硕士学位论文

融合外部知识的机器阅读理解方法研究

REASEARCH ON MACHINE READING COMPREHENSION METHODS BASED ON INCORPORATING EXTERNAL KNOWLEDGE

乐远

哈尔滨工业大学

2020 年 6 月

国内图书分类号：TP391.1　　学校代码：10213
国际图书分类号：004.8　　密级：公开

工程硕士学位论文

融合外部知识的机器阅读理解方法研究

硕士研究生：乐 远
导　　师：张宇教授
申请学位：学术硕士
学　　科：计算机科学与技术
所在单位：计算机科学与技术学院
答辩日期：2020 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.1
U.D.C: 004.8

Dissertation for the Master's Degree in Engineering

REASEARCH ON MACHINE READING COMPREHENSION METHODS BASED ON INCORPORATING EXTERNAL KNOWLEDGE

Candidate: Le Yuan
Supervisor: Prof. Zhang Yu
Academic Degree Applied for: Master of Engineering
Speciality: Computer Science and Technology
Affiliation: School of Computer Science and Technology
Date of Defence: June, 2020
Degree-Conferring-Institution: Harbin Institute of Technology

摘　要

人类阅读理解和机器阅读理解一个很大的差异是，人类很善于利用除了文本之外的一些外部知识，来辅助自己理解获取答案。然而当前的很多机器阅读理解方法更多的是在文本匹配层面，仅仅是根据阅读理解所提供的文本和问题来寻找答案。但是现实世界中的机器阅读理解任务很复杂，仅仅根据所提供的文本和问题，无法获得问题的答案，需要借助一些常识性的外部知识信息。

本文以为机器阅读理解引入外部知识为切入点，通过检索机器阅读理解任务相关的外部知识信息，然后设计某种方法将其加入到机器阅读理解的获取问题答案的过程中，从而提高机器阅读理解获取问题的答案的性能。主要进行了以下三项研究工作：

（1）基于预训练语言模型的隐式的引入外部知识的方法。由于预训练语言模型借助优秀的深度学习架构，能够很好的利用大量的无标注的文本，而这些大量的无标注的文本中已经包含很多的知识，因此，直接使用预训练模型构建机器阅读理解模型来隐式地引入外部知识，相比于的传统的阅读理解方法，在实验的测试集合上取得了很不错的效果。

（2）基于注意力机制的显式的引入外部知识的方法。针对当前很多外部知识库如 NELL、WordNet 等都包含丰富的知识信息，使用适当的方法检索出相关的知识之后，利用注意力机制设计知识融合模块显示的将这些外部知识融合到现有的机器阅读理解模型当中，实验结果证明了该方法的有效性。

（3）结合实体感知增强的引入外部知识的方法。当前的以预训练语言模型为基础的机器阅读理解模型对文本进行分词之后会将部分实体拆开，而很多检索到的外部知识都是实体级别的，这会影响机器阅读理解融合实体相关的外部知识，针对这个问题，提出一个结合实体感知增强和外部知识的阅读理解，该方法添加了命名实体识别的辅助任务，该任务和机器阅读理解任务一起联合训练，增强了机器阅读理解模型的实体感知能力，从而进一步提高机器阅读理解获取问题答案的能力，实验结果表明了该方法的有效性。

关键词：机器阅读理解；预训练语言模型；注意力机制；命名实体识别；多任务学习

Abstract

A big difference between human reading comprehension and machine reading comprehension is that human beings are very good at using some external knowledge other than text to assist themselves in understanding and obtaining answers. However, many current methods of machine reading comprehension are more at the level of text matching, only to find answers based on the text and questions provided by reading comprehension. However, the task of machine reading comprehension in the real world is very complicated. Based on the text and questions provided, the answers to the questions cannot be obtained, and some common-sense external knowledge information is needed.

This article takes the introduction of external knowledge for machine reading comprehension as an entry point, by retrieving external knowledge information related to the machine reading comprehension task, and then designing a method to add it to the process of obtaining the answer to the question of machine reading comprehension, thereby improving the machine reading comprehension The performance of the answer to the question. Mainly carried out the following three research work:

(1) A method of implicitly introducing external knowledge based on pre-trained language models. Due to the excellent deep learning architecture, the pre-trained language model can make good use of a large amount of unlabeled text, and these large amounts of unlabeled text already contain a lot of knowledge, so directly use the pre-trained model to build machine reading comprehension, the model implicitly introduces external knowledge. Compared with the traditional reading comprehension method, it has achieved very good results in the experimental test set.

(2) An explicit introduction of external knowledge based on the Attention mechanism. In view of the fact that many current external knowledge bases such as NELL and WordNet contain rich knowledge information, after using appropriate methods to retrieve relevant knowledge, the attention mechanism is used to design the knowledge fusion module to display these external knowledge

into the existing machine reading comprehension. In the model, the experimental results prove the effectiveness of the method.

(3) The method of introducing external knowledge combined with entity perception enhancement. The current machine reading comprehension model based on the pre-trained language model will split some of the entities after word segmentation of the text, and many of the external knowledge retrieved are at the entity level, which will affect the machine reading comprehension fusion entity related external For this problem, a reading comprehension combining entity perception enhancement and external knowledge is proposed. This method adds an auxiliary task of named entity recognition. This task and the machine reading comprehension task are jointly trained together to enhance the entity perception of the machine reading comprehension model. The ability to further improve the machine reading comprehension ability to obtain the answer to the question, the experimental results show the effectiveness of the method.

Keywords: Machine Reading Comprehension, Pretrained Language Model, Attention Method, Named Entity Recognition, Multi-task Learning

目录

摘　要 I
Abstract II
第 1 章 绪　论 1
1.1 课题背景及研究目的和意义 1
1.1.1 课题背景 1
1.1.2 课题研究的目的和意义 2
1.2 国内外研究现状 3
1.2.1 国外研究现状 5
1.2.2 国内研究现状 6
1.3 主要研究内容 7
1.4 本文章节安排 8
第 2 章 基于预训练语言模型的隐式引入外部知识的方法 10
2.1 引言 10
2.2 实验数据集介绍 10
2.3 预训练语言模型介绍 11
2.3.1 传统的语言模型训练方法 11
2.3.2 预训练语言模型 13
2.4 基于预训练语言模型的机器阅读理解模型 16
2.5 实验结果与分析 16
2.5.1 基于预训练语言模型的机器阅读理解模型实验结果 16
2.6 本章小结 17
第 3 章 基于注意力机制的融合外部知识的方法 18
3.1 引言 18
3.2 机器阅读理解中常用的 Attention 机制 19
3.2.1 Attentive Reader 和 Impatient Reader 19
3.2.2 Attentive Sum Reader 20
3.2.3 斯坦福 Attentive Reader 21
3.2.4 AOA Reader 22
3.2.5 双向注意力机制 23
3.2.6 自匹配注意力机制 25
3.3 实验数据集与外部知识库 26

3.4 基于注意力机制的融合外部知识的机器阅读理解方法 26
3.4.1 外部知识检索 26
3.4.2 基于注意力机制的融合外部知识的机器阅读理解模型 27
3.5 实验设置与结果分析 30
3.5.1 评价指标 30
3.5.2 实验设置与结果分析 30
3.6 本章小结 31
第 4 章 结合实体感知增强的引入外部知识的方法 33
4.1 引言 33
4.2 引入外部知识的阅读理解主要技术介绍 34
4.2.1 知识感知的双向 LSTMs 34
4.2.2 命名实体识别主要技术介绍 35
4.4 结合实体感知增强和外部知识的机器阅读理解 37
4.4.1 命名实体识别辅助任务数据构造 37
4.4.2 外部知识检索 38
4.4.3 结合实体感知增强与外部知识的阅读理解模型 39
4.5 实验设置与结果分析 42
4.6 本章小结 45
结 论 46
参考文献 48
攻读硕士学位期间发表的论文及其它成果 53
哈尔滨工业大学学位论文原创性声明及使用授权说明 54
致 谢 55

第 1 章 绪　论

1.1 课题背景及研究目的和意义

1.1.1 课题背景

阅读理解是我们从中考到高考都熟悉不过的很常见的一种题型，该题型设置的目的在于广泛地考察读者的阅读理解能力，要求读者能读懂领会并解释词语、句子、段落的意思，而且要求能筛选出材料中的重要信息，分析相关现象和问题，并归纳整合，最后给出问题的答案。

学者 C.Snow 在 2002 年的一篇论文中定义阅读理解是“通过交互从书面文字中提取与构造文章语义的过程”。而机器阅读理解的目标是利用人工智能技术，使计算机具有和人类一样理解文章的能力。

近年来，人工智能快速发展，尤其是深度学习的迅猛发展，已在图像识别、语音识别和围棋等领域超越了人类水平。作为实现人工智能的核心技术的自然语言处理，借助于深度学习，在机器翻译、人机对话等方面也取得了重大突破。自然语言处理一直被认为是 AI 领域最难的任务之一，而机器阅读理解是自然语言处理领域最上层的任务，是自然语言处理的核心。

深度学习的快速发展，使得机器阅读理解的研究受到了学术界和工业界的广泛关注。一方面，机器阅读理解是人工智能的一个重要的体现形式，对其深入研究有助于推动人工智能和自然语言处理领域的发展；另一方面，机器阅读理解利用人工智能技术为计算机赋予了阅读、分析和归纳文本的能力，满足了信息爆炸时代人们自动化、智能化地获取信息的需求，在工业界中的众多领域和人们生活的方方面面都有着广阔的应用空间，如搜索引擎、智能客服机器人、智能音响等等。

从 2017 年至今，作为自然语言处理的最上层的任务，机器阅读理解相关的数据集和新的模型方法呈现爆发式的增长，这为机器阅读理解的发展奠定了坚实的基础，机器阅读理解的研究空前繁荣，这些也促使了很多机器阅读理解的工业应用的出现。

本课题来源于社会计算与信息检索研究中心 QA 组问答系统研究项目。

1.1.2 课题研究的目的和意义

当前的很多机器阅读理解的数据集任务，都是根据所提供的文本，来寻找问题的答案。传统的机器阅读理解技术采用基于规则的模式匹配方法，或者借鉴信息抽取的方法构造关系数据库来寻找答案。这些方法不仅效率低下，且准确率不高。随着深度学习技术的日趋成熟以及各种大规模的数据集的出现，基于深度学习的机器阅读理解也取得了巨大的进步，出现了很多优秀的机器阅读理解模型和方法。这些基于深度学习的机器阅读理解方法都有一个通用的一般架构，包含编码层、交互层、输出层等。其中编码层负责编码每个单词、短语和句子，交互层负责建立文章和问题之间的联系，输出层负责预测问题的答案，这些新的模型和方法在很多机器阅读理解任务上表现都很不错。

然而在现实世界的很多实际应用当中，很多阅读理解的问题都很复杂，文本内容也不是那么简单，仅仅通过有限的文本并不能很好的去理解，更不谈去准确的回答问题，因此机器阅读理解在实际应用中仍有很多缺陷，效果不太好。这其中很大的一个原因是，和机器阅读理解仅仅利用所提供的文本不同，人类在做阅读理解的时候，会很好的去利用一些外部经验和外部知识来辅助理解，这也正是机器阅读理解和人类阅读理解的一个巨大差异，这是当前很多机器阅读理解所存在的问题。

然而当前很多的机器阅读理解方法都只是关注所提供的文本，没有考虑结合外部知识。所以，如果能够设计某种方法让机器阅读理解能够很好的利用外部经验或者外部知识，将一定能提高机器阅读理解的理解能力，帮助机器阅读理解更准确的获取问题的答案，从而增强机器阅读理解的工业落地应用，使得机器阅读理解能够应用在更复杂的场景，这将对机器阅读理解的发展起到极大的促进作用。而且，融合外部知识的机器阅读理解的方法甚至还可以迁移到其他的自然语言处理任务上，这将对自然语言处理的发展乃至人工智能的发展都具有重大意义，因此，融合外部知识的机器阅读理解方法的研究课题是一个非常有意义而且很具有挑战性的课题。

1.2 国内外研究现状

早在上世纪 70 年代初期，MIT 人工智能实验室的 Eugene Charniak 发表了一篇题为“Toward a model of children’s story comprehension”的技术报告，是最早探索机器阅读理解技术的论文。此后，对机器阅读理解的研究未曾中断，研究者们陆陆续续发布了非常多的机器阅读理解论文、系统及数据集。近年来，由于深度学习技术的迅猛发展和大规模数据集的出现，机器阅读理解任务的研究热度更是达到了前所未有的新高度。

机器阅读理解是指让计算机能够像人一样阅读、理解文本，而且能够在理解的基础上，正确回答与所读文本相关的问题[1]。

机器阅读理解类似于传统的问答任务[2]，其主要目的都是为了考察机器对文本的理解能力和在理解基础之上的推理能力。与问答任务的区别在于，问答任务往往具有一个较大的文档数据库，回答问题将会在文档数据库中进行检索或抽取，而阅读理解任务则是在单个文档上进行文本理解和问题回答。

现阶段的各项底层自然语言处理研究任务，比如句法分析、语义角色标注、命名实体识别、文本蕴含识别、浅层语义分析等，研究者们进行了深入的研究，并取得了很大的进步。但这些底层的研究任务相对独立，对于达到自然语言处理的最终目标究竟有多大帮助，尚且不得而知。机器阅读理解恰恰是综合各项自然语言处理任务，并对其进行检测和评估的一种良好方式[3]，同时从篇章的角度探究语言理解的技术和方法，又能刺激其他自然语言处理相关任务的进步。

同时，机器阅读理解需要深度挖掘文本的语义，这使得对语义挖掘有深度需求的问答、文本推理等高层自然语言处理任务都有着直接帮助。对文章的深层理解将会使阅读理解不再是一个让人望而却步的问题，我们能够结合题干信息去准确的回答问题。

传统机器阅读理解方法作为早期探索的成果，有很多不足之处，比如机器阅读理解系统的实现需要大量的和应用场景有关的特征工程相关的工作，不仅不容易扩展到新的文本和新的应用领域，而且不适用于当今大规模的文本数据和快速处理的实际应用场景。

