formulate questions, rather than directly copying statements from the article. Each labeler's compensation is calculated by the time spent and the number of generated questions. In SQuAD dataset, each labeler is paid $9/hour to generate questions for at least 15 paragraphs.

The advantage of generating questions from articles is that the question is very relevant to the article and does not include external information. So this method is suitable for generating single-paragraph and multiparagraph MRC datasets. However, because the questions are artificially produced, it is inevitable that some questions are unnatural. This may cause discrepancies in focus points and language patterns between the generated questions and the questions raised by users in real applications.

1.6.1.2 Generate articles from questions

With the popularity of forums and search engines, online users issue a large number of questions and queries every day. By filtering appropriate questions from the history logs, one can collect a massive repository of questions for MRC tasks. Next, related articles are obtained by searching the web or large text corpora. The top retrieved documents and paragraphs can be used in the dataset.

The advantage of generating articles from questions is that all the questions come from real user queries, which are similar to questions in real applications. However, this approach cannot guarantee that the correct answer exists in the retrieved article, so a crowdsourced validation is required.

1.6.2 Generation of correct answers

Most MRC comprehension datasets employ manual answer generation, often in the form of crowdsourcing. The labelers need to provide concise answers in the required form, for example, extractive and freestyle. To ensure the quality of the answer, it is often necessary for multiple labelers to annotate the answer to the same question. However, due to subjective bias of the labelers, it is difficult to guarantee consistency among all answer candidates. One solution is majority voting, that is, use the answer that most people agree with. However, if there is too much disagreement

among labelers, the answer has to be relabeled. Here, the degree of agreement between answers can be estimated by exact match, *F*1 score, or ROUGE, introduced in Section 1.4.

Labelers are typically given a software or web environment to annotate answers in the specified format. For example, the labeling interface for the conversational MRC dataset CoQA includes articles, questions, and answers in previous rounds, answer specifications (e.g., prefer shorter answers, reuse words from the article), and compensation information.

Since MRC datasets are labeled by humans, we often see descriptions like "human performance" or "human level" in related reports. In some datasets such as SQuAD and CoQA, the best machine learning models have already achieved performance beyond human levels. So, how is the human level defined?

First of all, we can never get absolutely correct answers to all questions, since each labeler has subjective bias and it is impossible to unify the length and variation of all answers. Even if multiple labelers vote, there is no guarantee that the answer will be accurate. Therefore the absolutely correct answers, that is, results with a true accuracy of 100%, are unattainable.

Therefore in datasets that require manual labeling, the set of answers or voting results from multiple labelers are deemed as correct answers with an accuracy of 100%. And **the human level refers to the accuracy of a single labeler**. For example, in SQuAD v1.1, three labelers write answers for each question, and the human level refers to the accuracy of the second labeler's answers, compared against the answers from the first and third labeler.

Because a single labeler is more likely to make mistakes in answers than multiple labelers, the human level score is generally less than 100%, which means that it is possible for a good machine learning model to exceed the human level. Fig. 1.2 shows the accuracy of computer models and human-level accuracy. Note that if a different group of labelers are employed, the correct answers may change, and the defined accuracy of computer models and human-level accuracy will alter accordingly.

1.6.3 How to build a high-quality machine reading comprehension dataset

A successful MRC dataset should be able to accurately assess whether a model has reading ability comparable to humans. It should also effectively identify and compare the strengths and weaknesses of different models. A high-quality dataset can greatly boost the development of MRC research.

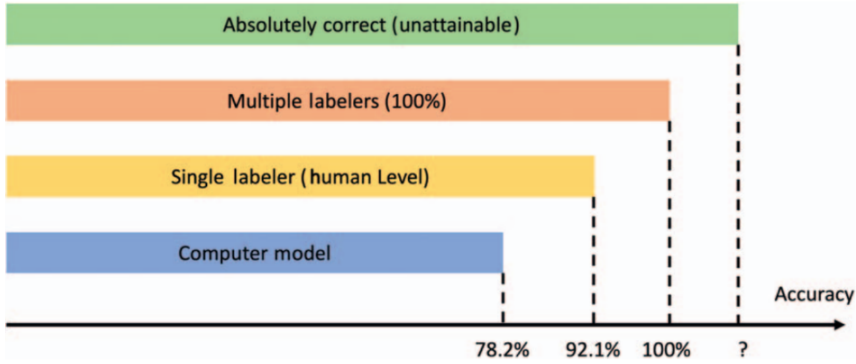


Figure 1.2 Comparison between the accuracy of computer model and human-level accuracy.