如今主流的机器阅读理解方法可分为基于特征工程的方法和基于深度学习的机器阅读理解方法。机器阅读理解和人在阅读理解过程中面临的问题是

类似的，不过为了降低任务难度，目前很多机器阅读理解的研究都不使用现存的网络文本数据，而是采用人工构造的比较简单的语料，并回答一些与语料原文相关的较为简单的问题。

而关于需要融和外部知识的机器阅读理解，有两个大的难点需要考虑：第一，如何准确的检索到相关的外部知识。外部知识有很多种，大多数以知识图谱的形式进行存储，如何从知识图谱中检索到相关的外部知识直接影响到最后机器阅读理解的效果。例如，“苹果”一词可以指“水果”也可以指“苹果电脑”，那么引入该词相关的哪个语义直接决定着最后的结果。第二，怎么更好的融合外部知识。问题和文本都的表示形式是文本，它们是非结构化的，然而许多外部知识的表示形式是知识图谱，它们的表示都是结构化的，因此如何把这些结构化的知识和非结构化的文本表示结合起来是一个很具有挑战性的问题。

近几年机器阅读理解相关的数据集和论文呈现爆发式增长，可以看到这个领域的火爆，当然从另一个侧面也可以看出该项任务的极大的研究价值。像 2015 年发布的 CNN/DM[4]数据集、2016 年的 SQuAD[5]数据集、2017 年的 SearchQA[6]数据集、2017 年的 TriviaQA[7]以及 2016 年的 MS-MARCO[8]数据集，这些高质量的数据集的发布极大的促进了机器阅读理解领域的发展，促使了很多端到端的机器阅读理解模型的出现，如 Match-LSTM[9]、BiDAF[10]、AOA-Reader[11]、DCN[12]、R-Net[13]以及 QA-Net[14]。

最近，随着预训练语言模型的进一步发展，像 ElMo[15]、GPT[16]以及 Bert[17]等的出现，再一次刷新了机器阅读理解的各项指标，甚至像 SQuAD 等任务首次超过人类。然而对于一些需要外部经验和知识的阅读理解任务来说，这些端到端的模型以及预训练语言模型也并不是那么万能。在人类阅读理解过程中，当有些问题不能根据给定文本进行回答时，人们会利用常识或积累的背景知识进行作答，而在机器阅读理解任务中却没有很好的利用外部知识，这是机器阅读理解和人类阅读理解存在的差距之一。因此，如果能够找到比较好的融合外部知识的机器阅读理解方法，将会提高机器阅读理解的理解能力、推理能力以及可解释性，这会对机器阅读理解乃至整个自然语言处理的发展具有重要意义。

1.2.1 国外研究现状

从2015年开始至2018年,陆续出现了很多机器阅读理解相关的数据集，像 2015 年发布的 CNN/DM[4]数据集、2016 年的 SQuAD[5]数据集、2017 年的SearchQA[6]数据集、2017年的TriviaQA[7]以及2016年的MS-MARCO[8]数据集，这些数据集的发布极大的促进了端到端的机器阅读理解技术的发展，产生了一系列的经典的机器阅读理解模型，像 Match-LSTM[9]、BiDAF[10]、AOA-Reader[11]、DCN[12]、R-Net[13]以及QA-Net[14]。这些端到端的机器阅读理解模型具有相似的架构，首先使用一个编码层获得问题和文章的向量表示，然后在接一个基于注意力机制的一个交互层获得问题和文章的交互信息表示，最后再使用一个预测层例如常用的Pointer-Network获得最后的答案片段。

2018年，预训练语言模型例如ElMo[15]、GPT[16]以及Bert[17]等的出现再一次刷新了机器阅读理解模型的多项指标，这些预训练的语言模型可以在大规模的未标注的语料中获得比较好的文本表示，然后再这些文本表示之上去做各种下游任务，包括机器阅读理解。这些预训练模型不管是 feature-based还是fine-tuning的方法都极大的提升了机器阅读理解的指标。

从2018年至今，出现了很多个高质量的需要借助外部知识的机器阅读理解的数据集，包括 ReCoRD[18]、ARC[19]、MCScripts[20]、OpenBookQA[21]以及CommonsenseQA[22]。ReCoRD 可以看作是抽取式的机器阅读理解的数据集，而其他几个数据集都是多项选择式的数据集。

Long[23]等为了让深度学习模型利用外部知识，提出了一种新的任务叫做稀有实体预测，该任务需要预测缺失的命名实体，类似与完形填空任务。但是该任务仅仅通过提供的文本内容无法预测出缺失的实体，需要借助从Freebase知识库中提取的实体描述来帮助实体预测。

Yang[24]等为了防止在引入外部知识的过程中，引入了与文本内容无关甚至是误导的知识，设计了注意力机制来决定外部知识是否应该被引入。

Mihaylov[25]等和Sun[26]等人利用Key-Value的Memory Networks[26]来找出相关的外部知识。所有可能用到的外部知识首先被选择放到知识库中的内存槽中作为键值对，键是用来和query做匹配，对应的值是相关知识的权重求和表示。

Wang[27]等针对英语单词知识库 WordNet 中的语义关系提出了一种数据丰富的方法，它对于文章和问题中的每一个词，试图去WordNet中找到文章中

每个单词的位置，这个位置直接或者间接地对该词有语义关系，而这个位置信息就可以作为外部知识加入到机器阅读理解模型中辅助地去预测答案。

Weissenborn[28]等提出了一种动态融合外部知识的方法，首先通过一个通用的阅读模块以文本的形式读入外部知识以及面向具体任务的文本内容，然后再根据具体任务来修正文本的内容表示，再去预测问题的答案。

Bauer[29]针对多跳的生成式的 NarativeQA 任务提出了一种新颖的方法，该方法通过逐点互信息和基于词频的打分函数，从常识知识图谱 ConceptNet 来选择多跳的相关的常识知识，来辅机器阅读理解生成相应的答案。

Mihaylov[30]等针对完形填空式的阅读理解提出了一种将外部知识编码为 key-value 内存表示的 Knowledgeable Reader 网络，然后在推断出最后的答案之前将这些知识和文章的上下文表示进行融合，提高了机器阅读理解的效果。

Pan[31]等提出在科学问答数据集如 ARC 和 OpenBookQA 中利用信息检索技术引入无结构化的外部知识，并将该知识和 subject 相关的描述知识进行融合，辅助最后的答案选择。

Chen[32]等提出 KIM 模型，该方法通过 Co-Attention 机制、收集局部推理信息然后作出推断，在自然语言推理任务 NLI 和 Multi-NLI 等任务中取得了不错的效果，该方法在其他领域仍有很多借鉴意义。

Wang[33]等利用三路 Attention 机制，得到问题和文档的表示之后，使用“硬编码”的方式融合了来自常识知识图谱 ConceptNet 中的知识，最后辅助机器阅读理解选择出合适的答案。

Yang[34]等提出了一种 KT-NET 模块来融合外部知识，这些外部知识首先通过从 WordNet、Nell 等知识图谱中检索得到，然后在 KT-NET 中通过 BiLinear 函数和 Attention 机制融合外部知识，最后辅助机器阅读理解预测答案的起始位置。

1.2.2 国内研究现状

中文领域的机器阅读理解发展相对较晚，这其中很重要的原因是中文阅读理解语料库的缺乏。最早的工作是郝晓燕[35]等人，人工构建了首个中文阅读理解的数据集，该数据集包含 121 篇文章，涉及到 14 个领域，而且还详细介绍了阅读理解数据集的构造方法，这为中文阅读理解的研究奠定了很好的基础。

进一步，受限于语料库的规模，早期的机器阅读理解方法更多的是基于特征工程和传统的机器学习的方法。李济洪[36]借助最大熵模型，根据阅读理解中问题和候选答案的关系，构造了35个词法层面的和句法层面的人工特征，在山西大学所提出的中文阅读理解数据集上达到了80.18%的准确率。

然而这种方法的特点是在特定任务特定场景下准确率很高，但是在复杂多变的应用场景下适用性不强。

另一个极大的促进中文领域的机器阅读理解的发展的是哈工大讯飞联合实验室的相关工作，该组织简称为HFL-RC。2019年HFL-RC[37]首次发布大规模的中文机器阅读理解的数据集，该数据集使用了机器自动挖词的方式来构造的，使用的语料库为人名日报和儿童读物。基于该构造的中文阅读理解数据集，作者还设计了一种基于注意力机制的深度阅读理解模型，在该数据集上取得了还不错的效果。

同一年HFL-RC[38]还发布了中文司法领域的阅读理解数据集，该数据集聚焦于司法领域，要求基于中文裁判文书来回答相关问题。这些数据集的发布极大地促进了中文机器阅读理解的发展。

1.3 主要研究内容

本文的主要研究思路是设计某种融合外部知识的方法，以此来构建融合外部知识的机器阅读理解模型，根据阅读理解提供的问题和文档内容，将检索的外部知识融合到机器阅读理解模型获取答案的过程中，从而辅助机器阅读理解更好地得到问题的答案。

本文的研究内容主要分为以下三个部分：

（1）基于预训练模型的隐式的引入外部知识的方法。近年来，随着预训练模型的快速发展，很多的机器阅读理解任务指标一次又一次的被刷新，这其中很大程度上得益于预训练模型能够很好的利用大量的无标注文本，从而获取文本更深层的语义信息。而且很多外部知识其实都蕴含在这些大量的无标注的文本当中，然而普通的机器阅读理解模型难以获得这些知识，针对这个问题，基于预训练模型构建机器阅读理解模型来隐式地引入外部知识，辅助机器阅读理解寻找问题的答案。这一部分的研究结果将为后面的研究内容打下基础。

（2）基于Attention机制的显式的引入外部知识的方法。知识不仅仅存在于大量的无标注的文本当中，还存在于结构化的知识库中，如NELL、WordNet等知识库。针对这个问题，首先检索出与文档和问题相关的外部知

识，再利用 Attention 机制设计带知识融合模块的机器阅读理解模型，将知识融合到机器阅读理解模型中，以进一步增强机器阅读理解获取问题答案的性能。

（3）结合实体感知增强的显示的引入外部知识的方法。当前的以预训练模型为基础的机器阅读理解模型将文本进行分词的时候会将实体词拆开，然而很多存在于外部知识库中的知识都是实体级别的，这将导致引入外部知识的过程中会添加很多噪声，这将很不利于外部知识与文本信息的融合，从而影响机器阅读理解的性能。针对这个问题，提出了一个结合实体感知增强和外部知识的机器阅读理解模型 FSNER-net，该模型添加了一个命名实体识别的辅助任务，该任务和机器阅读理解中的阅读理解任务一起进行联合训练，以增强机器阅读理解模型的实体感知能力，从而增强模型引入外部知识的能力，进一步提高机器阅读理解获取答案的能力。

1.4 本文章节安排

本文主要从基于预训练模型的隐式的引入外部知识的方法、基于 Attention 机制的显式的引入外部知识的方法、结合实体感知增强的显示的引入外部知识的方法等几个方面开展讨论和研究。

本文的主要内容及结构安排如下：

第一章为本文的绪论。首先阐述课题的研究背景和意义，对课题的实际应用需求以及技术背景进行介绍，从而分析课题任务的应用价值以及研究的可行性。然后对融合外部知识的机器阅读理解的国内外相关研究进行分析和总结，针对现有融合外部知识的方法的问题进行探讨，从而确定本文的主要研究内容。

第二章主要介绍基于预训练模型的隐式的引入外部知识的方法。首先介绍本文的数据来源，并对其进行预处理，得到本课题的数据集。然后对现有的流行的预训练模型进行简要介绍，并分析各个预训练模型的特点。之后，利用这些预训练模型设计机器阅读理解模型，进行实验，根据实验结果进行分析调整，确定后续任务使用的预训练的模型。

第三章主要介绍基于 Attention 机制的显式的引入外部知识的方法。首先介绍机器阅读理解中常用的 Attention 机制，然后介绍几个常见的外部知识库以及其特点。根据阅读理解数据集的特点选择对应的合适的知识库，并设计方法检索出相关的外部知识，最后介绍基于 Attention 机制的融合外部知识的机器阅读理解模型，进行实验，并给出分析结果。

第四章主要介绍结合实体感知增强的引入外部知识的方法。首先介绍引入外部知识的机器阅读理解主要技术以及常见的命名实体识别方法，然后介绍结合实体感知增强和外部知识的阅读理解方法，该部分包含命名实体识别的辅助任务数据集构造、外部知识检索、结合实体感知增强和外部之知识的阅读理解模型介绍，最后进行实验并给出分析结果。

第 2 章 基于预训练语言模型的隐式引入外部知识的方法

2.1 引言

近几年，随着预训练语言模型的快速发展，像 BERT[17]、XLNET[39]等，很多的自然语言处理任务的 SOTA 指标一次再一次地被刷新，尤其是机器阅读理解任务的效果提升幅度显著。这其中很大的原因是预训练语言模型借助优秀的深度学习架构，如 Transformer[40]等，能够很好的利用大量的无标注的文本，从而能够获取文本更深层的语义信息。而且很多的外部知识其实都已经蕴含在这些大量的无标注的文本当中，也就是说预训练语言模型其实已经从这些大量的无标注的文本当中学习到了一些外部知识信息，这些外部知识信息将会很有利于需要外部知识的机器阅读理解任务。为此，本章直接利用预训练模型来构建机器阅读理解模型，隐式地引入外部知识，从而提升机器阅读理解的任务效果。

本章的主要安排如下：第 2.2 节主要介绍实验使用的数据集，对数据进行的预处理操作；第 2.3 节简要介绍了预训练语言模型的发展，并对本章使用的预训练语言模型方法原理加以阐述；第 2.4 节介绍基于预训练模型构建的机器阅读理解模型方法;第 2.5 节给出了本章的实验结果并对其进行分析；第 2.6 节则是对本章内容进行一个总结。

2.2 实验数据集介绍

本章使用的语料为 ReCoRD 英文的阅读理解数据集，该数据集主要来自于 CNN、Daily Mail 的一些英文文章，该数据集任务要求给定一个文章和问题，要求机器能够找到问题的答案，仅仅根据文章内容可能无法准确获得问题的答案，需要借助一定的外部知识。

ReCoRD 数据集任务需要借助外部知识推理，该外部知识推理大致有 5 种类型：意译（3%）、部分线索（10%）、多个句子的推理（6%）、常识推理（75%）、歧义（6%）。这其中常识推理再细分包括：概念知识（49.3%）、因果推理（32.0%）、通俗心理学（28.0%）、社会规范和空间推理等（12.0%）。根据此数据集的特点，选取了两种外部知识库 WordNet、Nell，其中 WordNet 存储的是同义