In this section, we will analyze what features a high-quality MRC dataset should have.

1.6.3.1 Distinguishing comprehension-based and matching-based models

After models achieved excellent results in various MRC competitions, researchers analyzed the mechanism by which the model generated the answers. The results reveal that the most models heavily rely on lexical matching between the question and the article, rather than understanding of the content, which often leads to comprehension errors. These errors indicate that MRC models are overdependent on the text matching between the question and the article when seeking answers.

The researchers further found that if a misleading sentence containing question keywords yet semantically irrelevant with the question is appended to each article in the SQuAD dataset, the *F*1 scores for all models drop by as much as 20%–40% [15]. However, this misleading statement has little effect on human performance.

Therefore if a dataset mostly contains questions that can be answered by simple text matching, it is hard to distinguish high-quality models from text-matching algorithms. It follows that a high-quality MRC dataset should punish text-matching models in the evaluation. For example, when building the SQuAD v2.0 dataset, the labelers were asked to generate "unanswerable" questions using a combination of keywords in the article. These questions account for 35% of all the questions in the dataset, which effectively reduces the accuracy of text-matching models. This design has been adopted as a new standard in many MRC datasets.

1.6.3.2 Evaluate the reasoning capability

Humans can reason and induct when reading articles. One example is that they can collect and summarize various clues in the article and infer the answer. For instance, an article mentions "David is from California" in one place, and "California is a state in US" in another place. So the answer to the question "Which country does David come from" should be "US." This type of question requires multistep reasoning and cannot be inferred by matching-based models. Thus reasoning is one of the key factors that MRC datasets should assess the models on.

In current MRC datasets that promote reasoning, for example, HotpotQA in Section 1.5, most questions are manually generated by labelers to include information from multiple paragraphs. However, questions created in this way may be unnatural and very different from those of real users. Another possible approach is to pick questions that require reasoning among real user queries, retrieve relevant documents, and then divide the text into paragraphs for multistep reasoning.

1.6.3.3 Assess common sense

Common sense is one of the most important cognitive abilities of humans. Common sense includes spatial knowledge like "big objects can't be put into a smaller space," temporal knowledge like "the year of 2016 is before 2019," and physical knowledge like "under normal conditions, the water will boil when heated to 212°F." The understanding and application of common sense is so far the weak point of NLP and machine learning. The reasons are that (1) common sense knowledge is massive in scale, vaguely defined, and difficult to summarize in its entirety, and (2) it is difficult to effectively express and apply common sense.

In reading comprehension, many questions involve common sense logic. For example, suppose the article is "David is having fun with Mary and Tom at home. After a while Tom left," and the question is "How many people are now in David's home?" To answer the question, the model should understand that "left" means that Tom is no longer in David's home, and it needs to do some math to figure out the answer is "two people." In another example, the article is "Kevin needs to go to the company for an interview at 9:00 a.m. It is lunchtime now, but he's still at home," and the question is "Did Kevin go to the interview?" Clearly, the model has to understand that "lunchtime" is after "9:00 a.m." and "at home" indicates that Kevin has not visited the company. Therefore the correct answer is "No."

Although common sense is commonly used in reading comprehension, few MRC datasets exist to investigate this area. In 2018 researchers from Tel

Aviv University and Allen Institute of Artificial Intelligence launched a common sense MRC dataset CommonsenseQA with 12,000 questions, based on the knowledge graph ConceptNet. Thus CommonsenseQA examines a model's understanding of structured knowledge.

MRC datasets for common sense also need to discern and punish simple text-matching methods. One possible approach is to let the model give the employed common sense when giving answers.

1.6.3.4 Other comprehension skills

In 2017 some researchers pointed out that MRC models should have 10 basic skills, such as understanding coreference, logical reasoning, common sense, and mathematical knowledge. A comprehensive list is given in Table 1.3.