词之间的词法关系，例如（organism，hypernymof，animal）；Nell 存储的是实体的相关信息，例如（CoCa Cola，isa，company）。该数据集包含训练集（100K）、验证集（10K）、测试集（10K）。一个简单的 example 如图 2-1 所示：

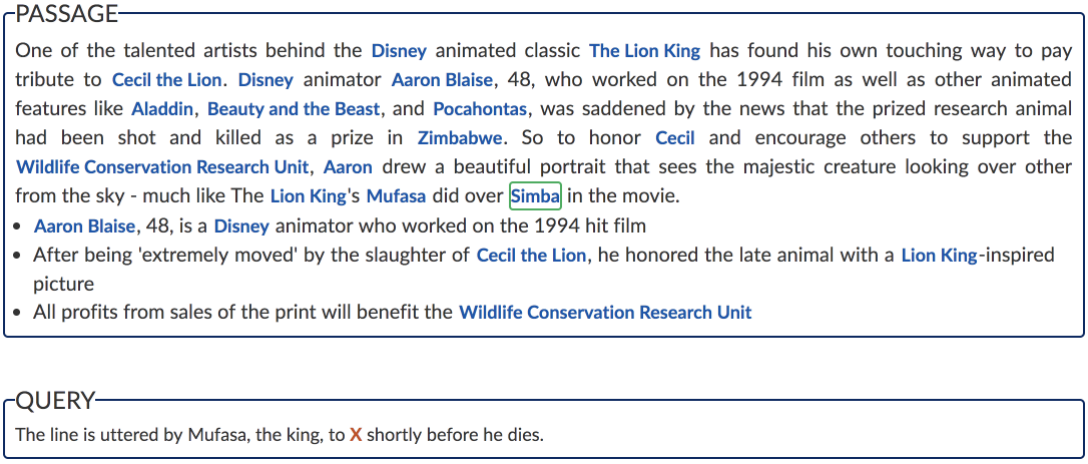

图 2-1 ReCoRD 例子

2.3 预训练语言模型介绍

近几年，在自然语言处理领域中，使用语言模型预训练方法在多项 NLP 任务上都取得了不错的提升，受到了学术界和工业界的广泛关注。这一小节将先介绍传统的语言模型方法，然后再介绍当前流行的语言模型预训练方法，特别是重点介绍 BERT 预训练语言模型。

2.3.1 传统的语言模型训练方法

语言模型是自然语言处理中一个很重要的概念，可以说自然语言处理的发展很大程度上是由于语言模型的发展所决定的。那么什么是语言模型呢？简单说语言模型是指一串词序列的概率分布。严格的定义是，语言模型表示一个长度为n的文本的为一段自然文本的概率，我们一般用P表示，用公式表示如下：

$$\mathrm{P}(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)$$
$$\ldots P(w_i|w_1, w_2, \ldots, w_{i-1}) \ldots P(w_n|w_1, w_2, \ldots, w_{n-1}) \quad (2-1)$$

然而在实际的文本当中，如果文本比较长，$P(w_i|w_1, w_2, \ldots, w_{i-1})$的估计将会很困难。针对这个问题，研究者们提出了一个简化的语言模型，称做 N-gram

语言模型，即在估计条件概率的大小时，不需要对当前词的前面所有的词进行计算，只需要对当前词的前面 N 个词进行计算，公式表示如下：

$$P(w_i|w_1,w_2,\dots,w_{i-1}) \approx P(w_i|w_{i-(n-1)},\dots,w_{i-1}) \quad (2-2)$$

$$P(w_i|w_{i-(n-1)},\dots,w_{i-1}) = \frac{count(w_{i-(n-1)},\dots,w_{i-1},w_i)}{count(w_{i-(n-1)},\dots,w_{i-1})} \quad (2-3)$$

在 N-gram 语言模型当中，传统的方法估算 N 元条件概率一般采用频率计数的比例来估算。然而这个方法也有很大的缺陷，当 N 较大的时候，会有数据稀疏的问题，这会导致估算结果不准确，因此在实际中，一般常采用二元语言模型和三元语言模型，然而该方法所起到的作用仍然有限。

2003 年 Bengio 等人为了缓解 N-gram 语言模型估计概率时所碰到的数据稀疏的问题，提出了经典的神经网络语言模型[41]，该语言模型使用了前馈的全连接的神经网络建模，首层参数可以作为词向量表示。

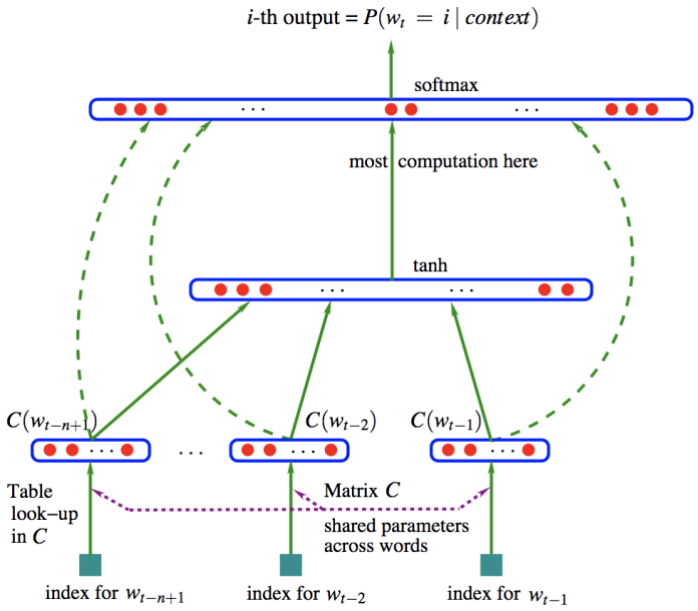

图 2-2 NNLM 模型结构图

词向量可看作是 NNLM 的一个副产品，2013 年 Mikolov 等人提出 word2vec[42]，该方法使用一些优化技巧专注于词向量的产生。在这基础之上，研究者又提出通过共现语料矩阵进行高效分解产生的 glove 词向量。

然而 word2vec、glove[43]等传统的语言模型训练方法，产生的是静态词向量，未考虑一词多义、无法理解复杂语境，这将导致在很多 NLP 任务上不能达到一个很好的效果。针对这个问题，研究者提出了新的预训练语言模型的方法，能够产生上下文相关的特征表示，即动态词向量，下一节将介绍新的预训练语言模型的方法。

2.3.2 预训练语言模型

第一个最具有代表性的预训练语言模型应该是华盛顿大学的研究者的工作。2018 年，他们提出了有名的 ELMo[15]，该模型主要是为了解决词向量的一词多义的问题，而且词向量编码的信息应该包含句法信息和语义信息。因此，ELMo 借助语言模型来获得一个上下文相关的预训练的表示。而且 ELMo 的基本结构使用的是一个双向的 LSTM 语言模型，该模型由一个前向和一个后向语言模型构成，目标函数是这两个方向语言模型的最大似然估计。具体的，给定一个包含 N 个词的句子，前向语言模型计算的概率为：

$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2, \dots, t_{k-1}) \tag{2-4}$$

后向语言模型计算的概率为:

$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^{N} p(t_k | t_{k+1}, t_{k+2}, \dots, t_N) \tag{2-5}$$

取前向和后向语言模型两个方向的最大似然：

$$\sum_{k=1}^{N} (logp(t_k | t_1, t_2, \dots, t_{k-1}; \theta_x, \theta^{\rightarrow}_{LSTM}, \theta_s)$$

$$+ logp(t_k | t_{k+1}, t_{k+2}, \dots, t_N; \theta_x, \theta^{\leftarrow}_{LSTM}, \theta_s) \tag{2-6}$$

在训练好这个语言模型之后，对于每个词，ELMo 计算双向语言模型的每一个中间层的和作为该词的表示，也可以直接使用最上层的表示。另外，如果是有监督 NLP 任务，可以直接以特征添加的方式来使用。

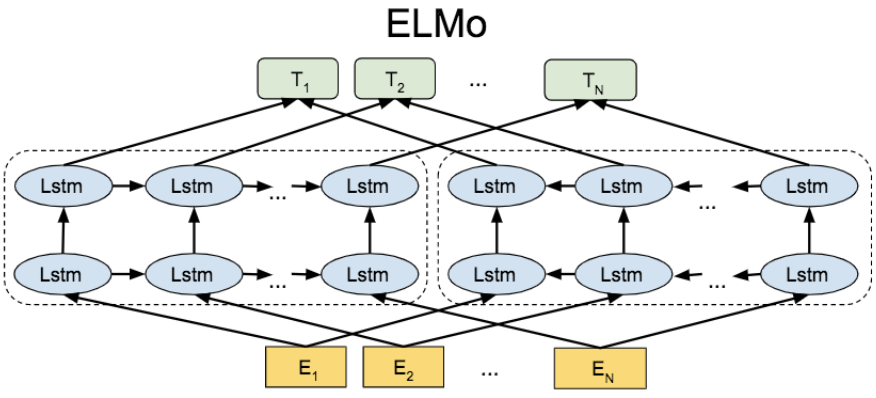

图 2-3 ELMo 模型结构图

可以看到，不像传统的词向量，一个词只对应一个词向量，ELMo 利用双向的语言模型架构，然后根据具体输入从 ELMo 中得到上下文依赖的当前词表示，然后针对具体的任务当成特征加入进去。

OpenAI 团队提出预训练语言模型 GPT[16]，该模型利用了优秀的 Transformer 网络，取代了传统的 LSTM 作为语言模型，以便于更好的捕获长距离语言结构，针对具体的下游任务，进行微调，将语言模型的训练目标作为附属任务。

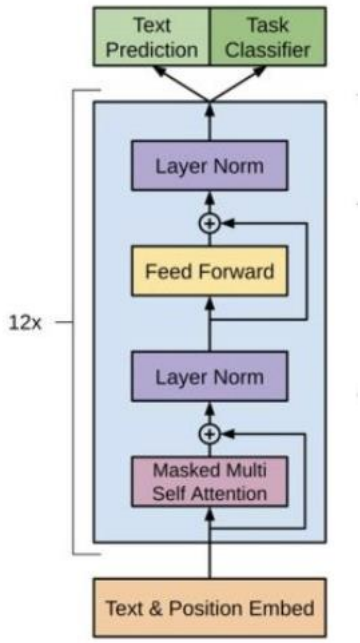

图 2-4 GPT 预训练语言模型图

语言模型中最具代表性的工作是 Google 等团队的工作，他们提出 Bert[17]，该模型将预训练语言模型分为了两大类，一个是基于特征，一个是基于微调，前者代表性的工作是 ELMo，后者是 GPT，但是这两者预训练的语言模型都是单向的。Bert 的基本思想和 GPT 很类似，最大的不同就是语言模型是双向的。而且在语言模型预训练的时候，添加了两个辅助任务，一个是 masked 语言模型任务，一个是 NSP（预测下一句）的任务。

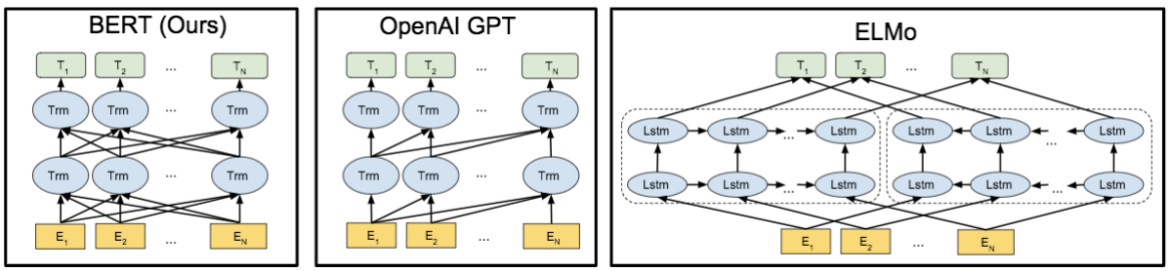

图 2-5 预训练语言模型结构图

而且，Bert 和其它三种语言模型基本组成结构不同，Bert 使用的是 transformer 编码器，该编码器是一种全连接的自注意力机制，而且在结构上 Bert 是真正的双向的语言模型。

另外，BERT 在模型的输入方面也很有讲究，Bert 做了很多的细节，如下图所示。第一是使用三个向量，一个是 WordPiece 的词向量，另外两个是位置向量和句子切分向量，以编码相关信息。此外，作者还加入了特殊分隔符，每一个文本输入都加入[CLS]、[SEP]向量，作为特殊的分隔符向量。

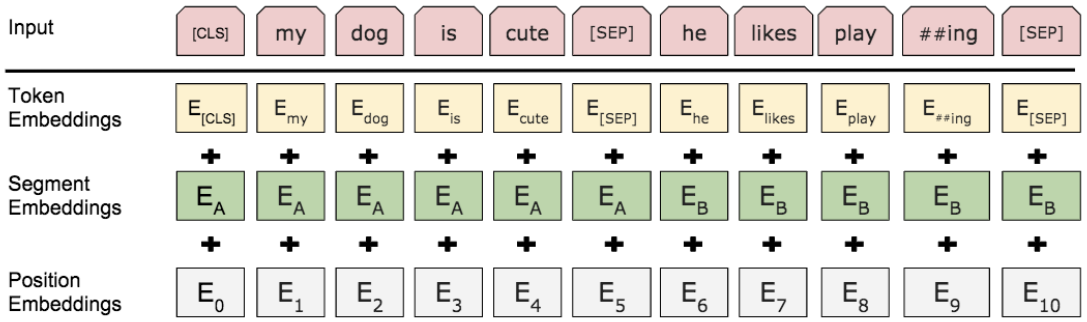

图 2-6 Bert 词向量构成

在语言模型预训练上，Bert 并没有使用标准的从左到右预测句子的下一个可能的词作为目标任务，而是使用了两个新的替代任务。第一个任务称为 masked 语言模型，即在输入的词序列中，随机地 mask 掉 15%的词，然后模型训练的任务就是去预测 masked 的这些词，可以发现，相比传统的语言模型预测目标函数，masked 语言模型即可以从左往右去预测这些挡住的词，又可以从右往左去预测挡住的这些词，不是单向的而是双向的。然而这样会有两个很明显的缺陷：

（1） 预训练过程使用[mask]挡住的词，在微调阶段是没有使用[mask]这个词的，所以会出现训练阶段和微调阶段不一致的问题；

（2） 预测 15%的词，而不是去预测整个句子，这将导致预训练的收敛过程更慢。但是虽然预训练过程变慢了，但是效果的提升可以弥补一些缺陷。

对于第一个缺，Bert 采用下面的技术来缓解：不总是使用[mask]去替换挡住的词，设定一个概率，80%的情况使用[maks]去替换，10%的情况使用一个随机词替换，10%的情况就使用这个词本身。

另外，传统的语言模型并没有考虑句子之间的关系。因此，Bert 为了让预训练语言模型学习到句子之间的关系。Bert 还引入了预测下一句的第二个辅

助任务，该任务本质上是一个二分类任务。

最后详细看一下 Bert 预训练模型所使用的语料。Bert 使用了 BooksCorpus 以及英语的 Wikipedia，BooksCorpus 总共包含 800M 单词，英语的 Wikipedia 包含 2500M 单词。可以看到，Bert 借助优秀的深度学习架构，以及很好的利用咯额这么多大量的无标注的文本，其实已经学习到了大量的外部知识，如果将预训练模型直接用于下游任务，一定会有很大的提升效果。

2.4 基于预训练语言模型的机器阅读理解模型

由于 Bert 预训练语言模型的结构特点，以及 ReCoRD 的数据集特点，将 ReCoRD 建模成抽取答案片段式地机器阅读理解任务，根据 Bert 改造成机器阅读理解模型如下图：

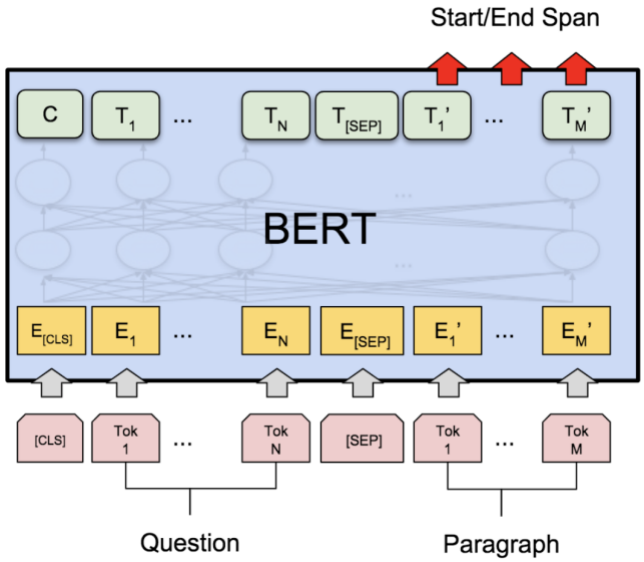

图 2-6 Bert for QA 模型结构图

Bert for QA 模型结构图直接将输入数据的每个 sample 的 question 和 paragrah 进行拼接，得到[CLS]question[SEP]paragraph 的序列输入到 Bert 中，在最上层直接接一个全连接层，以计算答案片段的起始位置和终止位置的概率。

2.5 实验结果与分析

2.5.1 基于预训练语言模型的机器阅读理解模型实验结果

为了更好的做比较，这里选取了机器阅读理解中的典型模型 QANet[14]和 DocQA[44]作为 baseline，并选取 EM 值和 F1 值作为评价指标。对于 BERT，经

过多次实验，学习率选择 3e-5，batch_size 选择 24，直接 fine-tuning 两轮，最终得到的结果如下表所示。

表 2-1 模型实验结果

模型	EM_{dev}	$F1_{dev}$
QANet	35.38	36.75
DocQA w/o ELMO	36.59	37.89
DocQA w/ ELMO	44.13	45.39
BERT(base)	**54.03**	**55.99**
BERT(large)	**64.28**	**66.60**

从上表的实验结果可以看出，相比于传统的机器阅读理解方法，BERT 提升的效果非常显著，大约提升 10 个点；而且就 BERT 而言，BERT large 的结果相比于 BERT base 的结果也提升显著，大约 10 个点。这也说明了预训练语言模型很好的利用了大量的无标注文本，这些无标注的文本可能就包含很多知识，这些将使得预训练语言模型能够获得文本更深层的语义，因而能在需要外部知识的机器阅读理解任务上取得不错的效果。

2.6 本章小结

近几年，随着预训练语言模型的快速发展，像 BERT、XLNET 等，很多的自然语言处理任务的 SOTA 指标一次再一次地被刷新，尤其是机器阅读理解任务的效果提升幅度显著。这其中很大的原因是预训练语言模型借助优秀的深度学习架构，如 Transformer 等，能够很好的利用大量的无标注的文本，从而能够获取文本更深层的语义信息。而且很多的外部知识其实都已经蕴含在这些大量的无标注的文本当中，也就是说预训练语言模型其实已经从这些大量的无标注的文本当中学习到了一些外部知识信息，这些外部知识信息将会很有利于需要外部知识的机器阅读理解任务。为此，本章直接利用预训练模型来构建机器阅读理解模型，隐式地引入外部知识，从而提升机器阅读理解的任务效果。最后的实验结果表明了假设的正确性，而且也说明预训练模型所使用的数据量越大，学到的知识也会越多，然而缺点就是代价太高。

第 3 章 基于注意力机制的融合外部知识的方法

3.1 引言

机器阅读理解，顾名思义，需要机器来理解文本并回答问题的答案，是自然语言处理中的关键任务。近几年，随着深度学习的迅猛发展，机器阅读理解取得了显著的进步。

最近，语言模型的预训练方法在机器阅读理解的研究领域引起了轰动。这些语言模型首先在无标注的文本上进行预训练，然后再应用到机器阅读理解当中，或者是以一个基于特征的方式，或者是一个基于 fine-tuning 的方式，这些方法相比于传统的机器阅读理解方法都取得了显著的性能提升。在这些不同的预训练机制里面，Bert 利用优秀的深度学习架构 Transformer，训练了一个双向的语言模型，是到目前为止最经典最成功的预训练语言模型之一，让很多的机器阅读理解任务都达到了一个新的 SOTA，而且在自然语言处理的其他任务中当中也具有广泛的应用。由于大量的未标记数据和预训练期间使用的足够深的深度学习架构，诸如 Bert 之类的高级的预训练语言模型能够捕获复杂的语言现象，比传统的语言模型更好地理解语言的深层语义信息。

但是，众所周知，真正的阅读理解不仅需要语言理解，而且还需要支持复杂的知识推理。因此，尽管预训练语言模型非常强大，但是如果能够采取合适的方式引入外部知识，将一定能够进一步增强机器阅读理解的性能。

传统的机器阅读理解模型具有相似的基本架构，都包含编码层、交互层、输出层等，其中编码层负责对输入的文本进行编码，交互层负责计算问题和文本的交互，输出层负责预测问题的答案，其中不同的机器阅读理解模型区别较大的在于交互层，交互层很多都用到了各种不同的注意力机制。这些传统的机器阅读理解模型架构在很多机器阅读理解任务上也都能达到不错的效果，但是在涉及到需要外部知识的机器阅读理解任务中，表现效果并没有那么好，和人类相比仍然有很大的提升空间，而且也没有考虑设计知识融合相关的模块。

因此，在本章中我们提出基于 Attention 机制的融合外部知识的机器阅读理解模型 FS-NET（Fusion Net），该模型是利用 Attention 机制给预训练语言模型添加知识图谱类型的外部知识。目的是为了让机器阅读理解模型不

仅能够利用深层的文本语义信息，而且还能够利用高质量的结构化的外部知识。最后在机器阅读理解数据集 ReCoRD 上证明了我们的方法的有效性。

本章的主要安排如下：第 3.2 节主要介绍本章实验使用的数据集与外部知识库；第 3.3 节主要介绍机器阅读理解中常用的 Attention 机制；第 3.4 节主要介绍基于 Attention 机制的融合外部知识的方法；第 3.5 节给出了本章的实验结果并对其进行分析；第 3.6 节则是对本章内容进行一个总结。

3.2 机器阅读理解中常用的 Attention 机制

3.2.1 Attentive Reader 和 Impatient Reader

Attentive Reader[45]的基本结构如图 3-1 所示，实际上也比较简单，属于一种细粒度的注意力机制。

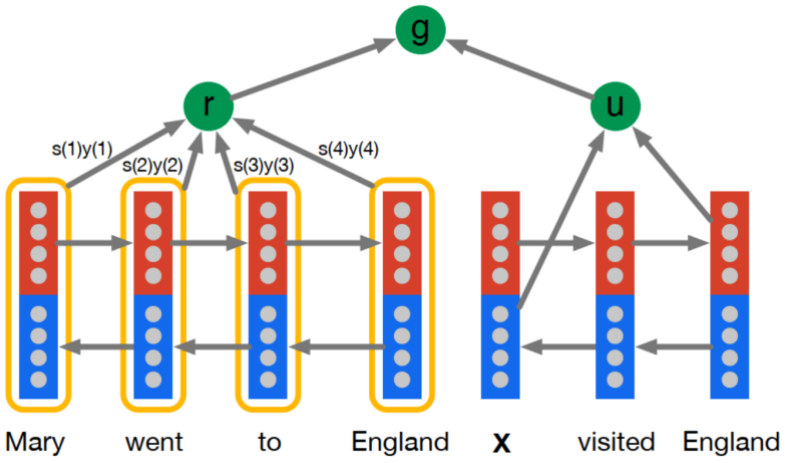

图 3-1 Attentive Reader 模型结构图

u是问题Q在经过双向 LSTMs 编码后的最后一个前向输出状态和最后一个后向输出状态的拼接，$y_d(t)$是文档D中第t个词经过双向 LSTMs 编码后的前向输出状态和后向输出状态的拼接。将文档D的单词表示$y_d(t)$作为 Key，将问题表示u作为 Query，输入一个注意力层，得到问题对文档的注意力加权表征r。

$$m(t) = \tanh\left(W_{ym}y_d(t) + W_{um}u\right) \quad (3-1)$$

$$s(t) \propto \exp\left(W_{ms}^{T}m(t)\right) \quad (3-2)$$

$$r = y_d s \quad (3-3)$$

模型最后将文档表示r和问题表示u通过一个非线性函数进行结合进行判断，计算公式如下：

$$g^{AR}(d,q) = \tanh\left(W_{rg}r + W_{ug}u\right) \quad (3-4)$$

Impatient Reader[45]也采用了注意力机制，其基本结构图如图 3-2 所示。

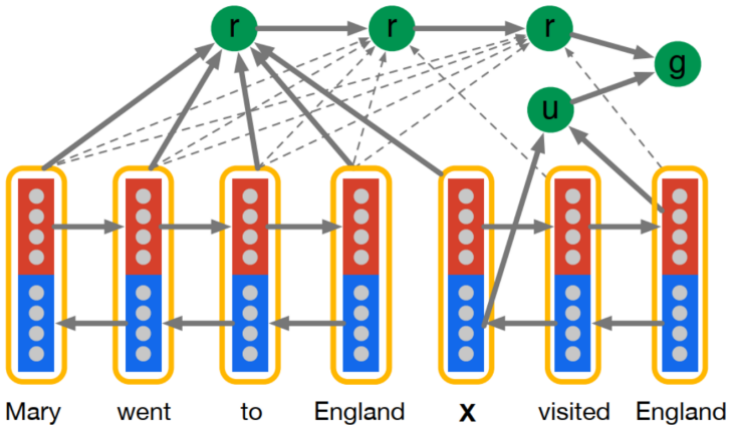

图 3-2 Impatient Reader 模型结构图

Impatient Reader 在计算注意力的时候，将每个词当作一个单独的 Query，从而计算该词对于文档每个词的注意力加权表征，并用非线性变换将所有的r进行反复累积，代表单词的重阅读能力，计算公式如下：

$$m(i,t) = \tanh\left(W_{dm}y_d(t) + W_{rm}r(i-1) + W_{qm}y_q(i)\right), 1 \le i \le |q| \quad (3-5)$$

$$s(i,t) \propto \exp\left(W_{ms}^T m(i,t)\right) \quad (3-6)$$

$$r(0) = r_0 \quad (3-7)$$

$$r(i) = y_d^T + \tanh\left(W_{rr}r(i-1)\right), 1 \le i \le |q| \quad (3-8)$$

最后将文本表示$r(|q|)$和问题表示u进行非线性组合用于答案预测。

$$g^{IR}(d,q) = \tanh\left(W_{rg}r(|q|) + W_{qg}u\right) \quad (3-9)$$

可以看到，这两种注意力的计算方式还是很有区别的。第一种，直接讲问题编码成一个固定长度的向量，在计算注意力分数的时候，等效于直接计算文档 D 每个词在特定问题上下文向量中作为答案的概率，也正是在计算问题向量Q与文档各个词的匹配关系中形成的一维线性结构，这种方式属于一维匹配模型。而第二种，直接输出问题Q每一个词的编码，计算注意力的时候，计算文档Q中每一个词对D中每一个词的注意力，即形成了一个词-词的二维匹配结构。

3.2.2 Attentive Sum Reader

Attentive Sum Reader[46]基本上和 Attentive reader 十分类似，是一种一维匹配模型，不同的是在最后的答案预测应用了一种 Pointer Sum 注意力机制，模型结构如图 3-3 所示。