1.6.3.4.1 List/enumeration

Models need to identify, summarize, and sequentially output related concepts in the article, such as the answer to the question "What are the categories of organisms on Earth?"

1.6.3.4.2 Mathematical operations

Some MRC questions necessitate basic math knowledge. For example, suppose the article is "Brian was playing basketball when James and Linda joined. Then

Table 1.3 Ten skills of machine reading comprehension models [16].

Skills	Details
List/enumeration	Tracking, retaining, and list/enumeration of entities or states
Mathematical operations	Four arithmetic operations and geometric comprehension
Coreference resolution	Detection and resolution of coreference
Logical reasoning	Induction, deduction, conditional statement, and quantifier
Analogy	Trope in figures of speech, for example, metaphor
Spatiotemporal relations	Spatial and/or temporal relations of events
Causal relations	Relations of events expressed by why, because, the reason for, etc.
Commonsense reasoning	Taxonomic and qualitative knowledge, action, and event changes
Schematic/rhetorical clause relations	Coordination or subordination of clauses in a sentence
Special sentence structure	Scheme in figures of speech, constructions, and punctuation marks

Linda left with her mom." and the question is "How many people are playing basketball now?" Obviously, the answer cannot be simply extracted from the text but has to be obtained via mathematical computing.

1.6.3.4.3 Coreference resolution

Coreference resolution is an important task in NLP. Its goal is to understand the referred object of pronouns, for example, this, that, he, and she, to answer the questions.

1.6.3.4.4 Logical reasoning

The model needs to derive inferred facts from the article to get answers via reasoning. For example, from the statement "I asked Wendy whether she wanted to eat at home or at the restaurant, and she said she'd like to have a walk," we can infer that Wendy chose to eat at the restaurant.

1.6.3.4.5 Analogy

To answer certain questions, the model should understand common rhetorical techniques like metaphors and analogies.

1.6.3.4.6 Spatial–temporal relations

Spatial–temporal relations are common topics of questions. For instance, suppose the article is "I ask Bob to come to the company on Wednesday. Since he had some personal errands, Bob came the next day." The model should infer that Bob came to the company on Thursday.

1.6.3.4.7 Causal relations

Understanding causality is important to answer questions of why.

1.6.3.4.8 Common sense reasoning

Questions involving common sense and logic require the model to conduct common sense reasoning.

1.6.3.4.9 Schematic/rhetorical clause relations

Language structures like clauses contain rich semantic information, such as descriptions of entities and references, which are challenging for MRC models.

1.6.3.4.10 Special sentence structure

Some less frequent language structures such as inverted order and subjunctive make it difficult to understand the article.

These reading comprehension skills cover many reading skills that humans acquire, in the aspects of language, rhetoric, and external knowledge. Therefore an MRC task should evaluate the model's skills from all these aspects. Current researches, especially those related to deep learning, still heavily rely on statistical pattern recognition and matching. Thus most models lack many of the above skills and hence the ability to give reasonable explanations of the given answer. These are important topics to address in MRC research.

1.7 Summary

- **Machine reading comprehension** (**MRC**) is similar to the reading comprehension task for humans, where a model's ability to understand an article is assessed by answering related questions.
- MRC can be applied in scenarios where it is desired to automatically process a large amount of text data and understand the semantics.
- **Natural Language Processing** has many subfields closely related to MRC, such as information retrieval and question-and-answer systems.
- **Deep learning** is one of the hottest research directions in artificial intelligence. It greatly improves the accuracy of models in many areas. The majority of current MRC models are based on deep learning.
- In MRC, the answer types include multiple-choice, extractive, free-style, and cloze test.
- Depending on the form of articles, MRC datasets can be categorized into three types: **single-paragraph**, **multiparagraph**, and **corpus**.
- To build an MRC dataset, one can manually generate questions based on articles from public sources. Another way is to retrieve articles using questions from search engines and forums.
- **Human level** refers to the accuracy of a single labeler, judged by the set of annotated answers from multiple labelers.
- A high-quality MRC dataset should effectively distinguish between understanding-based and matching-based models. It should also access models' reasoning and common sense ability.

References

[1] Snow C. Reading for understanding: toward an R&D program in reading comprehension. Rand Corporation; 2002.
[2] Lehnert WG. The process of question answering (No. RR-88). New Haven, CT: Yale University, Department of Computer Science; 1977.