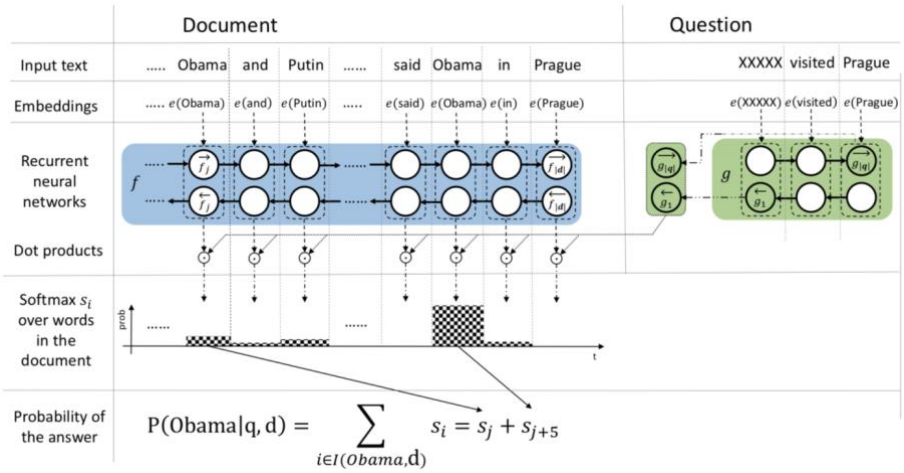

图 3-3 Attentive Sum Reader 模型结构图

与 Attentive Reader 一样，使用两个循环神经网络 GRU 对文档和问题分别进行编码，注意力层应用的是 Dot Attention，相对于 Attentive Reader 参数更少，即注意力权重计算公式为：

$$s(i,t) \propto exp f_i(d) \cdot g(q) \qquad (3-10)$$

一维匹配模型的注意力分数等效于直接文档d中每个词在特定问题上下文向量中作为答案的概率，该模型的做法是，在得到每个词 Softmax 归一化之后的分数后，将同类型的词的分数累加，得分最高的词即为答案，称为 Pointer Sum Attention，计算公式如下：

$$p(w|q,d) = \sum_{i \in I(w,d)} s_i \qquad (3-11)$$

这样将一个注意力分数累加的操作是基于一个假设，在阅读理解中出现次数越多的词越可能成为问题的答案，而且该模型的结构以及注意力的求解过程明显比 Attentive Reader 更简单，却去得了更好的效果，这也说明了并不是越复杂的模型效果更好，简单的结构在合适的场景下能取得非常好的结果。

3.2.3 斯坦福 Attentive Reader

斯坦福 Attentive Reader 同样是对 Attentive Reader 的改进，属于一种一维匹配模型，模型的基本结构图如图 3-4 所示。

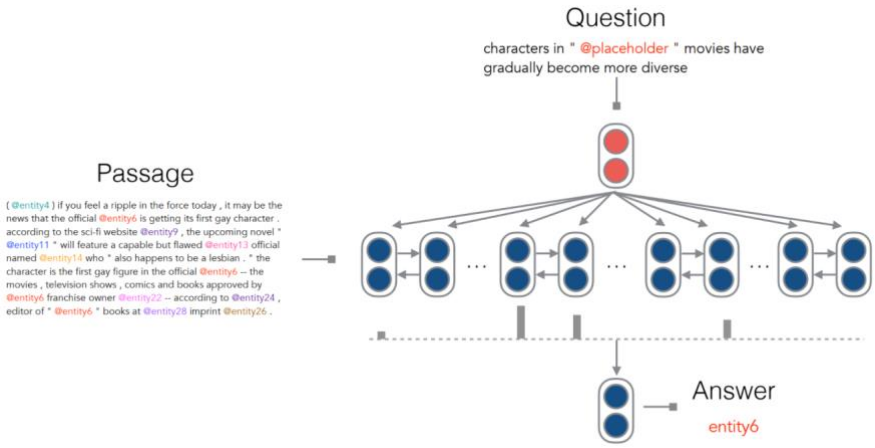

图 3-4 斯坦福 Attentive Reader 模型结构图

与 Attentive Reader 不同的是，注意力的计算方式为 bilinear，较点积的方式更灵活，计算公式如下：

$$a_i = softmax(q^T W_s p_i) \tag{3-12}$$

$$o = \sum_i a_i p_i \tag{3-13}$$

另外一点不同的是，斯坦福 Attentive Reader 得到注意力加权输出o之后，并没有与问题又做了一次非线性处理之后才预测的，而是直接用输出进行分类预测。而且，相比于 Attentive Reader 考虑所有出现在词表中的词用来做预测，二该模型只考虑出现在文本中的实体，进一步减少参数。

3.2.4 AOA Reader

AOA Reader[46]属于一种二纬匹配模型，该论文的亮点是将另一种注意力嵌套在现有注意力之上的机制，即注意力过度集中机制，模型的主要结构图如图 3-5 所示。

该模型首先对文档和问题分别使用双向循环神经网络进行编码，得到编码后的隐藏层表征h_{doc}和h_{query}。然后利用 pair-wise 匹配矩阵来计算得到注意力匹配分数，计算公式如下：

$$M(i,j) = h_{doc}(i)^T \cdot h_{query}(j) \tag{3-14}$$

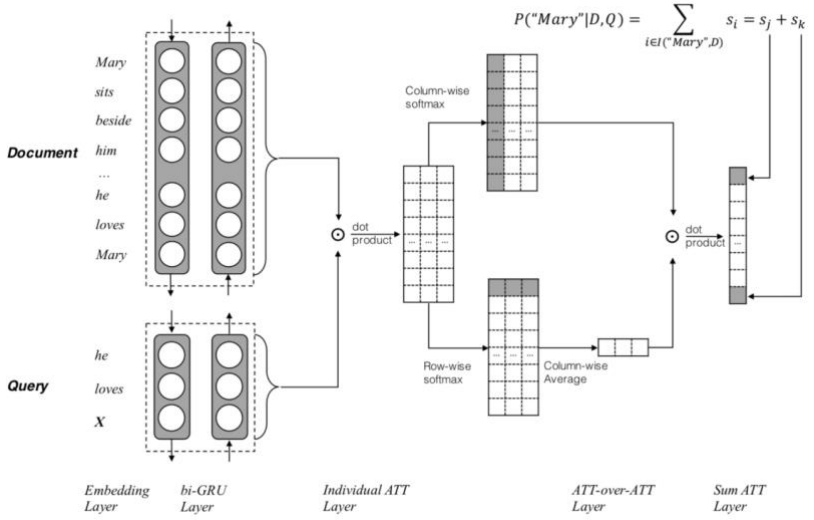

图 3-5 AOA Reader 模型结构图

在列的方向上进行 Softmax 归一化，注意上一个公式，每一列表示问题中的每一个词对文档所有词的注意力分数大小，得到所谓的问题到文档的注意力，计算公式如下：

$$\alpha(t) = softmax\big(M(1,t),\dots,M(|D|,t)\big) \quad (3-15)$$

$$\alpha = [\alpha(1),\alpha(2),\dots,\alpha(|Q|)] \quad (3-16)$$

在行的方向上进行 Softmax 归一化，得到文档到问题的注意力，计算公式如下：

$$\beta(t) = softmax\big(M(t,1),\dots,M(\,t,|Q|)\big) \quad (3-17)$$

将文档对问题的注意力和刚刚得到的问题级别的注意力做点乘，得到文档中的每个词的分数。

$$s = \alpha^T\beta \quad (3-18)$$

与 Attentive Sum Reader 类似，最后预测答案词的方式是将同类型的词的分数累加，得分最高的词即为答案，下式中，V为词表：

$$P(w|q,d) = \sum_{i \epsilon I(w,d)} s_i, w \epsilon V \quad (3-19)$$

3.2.5 双向注意力机制

双向注意力机制最先出现在机器阅读理解模型 BiDAF[48]当中，BiDAF 的模型结构图如图 3-6 所示，该模型包含字符嵌入层、词嵌入层、上下文嵌入层、双向注意力流层、建模层、输出层，其中最重要的双向注意力流层。

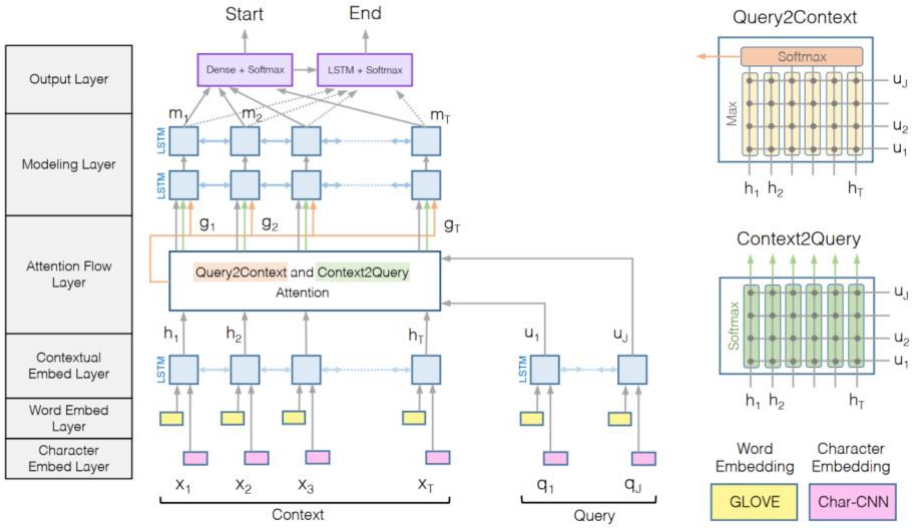

图 3-6 BiDAF 模型结构图

注意力流层是 BiDAF 模型最重要的部分，注意力流层负责链接与融合问题和上下文的信息，和以前流行的注意力机制不一样，BiDAF 不会把问题和上下文变成一个单一的特征向量，而是将每个时间步的注意力向量都与之前层的嵌入向量,一起输入建模层,这可以减少因为早起概要引起的信息损失。

在该层的计算中，计算了两个方向上的注意力，上下文到问题的注意力，问题到上下文的注意力。首先，构造一个共享相似度矩阵：

$$S_{tj} = \alpha\left(H_{:t}, U_{:j}\right) \in R^{T*J} \qquad (3-20)$$

$$a(h,u) = w_{(S)}^{T}(h;u;h \cdot u) \qquad (3-21)$$

然后使用得到的共享相似度矩阵S来计算两个方向上的注意力大小，一个是上下文到问题的注意力，计算问题上的一个词对上下文上的每个词的注意力大小，与 AOA 模型中的做法有点类似，对行方向进行归一化，再对问题进行注意力加权，包含所有问题信息：

$$a_t = softmax(S_{t:}) \in R^{J} \qquad (3-22)$$

$$U_{:t} = \sum_{j} a_{tj} U_{:j} \in R^{2dxT} \qquad (3-23)$$

一个是问题到上下文的注意力，计算上下文上的一个词对问题上的每个词的注意力，这些上下文单词对回答问题都很重要。直接取相关性矩阵每一列的最大值，再将其进行 softmax 归一化，对上下文加权，并在列方向迭代 T 次，最后得到的矩阵维度为$H \in R^{2d*T}$，包含所有的上下文信息，计算公式如下：

$$b = softmax\left(max_{col}(S)\right) \in R^{T} \qquad (3-24)$$

$$h = \sum_t b_t H_{:t} \in R^{2d} \quad (3-25)$$

3.2.6 自匹配注意力机制

自匹配注意力机制来自于 R-NET[49]机器阅读理解模型，该模型包括编码层、门控注意力循环神经网络、自匹配注意力层、边界预测层，其中最重要的是自匹配注意力层。

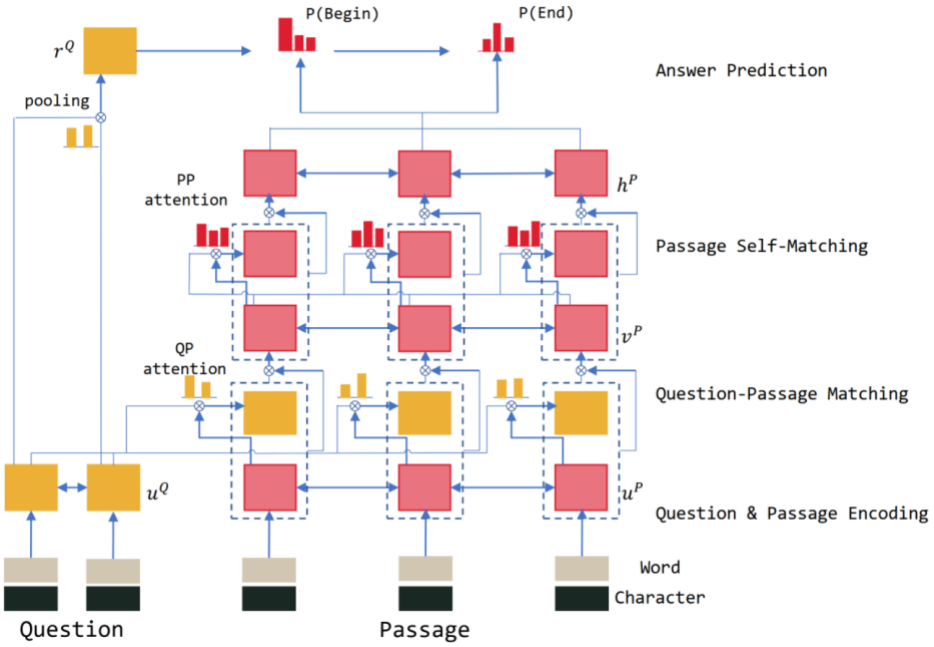

图 3-7 R-NET 模型结构图

自匹配注意力层的提出是动态地收集整个段落的信息给当前的词语，把与当前的段落词语相关的信息和其匹配的问题信息编码成段落表示：

$$h_i^p = BiRNN\left(h_{t-1}^P, \left[v_t^p, c_t\right]\right) \quad (3-26)$$

这里的$c_t = att(v^P; v_t^P)$为对整个文章的自注意力池化，另外，

$$s_j^t = v^T \tanh\left(W_v^P v_j^p + W_v^p v_t^p\right) \quad (3-27)$$

$$a_i^t = \frac{\exp(s_i^t)}{\sum_{j=1}^n \exp\left(s_j^t\right)} \quad (3-28)$$

$$c_t = \sum_{i=1}^n a_i^t v_i^p \quad (3-29)$$

3.3 实验数据集与外部知识库

本章仍然使用上一章所使用的 ReCoRD 数据集，从上一章中训练得到的基于预训练语言模型的机器阅读理解模型的实验结果可以看出，仅仅依靠预训练语言模型 Bert，仍然是缺乏知识的，因此在 ReCoRD 数据集上的效果还有提升的空间。针对数据集的特点，选择了两个常见的知识库：WordNet 和 NELL 知识库。

WordNet[1]是由 Princeton 大学的心理学家、语言学家和计算机工程师联合设计，是一种基于认知语言学的英语词典。WordNet 包含很多同义词集合，而且每个同义词集合都代表着一个基本的语义概念，并且这些概念之间也由各种关系连接。WordNet 包含描述概念含义，一义多词，一词多义，类别归属，近义，反义等问题。例如，利用 WordNet 可以得到 publish 单词的同义词，同义词有“print.v.01”、“publish.v.02”、“publish.v.03”、“published.a.01”、“promulgated.s.01”等。

NELL(Never-Ending Language Learner)[2]是卡内基梅隆大学开发的知识库。NELL 主要采用互联网挖掘的方法从 Web 自动抽取三元组知识。NELL 的基本理念是：给定一个初始的本体（少量类和关系的定义）和少量样本，让机器能够通过自学习的方式不断的从 Web 学习和抽取新的知识。目前 NELL 已经抽取将近 300 万条三元组知识。例如：一条三元组 special_events is a TV_show。

3.4 基于注意力机制的融合外部知识的机器阅读理解方法

3.4.1 外部知识检索

由于 ReCoRD 的数据特点，将其建模成抽取式地机器阅读理解任务。对于问题和文章中的每个词，去知识库中检索相关的知识，然后借助知识图谱训练知识的方法得到这些知识的一个初步表示。

具体的，对于 WordNet 知识库来说，会检索出每个词的同义词，作为该词的外部语义知识。例如对于词“husbands”会从 WordNet 知识库中检索出知识“husband.n.01”、“conserve.v.03”。

1 WordNet 官网 https://wordnet.princeton.edu/

2 NELL 官网 http://rtw.ml.cmu.edu/rtw/index.php?

对 NELL 知识库来说，会检索出每个命名实体的相关知识。细节处理步骤主要有以下几点：

（1） 将 ReCoRD 数据集里面的实体转换成字符串，去除标点符号并用_来替换空格

（2） 处理 NELL 知识库中的实体，档实体中有数字的时候，去除前后的 n 符号

（3） 若 ReCoRD 中的实体包含一个词以上，检索知识的时候使用精确匹配；若只有一个词，先对改词进行词干提取，再部分匹配

（4） 在 ReCoRD 的问题和文章中，如果一个实体 A 是实体 B 的前缀，则直接使用实体 B 代替实体 A。如使用实体“Donald Trump”代替实体“Trump”。

3.4.2 基于注意力机制的融合外部知识的机器阅读理解模型

预训练语言模型能够学习到文本的更深层的语义表示，基于 Attention 机制的融合外部知识的方法在这基础之上，将需要的外部知识表示成语义向量，利用 Attention 机制，显式的引入外部知识，将该模型称为 FS-NET，模型的基本结构图如下。

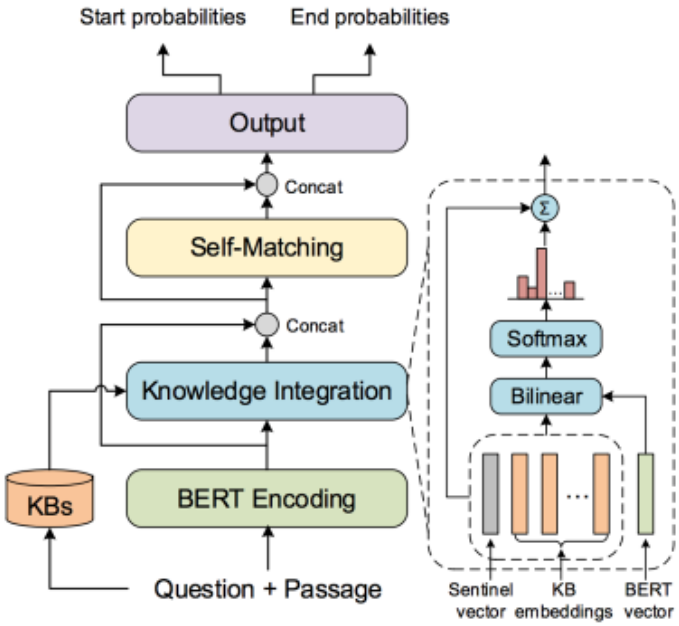

图 3-8 FS-NET 模型图

FS-NET 包含四层，分别是：

（1） BERT Encoding layer，计算 question 和 passage 的深度文本表示

（2） Knowledge Integration layer，从知识库 KB 中选择需要的知识向量

（3） Self-Matching layer，融合 BERT 和 KB 的表示

（4） Output layer，预测答案的起始位置和终止位置

接下来将详细说明 FS-NET 每一层是怎么做的。

第一层，BERT Encoding layer。假设 passage 表示为$\mathrm{P}=\{p_i\}_{i=1}^{m}$，问题表示为$\mathrm{Q}=\{q_j\}_{j=1}^{n}$，为了方便输入 BERT，将它们拼接成如下形式：

$$\mathrm{S}=[<CLS>,Q,<SEP>,P,<SEP>] \tag{3-30}$$

这种输入表示经过 BERT，得到最后的隐层表示$\{h_i^L\}_{i=1}^{m+n+3}$，L 表示 BERT 层数。

第二层，Knowledge Integration layer。过 BERT 之后的表示的每个词可以看作为一个 token，对于每个 token s_i，它的 BERT 表示$\mathrm{h_i^L}\in\mathrm{R^{d_1}}$。在训练之前，会给每个 token s_i 检索到一组相关的 KB concepts 集合称为$C(s_i)$，而且集合中的每个 concept c_j 都有一个 KB embedding $c_j\in R^{d_2}$ 。

为了更好的融合这些外部知识向量，这里采用了 Attention 机制。对于每个 token s_i ，为了衡量c_j与s_i的相关程度，这里采用 bilinear 操作来计算这个 Attention 权重，计算公式如下：

$$\alpha_{ij}\propto\exp\left(\mathrm{c_j^T W h_i^L}\right) \tag{3-31}$$

其中$W\in R^{d_2 x d_1}$。另外，由于不是每个 token s_i 都会有知识需要引入，因此，这里引入了一个知识哨兵(knowledge sentinel) $\bar{c}\in R^{d_2}$，并用同样的方式计算 Attention 权重为：

$$\beta_i\propto\exp\left(\bar{\mathrm{c}}^{\mathrm{T}}\mathrm{W h_i^L}\right) \tag{3-32}$$

对于 token s_i ，对齐检索的知识集合得到融合外部知识的知识状态向量k_i，

$$\mathrm{k_i}=\sum_{\mathrm{j}}\alpha_{ij}\mathrm{c_j}+\beta_i\bar{\mathrm{c}} \tag{3-33}$$

其中$\sum_j\alpha_{ij}+\beta_i=1$。

最后拼接 BERT 表示h_i^L和知识状态向量k_i得到$u_i=[h_i^L,k_i]\in R^{d_1+d_2}$， u_i就具有了上下文信息和相关外部知识信息。

第三层，Self-Matching layer。这一层主要是采用 Self-Attention 机制来进一步计算上下文的信息表示和相关外部知识信息的交互信息表示，而且这里同时使用了直接的交互信息表示和间接的交互信息表示。

直接的交互信息表示中，对于 token s_i 和 s_j，它们的外部知识信息表示分别为u_i 和 u_j ，这里使用 Trilinear 函数来计算相似矩阵，

$$r_{ij} = w^{\top}[u_i, u_j, u_i \odot u_j] \tag{3-34}$$

其中 $w \in R^{3d_1+3d_2}$是可训练的参数，这样得到一个矩阵R，然后对矩阵R做一个行的 softmax，得到 self-attention 的权重矩阵A，然后对于每个 token s_i 会得到一个计算注意力之后的 vector v_i ，

$$a_{ij} = \frac{\exp(r_{ij})}{\sum_j \exp(r_{ij})} \tag{3-35}$$

$$v_i = \sum_j a_{ij} u_j \tag{3-36}$$

间接的交互信息表示中，对原来的 self-attention 的权重矩阵进行一个自乘，然后得到每个 token s_i 的另外一个 attended vector $\overline{v_i}$ ，

$$\overline{A} = A^2 \tag{3-37}$$

$$\overline{v_i} = \sum_j \overline{a_{ij}} u_j \tag{3-38}$$

最后将结果拼接起来得到$o_i = [u_i, v_i, u_i - v_i, u_i \odot v_i, \overline{v_i}, u_i - \overline{v_i}] \in R^{6d_1+6d_2}$。

第四层，Output layer。这一层用来预测答案的起始位置和终止位置。将第三层的结果过一个线性层，然后再过一个标准的 softmax，计算预测答案的边界，即每个 token s_i 作为 answer span 的起始位置的概率还是终止位置的概率，即

$$p_i^1 = \frac{\exp(w_1^{\top} o_i)}{\sum_j \exp(w_1^{\top} o_j)}, \quad p_i^2 = \frac{\exp(w_2^{\top} o_i)}{\sum_j \exp(w_2^{\top} o_j)} \tag{3-39}$$

其中$w_1, w_2 \in R^{6d_1+6d_2}$是可训练的参数。

模型训练的目标函数采用最大似然，为

$$\mathcal{L} = -\frac{1}{N}\sum_{j=1}^{N}\left(\log p_{y_j^1}^1 + \log p_{y_j^2}^2\right) \tag{3-40}$$

在推断的时候，概率乘积最大的$p_a^1 p_b^2$所对应的 span（a, b）(a<b)作为最后预测的答案。

3.5 实验设置与结果分析

3.5.1 评价指标

在本章的构建模型里，由于将任务建模成抽取式的机器阅读理解 ，其评价指标通常使用 Exact Match(EM)和 F1-Score(F1)值来进行综合评价。

（1）F1 值的计算方法为：计算预测答案与原始答案字符之间的 overlap，根据 overlap 的数目与原始 ground truth answer 的字符数目计算召回率，overlap 的数目与预测出的所有字符数目计算准确率，

$$F1-Score=\frac{2*准确率*召回率}{准确率+召回率} \quad (3-41)$$

（2）Exact Match:表示预测答案与原始答案是否完全匹配，如果完全匹配则为 1，否则为 0。

3.5.2 实验设置与结果分析

这里首先在 BERT large 的基础上实现了 FS-NET，分别引入外部知识库 wordnet 和 nell，实验结果如下表所示：

表 2-2 FS-NET 实验结果表

模型	EMdev	F1dev
QANet	35.38	36.75
DocQA w/o ELMO	36.59	37.89
DocQA w/ ELMO	44.13	45.39
BERT(base)	54.03	55.99
BERT(large)	64.28	66.60
FS-NETlarge(wordnet)	65.72	67.68
FS-NETlarge(nell)	66.33	68.30

可以看到 FS-NET 的结果相比之前的传统的机器阅读理解方法都有较大幅度的提升，而且在 BERT 的基础上也有 1-2 个点的提升，由此可见引入外部知识库对于需要外部知识的机器阅读理解任务是很有效的；另外，从表中还可以看到引入 wordnet 知识库在 BERT large 上提升 1.5 个点左右，而引入 nell 知识库在 BERT large 上提升 2.1 个点左右，由此可见不同的知识库对于需要外部知识的机器阅读理解任务的作用也是不一样的，因为不同的知

识库包含的知识也不一样，wordnet 是一些包含实体同义词的一些知识信息，而 nell 包含的是实体的一些描述的知识信息。

为了更好的做比较，这里分别选取了预训练模型 BERT base 和 large 为基础实现了 FS-NET，并选取 EM 值和 F1 值作为评价指标，实验结果如下表，

表 2-3 基于 BERT base 实验结果表

模型	EMdev	F1dev
BERTbase	54.03	55.99
FS-NETbase(wordnet)	54.92	57.25
FS-NETbase(nell)	56.11	58.32

表 2-4 基于 BERT large 实验结果

模型	EMdev	F1dev
BERTlarge	64.28	66.60
FS-NETlarge(wordnet)	65.72	67.68
FS-NETlarge(nell)	66.33	68.30

从表中可以看到，在 BERT base 上引入外部知识库同样可以提升需要外部知识的机器阅读理解任务，这个和在 BERT large 的基础之上的结果是一致的；不过在 BERT base 上引入 wordnet 提升的幅度要小一些，只有 1 个点左右，而 nell 在 BERT base 和 large 上引入外部知识都提升 2 个点左右，这也可以看出 nell 知识库可能更适合作为 ReCoRD 这种需要外部知识的机器阅读理解任务的外部知识库。

3.6 本章小结

预训练语言模型为基础的机器阅读理解模型在需要外部知识的机器阅读理解任务上能够达到很不错的效果，这其中很重要的原因是预训练语言模型借助优秀的深度学习架构，能够很好的利用大量的无标注的文本，并从这些大量的无标注的文本上已经隐式地学习到了一些外部知识，因此能够在需要外部知识的机器阅读理解任务上达到不错的效果，这也从侧面说明了外部知识的引入的有效性，然而预训练语言模型为基础的机器阅读理解模型没有显示的去引入外部知识。

在本章中我们受传统机器阅读理解中的注意力机制的启发，提出基于注意力机制的显示的融合外部知识的机器阅读理解模型 FS-NET（Fusion Net），该模型是利用注意力机制给预训练语言模型添加知识图谱类型的外部知识。目的是为了让机器阅读理解模型不仅能够利用深层的文本语义信息，而且还能够利用高质量的结构化的外部知识。在需要外部知识的机器阅读理解数据集上的实验结果发现引入外部知识 wordnet、nell 能够进一步增强机器阅读理解的效果，而且引入不同的知识库提升的幅度也不相同，说明选择合适的知识库更有利于机器阅读理解效果的提升。总的来说，本章实验机过证明了基于注意力机制的引入外部知识的方法的有效性。

第 4 章 结合实体感知增强的引入外部知识的方法

4.1 引言

机器阅读理解是自然语言处理乃至 AI 界前沿的一个火热话题，它要求机器能够根据给给定的文本，来回答问题的答案，以此来衡量机器对于自然语言的理解能力。近几年，随着深度学习在自然语言处理中的迅猛发展，特别是预训练语言模型 Bert 等的出现，使得很多机器阅读理解任务的效果达到了一个新的高度，这其中很重要的原因就是预训练语言模型借助优秀的深度学习架构，能够很好地利用大量的无标注文本，从而能够获得文本更深层的语义信息。

然而在现实世界的很多实际应用当中，很多阅读理解的文本内容很复杂，涉及到很多复杂的背景知识，仅仅通过给定的文本内容并不能够很好的去理解，从而很难去准确的回答问题的答案，这就导致很多机器阅读理解模型在实际应用当中与人类仍然存在较大的差距。这其中很大的一个原因就是，和机器阅读理解不同，人类在做阅读理解的时候，不仅仅借助所提供的文本内容，还会很好的去利用一些外部经验或者外部知识来辅助理解作答，这也正是机器阅读理解和人类阅读理解的一个巨大差异。因此，给机器阅读理解引入外部知识是一个很有意义也很具有挑战性的方向。