[3] Hirschman L, Light M, Breck E, Burger JD. Deep read: a reading comprehension system. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics; June 1999. pp. 325–32.
[4] Lin CY. Rouge: a package for automatic evaluation of summaries. Text summarization branches out July 2004;74–81.
[5] Lai G, Xie Q, Liu H, Yang Y, Hovy E. Race: large-scale reading comprehension dataset from examinations. arXiv Preprint arXiv 2017;1704:04683.
[6] Trischler A, Wang T, Yuan X, Harris J, Sordoni A, Bachman P, et al. NewsQA: a machine comprehension dataset. arXiv Preprint arXiv 2016;1611:09830.
[7] Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, et al. Teaching machines to read and comprehend. Adv Neural Inf Process Syst 2015;1693–701.
[8] Rajpurkar P, Zhang J, Lopyrev K, Liang P. Squad: 100,000 + questions for machine comprehension of text. arXiv Preprint arXiv 2016;1606:05250.
[9] Reddy S, Chen D, Manning CD. CoQA: a conversational question answering challenge. Trans Assoc Comput Linguist 2019;7:249–66.
[10] Nguyen T, Rosenberg M, Song X, Gao J, Tiwary S, Majumder R, et al. Ms marco: a human-generated machine reading comprehension dataset; 2016.
[11] He W, Liu K, Liu J, Lyu Y, Zhao S, Xiao X, et al. Dureader: a chinese machine reading comprehension dataset from real-world applications. arXiv Preprint arXiv 2017;1711:05073.
[12] Welbl J, Stenetorp P, Riedel S. Constructing datasets for multi-hop reading comprehension across documents. Trans Assoc Comput Linguist 2018;6:287–302.
[13] Yang Z, Qi P, Zhang S, Bengio Y, Cohen WW, Salakhutdinov R, et al. Hotpotqa: a dataset for diverse, explainable multi-hop question answering. arXiv Preprint arXiv 2018;1809:09600.
[14] Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C, et al. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv Preprint arXiv 2018;1803:05457.
[15] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems. arXiv Preprint arXiv 2017;1707:07328.
[16] Sugawara S, Yokono H, Aizawa A. Prerequisite skills for reading comprehension: multi-perspective analysis of mctest datasets and systems. In: Thirty-first AAAI conference on artificial intelligence; February 2017.


image
not
available

In real applications, BPE or other vocabulary-independent tokenizers are usually applied when the model needs to generate text, for example, freestyle answers in MRC; while vocabulary-based tokenizer are used in classification tasks, for example, multiple-choice and extractive answers in MRC.

2.2 The cornerstone of natural language processing: word vectors

The input to a machine reading comprehension model is the text of article and question. However, computers conduct numerical operations and cannot directly manipulate on characters or strings. Therefore the text needs to be converted into numeric representations. A common approach is to represent each word by a list of numbers, that is, word vector. This section describes several methods to generate word vectors.

2.2.1 Word vectorization

In computers, any string needs to be converted into the binary format for computation and storage. For example, in C language, it takes 1 byte (8 bits) to store a character in ASCII code, for example, the letter *A* is represented by the 8-bit binary code *01000001*.

Since words are strings, it is convenient to store a word by concatenating the numeric representations of its characters into a vector, that is, word embedding. However, this approach has two drawbacks. First, this vector only represents the spelling, not the semantic meaning of the word. But it is essential to make the computer understand the meaning of words, sentences, and paragraphs. Second, the word embedding becomes longer when the word has more characters, making it difficult to understand long and complex words. Also the nonuniform length of representation for different words can cause problems for subsequent processing.

To solve the above problems in word vectorization, researchers propose one-hot embedding and distributed representation.

2.2.1.1 One-hot embedding

In **One-hot Embedding**, the embedding of a word is a vector of the same size of the dictionary. The vector contains zero everywhere except one at the position of the word in the dictionary. For example, suppose the dictionary is [apple, pear, banana, peach]. Because the size of the dictionary is 4, the one-hot embedding is a vector of length 4. As *pear* is the


image
not
available


image
not
available

Index

Note: Page numbers followed by "*f*" and "*t*" refer to figures and tables, respectively.

F

G

H

I