传统的方法中很大一部分是首先检索文本中某些词对应的外部知识，然后使用某种方法如知识图谱训练词向量的方法得到外部知识的一个词向量表示,然后再和原来文本的表示向量进行拼接或相加得到融合知识的文本表示，从而去增强机器阅读理解的性能。然而这种简单的融合方法可能会引入大量的噪声，反而导致效果不理想。近几年，随着预训练语言模型如 Bert 等的出现，有很多基于预训练语言模型设计的机器阅读理解模型，在很多机器阅读理解任务上都达到了不错的效果，然而使用预训练语言模型会导致文本中的实体词可能会被拆成几个片段，而大部分的知识都是实体级别的，这会导致机器阅读理解模型在融合外部知识的过程成引入更大的噪声，模型缺乏实体感知能力。

针对这个问题，本章提出了提出了结合实体感知增强与外部知识的机器阅读理解模型 FSNER-net，该模型添加了命名实体识别的辅助任务进行联合训练，增强了机器阅读理解的实体感知能力，使得机器阅读理解更多的关注

实体相关的信息，从而更好地融合实体相关的外部知识，最后在英文机器阅读理解数据集上的实验结果证实了我们方法的有效性。

本章的主要安排如下：第 4.2 节主要介绍引入外部知识的机器阅读理解的历史研究方法；第 4.3 节介绍命名实体识别任务的常用方法；第 4.4 节介绍结合实体感知增强和外部知识的机器阅读理解方法；第 4.5 节对本章的实验结果进行分析；第 4.6 节则是对本章内容进行一个总结。

4.2 引入外部知识的阅读理解主要技术介绍

4.2.1 知识感知的双向 LSTMs

双向 LSTMs 是编码文本常用的编码器，一个负责前向编码，一个负责后向编码。为了在编码机器阅读理解文本的过程中引入外部知识，Yang 等将 LSTM 改造成 KBLSTM[24]，KBLSTM 模型结构图如下：

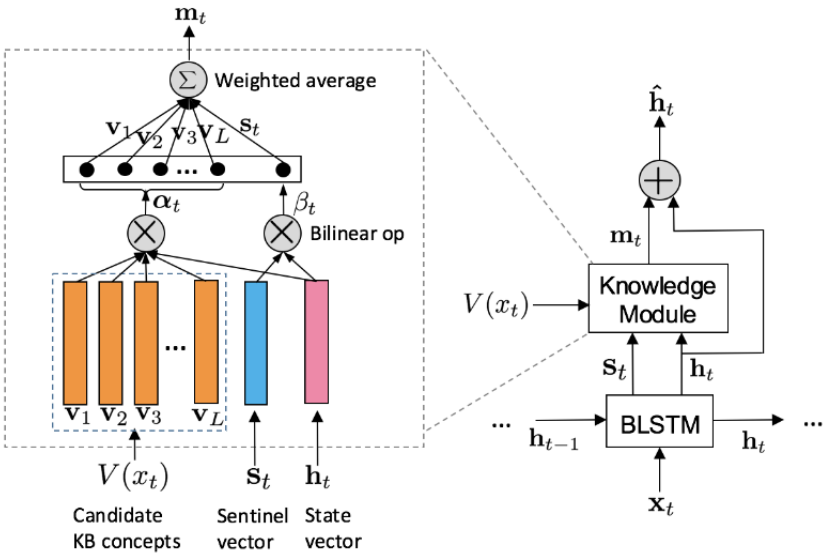

图 4-1 KBLSTM 模型结构图

KBLSTM 最核心的部分是知识模块，该模块负责将外部知识编码到双向 LSTM 中。对于输入的词x_t，t时刻的知识由很多候选的知识图谱向量$V(x_t)$组成，每一个候选的知识图谱知识都有一个向量v_i。对于每一个知识向量，我们计算一个注意力权重α_{ti}，该权重反映了知识i和当前上下文h_t的相关性，

$$\alpha_{ti} \propto \exp(v_i^T W_v h_t) \qquad (4-1)$$

其中W_v是需要学习的矩阵参数。

为了在引入外部知识和当前的文本信息取得一个折中，还引入了一个 sentinel 向量s_t，公式如下：

$$b_t = \sigma(W_b h_{t-1} + U_b x_t) \quad (4-2)$$

$$s_t = b_t \tanh \odot (c_t) \quad (4-3)$$

其中W_b和U_b是可以学习的权重参数。该权重在当前的上下文上计算公式如下：

$$\beta_t \propto \exp(s_t^T W_s h_t) \quad (4-4)$$

其中W_s是要学习的参数矩阵。该混合模型定义为如下：

$$m_t = \sum_{i \in V(x_t)} \alpha_{ti} v_i + \beta_t s_t \quad (4-5)$$

其中$\sum_{i \in V(x_t)} \alpha_{ti} + \beta_t = 1$。$m_t$能够被作为一种编码外部知识信息的知识状态向量。使用 KBLSTM 直接同时编码文本和外部知识的信息。

4.2.2 命名实体识别主要技术介绍

命名实体识别是自然语言处理中的一项非常基础的任务，在实际的工业应用中应用非常广泛。命名实体识别一般指识别文本中人名、地名、组织机构名、日期时间、专有名词等实体。

命名实体识别一直是自然语言处理中的研究热点，早期主要采用基于词典和规则的方法，后来流行使用传统机器学习的方法，到近来逐渐使用基于深度学习的技术。

深度学习中比较流行的方法是 DL-CRF 模型，在神经网络的上层再加一层 CRF[50]层，该层主要是为了在句子级别去预测标签，增加前后标签之间的依赖性。对文本进行编码一般使用长短期记忆网络，简称 LSTM[51]，是 RNN 的一种特殊类型。它的设计是用来解决长距离依赖问题。

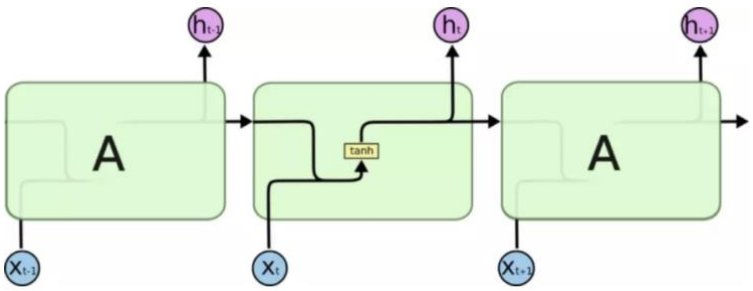

图 4-2 传统 RNN 结构图

LSTM 的结构类似于 RNN，但是不同于普通 RNN，它由四个部分组成，采用特殊的方式进行交互计算。具体的，LSTM 包含输入门、遗忘门以及输出门三个门，其中遗忘门负责选择性地遗忘部分的历史信息，输入门负责处理加入部分当前输入信息，输出门负责整合当前状态并产生输出状态。

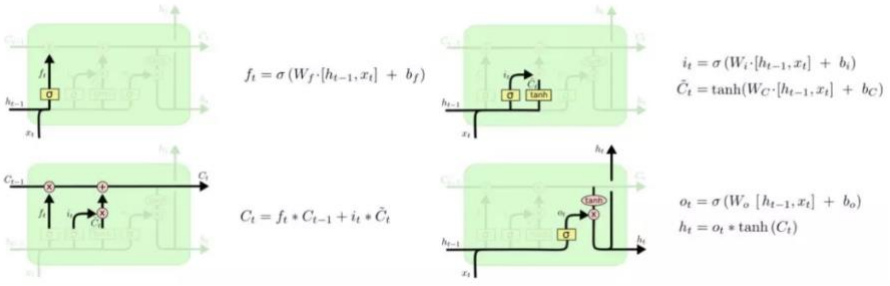

图 4-3 LSTM 各个门控结构图

命名实体识别中经典的 BiLSTM-CRF 模型主要由 Embedding 层、双向 LSTM 层，以及最后的 CRF 层构成，该模型结构图如图 4-4 所示。大量的实验结果表明 BiLSTM-CRF 深度学习模型几乎已经达到或者超过了加入丰富特征的 CRF 模型，成为目前基于深度学习的命名实体识别方法中的最主流的模型。就特征方面来说，BiLSTM-CRF 模型继承了深度学习方法的优点，省去了繁杂的特征工程，仅仅使用词向量以及字符向量就可以达到很好的效果，而且如果有高质量的词典特征，还能够进一步获得提高实体识别的效果。

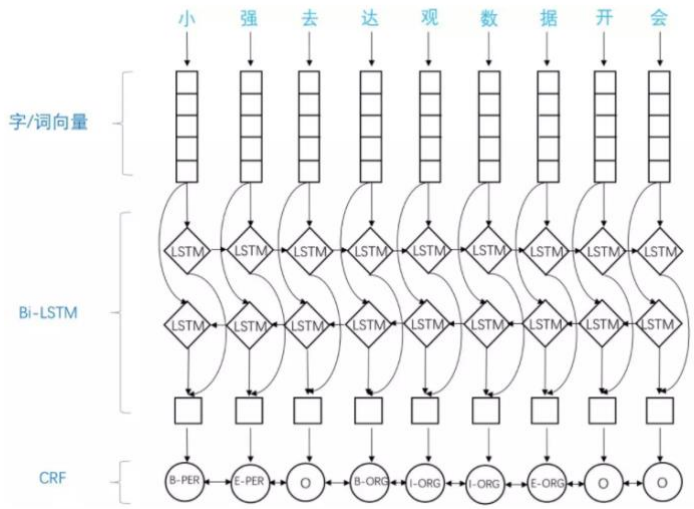

图 4-4 BiLSTM-CRF 模型结构图

近几年，随着预训练语言模型的出现，命名实体识别所使用的方法也发生了些许变化。因为，预训练语言模型借助优秀的深度学习架构，能够利用大量的无标注的文本，能够获得文本更深层的语义信息，这将很有利于命名实体识别等下游任务。基于预训练语言模型的命名实体识别模型也很简单，直接在 Bert 上面加一层线性层，然后计算预测标签的概率分布，模型结构图如图 4-5 所示。

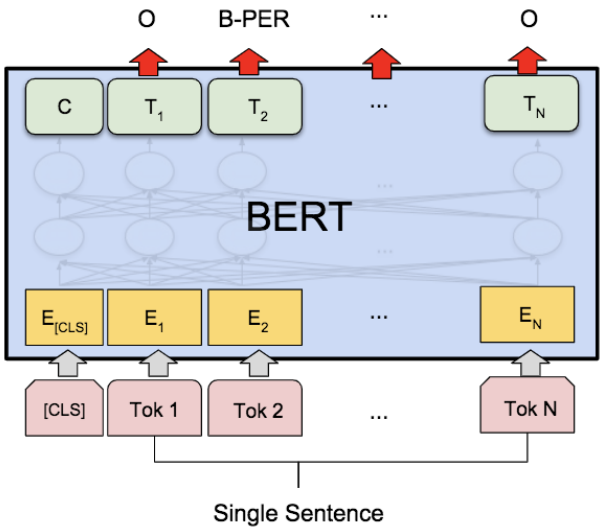

图 4-5 Bert for NER 模型结构图

4.4 结合实体感知增强和外部知识的机器阅读理解

结合实体感知增强与外部知识的机器阅读理解模型 FSNER-net 包括命名实体识别的辅助任务和引入外部知识的机器阅读理解任务两大部分，为了训练这两个任务，需要两个关键的预处理：命名实体识别辅助任务数据构造和外部知识检索，接下来将先介绍这两部分，然后再着重介绍我们提出的模型 FSNER-net 的基本结构。

4.4.1 命名实体识别辅助任务数据构造

为了让机器阅读理解模型更多的关注实体相关的信息，增强机器阅读理解模型的实体感知能力，我们在机器阅读理解任务基础上添加了命名实体识别的辅助任务。然而机器阅读理解的原始数据集是没有命名实体识别标签数据的，因此我们在原始数据集上设定以下规则来构造命名实体识别数据，规则如下：

（1）文章中的实体标注规则。根据机器阅读理解文章中的实体词信息，首先将文章中的词标注为实体，这里为了简便，只使用标签‘B’、‘I’、‘O’。如果是实体，使用标签‘B’和‘I’标注；如果不是实体，直接使用‘O’进行标注。例如‘Donald Trump’会被标注为‘B’和‘I’。并将标注的实体保存到一个列表里。

（2）问题中的实体标注规则。根据列表里的实体词，来对问题文本中的实体进行标注。如果是实体，使用标签‘B’和‘I’标注；如果不是实体，直接使用‘O’进行标注。通过这样的标注规则就得到了有命名实体识别标签的机器阅读理解数据集。

4.4.2 外部知识检索

为了在机器阅读理解中引入相关的外部知识，以增强机器阅读理解的性能，我们首先检索出机器阅读理解相关的外部知识。由于机器阅读理解数据集中很多的外部知识都是和实体相关的，所以我们采用 NELL[1]知识库作为机器阅读理解的外部知识库。

NELL(Never-Ending Language Learner)是卡内基梅隆大学开发的知识库，主要采用互联网挖掘的方法从 Web 自动抽取三元组知识，它的基本理念是：给定一个初始的本体（少量类和关系的定义）和少量样本，让机器能够通过自学习的方式不断的从 Web 学习和抽取新的知识。目前 NELL 已经抽取将近 300 万条三元组知识。例如：special_events is a TV_show。

为了更好的检索出外部知识，针对 NELL 知识库的特点以及阅读理解数据集的特点，我们制定了以下处理规则来检索外部知识：

（1）机器阅读理解数据集预处理；将机器阅读理解数据集里面的实体转换成字符串，去除标点符号并用_来替换空格；

（2）NELL 知识库预处理：处理 NELL 知识库中的实体，当实体中有数字的时候，去除前后的 n 符号；

（3）检索规则：若机器阅读理解数据集中的实体包含一个词以上，检索知识的时候使用精确匹配；若只有一个词，先对该词进行词干提取，再部分匹配；

（4）检索后处理：若机器阅读理解问题和文章中，如果一个实体 A 是实体 B 的前缀，则直接使用实体 B 代替实体 A。如使用实体“Donald Trump”代替实体“Trump”。

根据该方法，我们检索出了文本中的一些实体的相关知识，抽选出一部分的结果如表 4-1 所示。

可以看到，这些检索出来的知识是对原始文本词的语义补充，将这些知识作为机器阅读理解数据的外部知识库。

1 NELL 官网 http://rtw.ml.cmu.edu/rtw/index.php?

表 4-1 实体知识检索结果表

实体名称	检索知识列表
Anthony Watson	athlete
England	organization;island;trainstation; geopoliticalorganization
cheek	muscle; bodypart
far from the madding crowd	book
department of justice	governmentorganization
donald trump	celebrity; ceo; male

为了更好的得到知识词向量的初步表示，我们采用经典的 TransE 模型来预训练知识图谱的词向量，这里我们直接使用了清华开源的 OpenKE1工具，将训练完的词向量作为外部知识的初始化表示。

4.4.3 结合实体感知增强与外部知识的阅读理解模型

我们将提出的结合实体感知增强与外部知识的机器阅读理解模型简称为 FSNER-net，模型的结构图如图 3-1 所示，该模型主要包含 5 个部分，分别是：（1）BERT 编码层(BERT Encoding layer)，用来计算问题和文章的深度文本表示；（2）知识融合层(Fusion Knowledge layer)，用来融合文本和知识的表示；（3）自注意力层(Self-Attention layer)，用来计算文本知识表示的自注意力机制；（4）序列标注层(Sequence Tagging layer)，用来对文本进行序列标注。（5）输出层(Output layer)，用来预测答案的起始位置和终止位置；接下来，将详细说明模型 FSNER-net 每一部分具体是怎么做的。

（1）BERT 编码层

为了更好的得到问题和文章的深度文本表示，我们使用 BERT 编码层编码问题和文章。假设文章表示为$\mathrm{P}=\{p_i\}_{i=1}^{m}$，问题表示为$\mathrm{Q}=\{q_j\}_{j=1}^{n}$，为了方便输入 BERT，将它们拼接成如下形式：

$$S=[<CLS>,Q,<SEP>,P,<SEP>] \tag{4-6}$$

这种输入表示经过 BERT,得到最后的隐层表示$\{h_i^L\}_{i=1}^{m+n+3}$,其中 L 表示 BERT 层数。

（2）知识融合层

1 清华知识图谱 OpenKE 工具箱: https://github.com/thunlp/OpenKE

得到了问题和文章的深度文本表示之后，为了得到更好的融合外部知识的文本表示，将知识表示和文本表示通过知识融合层来计算融合知识的文本表示。经过 BERT 之后的文本表示的每个词可以看作为一个 token，对于每个 token s_i，它的 BERT 表示$\mathrm{h}_\mathrm{i}^\mathrm{L} \in \mathrm{R}^{\mathrm{d}_1}$。在训练之前，会按照前面提到的检索方法给每个 token s_i 检索到一组相关的 KB concepts 集合称为$C(s_i)$，而且集合中的每个 concept c_j 都有一个预训练的 KB embedding $c_j \in R^{d_2}$ 。

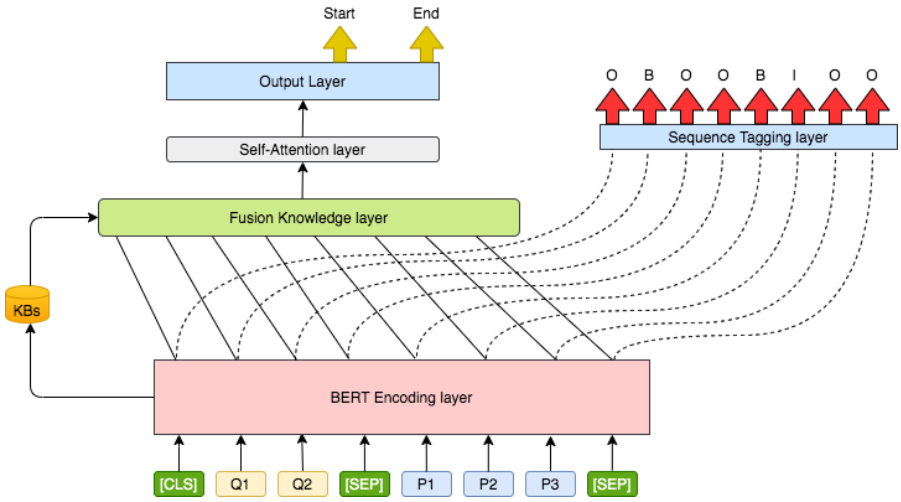

图 3-1 FSNER-net 模型结构图

为了更好的选择性地融合这些外部知识向量，这里采用了注意力机制。对于每个 token s_i ，为了衡量 c_j与s_i的相关程度，采用 bilinear 操作来计算这个注意力权重，计算公式如下：

$$\alpha_{\mathrm{ij}} \propto \exp\left(\mathrm{c}_\mathrm{j}^\mathrm{T} \mathrm{W} \mathrm{h}_\mathrm{i}^\mathrm{L}\right) \qquad (4-7)$$

其中$W \in R^{d_2 x d_1}$。另外，由于不是每个 token s_i 都会有知识需要引入，因此，这里引入了一个空知识向量 $\bar{c} \in R^{d_2}$，并用同样的方式计算注意力权重为：

$$\beta_{\mathrm{i}} \propto \exp\left(\bar{\mathrm{c}}^\mathrm{T} \mathrm{W} \mathrm{h}_\mathrm{i}^\mathrm{L}\right) \qquad (4-8)$$

对于 token s_i ，对齐检索的知识集合得到融合外部知识的知识状态向量 k_i，

$$\mathrm{k}_\mathrm{i} = \sum_\mathrm{j} \alpha_{\mathrm{ij}} \mathrm{c}_\mathrm{j} + \beta_\mathrm{i} \bar{\mathrm{c}} \qquad (4-9)$$

其中$\sum_\mathrm{j} \alpha_{\mathrm{ij}} + \beta_\mathrm{i} = 1$。

最后拼接 BERT 表示h_i^L和知识状态向量k_i得到$u_i = [h_i^L, k_i] \in R^{d_1+d_2}$， u_i就具有了上下文信息和相关外部知识信息。

（3）自注意力层

为了得到更深度的融合知识文本的交互信息表示，采用 Self-Attention 机制来进一步计算上下文的信息表示和相关外部知识信息的交互信息表示，而且这里同时使用了直接的交互信息表示和间接的交互信息表示。

在直接的交互信息表示中，对于 token s_i 和 s_j，它们的外部知识信息表示分别为u_i 和 u_j ，这里首先拼接u_i 与 u_j的差、和以及它们之间的哈达马乘积，然后再过一个线性层来计算相似矩阵，

$$r_{ij} = w^T[u_i + u_j, u_i - u_j, u_i \odot u_j] \quad (4-10)$$

其中 $w \in R^{3d_1+3d_2}$是可训练的参数，这样得到一个矩阵R，然后对矩阵R做一个行的 softmax，得到自注意力的权重矩阵A，然后对于每个 token s_i 会得到一个计算注意力之后的 vector v_i ，

$$a_{ij} = \frac{\exp(r_{ij})}{\sum_j \exp(r_{ij})} \quad (4-11)$$

$$v_i = \sum_j a_{ij} u_j \quad (4-12)$$

间接的交互信息表示中，对原来的自注意力的权重矩阵 A 进行一个自乘，然后得到每个 token s_i 的另外一个信息表示向量 $\overline{v_i}$ ，

$$\overline{\mathbf{A}} = \mathbf{A}^2 \quad (4-13)$$

$$\overline{v_i} = \sum_j \overline{a_{ij} u_j} \quad (4-14)$$

最后将结果拼接起来得到$o_i = [u_i, v_i, u_i - v_i, u_i \odot v_i, \overline{v_i}, u_i - \overline{v_i}] \in R^{6d_1+6d_2}$。

（4）序列标注层。

为了让机器阅读理解模型更多的关注实体相关的信息，增强模型的实体感知能力，这里引入了命名实体识别的辅助任务，它与机器阅读理解部分共享 BERT 编码层，两个任务联合训练。具体的，这一层用来对文本进行序列标注，得到问题和文章的深度表示之后过一个线性层，计算每个 token s_i标注为标签c_k的概率，使用交叉熵损失函数，为：

$$p_{c_k} = \mathrm{softmax}(w^T h_i^L) \quad (4-15)$$

这一部分的目标函数使用交叉熵损失函数，

$$\mathcal{L}_1 = -\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{C}(y_k log p_{c_k}) \quad (4-16)$$

（5）输出层

这一层用来计算预测答案的起始位置和终止位置的概率分布。将自注意力层的输出结果过一个线性层，然后再过一个标准的 softmax，计算预测答案的边界,即每个 token s_i 作为答案片段的起始位置的概率和终止位置的概率，即

$$\mathrm{p}_{\mathrm{i}}^{1}=\frac{\exp(\mathrm{w}_{1}^{\mathrm{T}}\mathrm{o}_{\mathrm{i}})}{\sum_{\mathrm{j}}\exp(\mathrm{w}_{1}^{\mathrm{T}}\mathrm{o}_{\mathrm{j}})},\quad \mathrm{p}_{\mathrm{i}}^{2}=\frac{\exp(\mathrm{w}_{2}^{\mathrm{T}}\mathrm{o}_{\mathrm{i}})}{\sum_{\mathrm{j}}\exp(\mathrm{w}_{2}^{\mathrm{T}}\mathrm{o}_{\mathrm{j}})} \tag{4-17}$$

其中$w_1,\mathrm{w}_2\in R^{6d_1+6d_2}$是可训练的参数。

这一部分模型训练的目标函数采用最大似然，为

$$\mathcal{L}_2=-\frac{1}{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}}\left(\log \mathrm{p}_{\mathrm{y}_{\mathrm{j}}^{1}}^{1}+\log \mathrm{p}_{\mathrm{y}_{\mathrm{j}}^{2}}^{2}\right) \tag{4-18}$$

在推断的时候，将概率乘积最大的$p_a^1p_b^2$所对应的片段（a，b）(a<b)作为最后预测的答案。

由于是联合训练命名实体识别辅助任务和融合外部知识的机器阅读理解任务，所以本模型最终的目标函数包含了两部分，一部分是机器阅读理解预测起始位置和终止位置的极大似然，另外一部分是命名实体识别部分标注的交叉熵损失，总的目标函数是：

$$\mathcal{L}=-\frac{1}{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}}\left(\log \mathrm{p}_{\mathrm{y}_{\mathrm{j}}^{1}}^{1}+\log \mathrm{p}_{\mathrm{y}_{\mathrm{j}}^{2}}^{2}\right)-\frac{1}{\mathrm{N}}\sum_{\mathrm{j}=1}^{N}\sum_{k=1}^{C}\left(y_k log p_{c_k}\right) \tag{4-19}$$

4.5 实验设置与结果分析

4.5.1 数据来源

我们在英文的机器阅读理解 ReCoRD 数据集进行实验。ReCoRD 数据集任务需要借助外部知识推理，该外部知识推理大致有 5 种类型：意译（3%）、部分线索（10%）、多个句子的推理（6%）、常识推理（75%）、歧义（6%）。这其中常识推理再细分包括：概念知识（49.3%）、因果推理（32.0%）、通俗心理学(28.0%)、社会规范和空间推理等(12.0%)。它的形式类似于 SQuAD，同样是给一篇文章和一个问题，要求从文章中找出该问题的答案，答案是文章中的一个片段。该数据集包含训练集 100K，验证集 10K，测试集合 10K。

4.5.2 评价指标

由于将 ReCoRD 数据集任务为抽取式的机器阅读理解 ，所以采用 EM 值（Exact Match)和 F1 值(F1-score)的评价指标来进行综合评价。

（1)F1 值的计算方法为：计算预测答案与原始真实答案字符之间的重叠，根据字符重叠的数目与原始真实答案的字符数目计算召回率，字符重叠的数目与预测出的所有字符数目计算准确率，F1 值计算公式如下：

$$F_1 = \frac{2 * 准确率 * 召回率}{准确率 + 召回率} \qquad (4-20)$$

（2）EM 值的计算方法：表示预测答案与原始真实答案是否完全匹配，如果完全匹配则为 1，否则为 0。

4.5.3 实验结果及分析

4.5.3.1 机器阅读理解实验结果

我们在 ReCoRD 数据集上进行实验，实验结果如表 4-1 所示。

表 4-1 FSNER-net 及其相关模型实验结果表

模型	Dev		Test	
	EM	F1	EM	F1
QANet[14]	35.38	36.75	36.51	37.79
DocQA[45] w/o ELMO	36.59	37.89	38.52	39.76
DocQA[45] w/ ELMO	44.13	45.39	45.44	46.65
BERT (base)	54.03	55.99	54.04	56.07
BERT (large)	64.28	66.60	-	-
BERT(large-wwm)	71.59	73.69	-	-
FSNER-net$_{bert-large-wwm}$ +NELL(ours)	73.17	75.05	73.57	75.63

表中 QANet[14]是机器阅读理解中比较有名的模型，它的编码器仅由卷积和自注意力机制组成，是相比于传统问答经典架构的一种新型架构，在 SQuAD 数据集上能达到很不错的效果，然而在 ReCoRD 数据集上的实验结果欠佳，这也说明了 ReCoRD 数据集的难度。DocQA 是针对阅读理解中有的文章无法回答问题而设计的模型，ELMo[15]是基于特征的预训练语言模型，同样是利用了大量的无标注的文本预训练。可以看到单凭 DocQA 模型达到的效果也是差强人意，而加上 ELMo，效果提升幅度较大，这也说明了基于特征的预训练语言模型的有效性。

我们提出的 FSNER-net 模型是基于预训练语言模型而设计的，为了选择更强的预训练语言模型，首先使用单纯的预训练语言模型进行实验。使用预训练语言模型 BERT 之后，可以看到效果相比于传统的阅读理解方法有大幅度提升，而且 BERT-large 的效果相比于 BERT-base 的效果也有大幅度地提升，这也说明了 BERT 等预训练语言模型借助优秀的深度学习架构，能够很好的利用大量的无标注的文本，而且这些大量的大量的无标注的文本中已经包含了很多的外部知识，因此预训练模型在该任务上表现不错。另外，我们还使用更换了原始 BERT 的 mask 机制的全词 mask(whole word masking)策略预训练的 BERT 语言模型，相比于原始的 BERT，提升效果也很明显，这也表明 ReCoRD 数据集中的文本需要借助很多实体相关的外部知识，而预训练语言模型全词 mask 策略更有利于这些实体信息的学习，从而能够让预训练语言模型学习到更多实体相关的知识信息。

因此，我们提出的 FSNER-net 机器阅读理解模型以 BERT-large-wwm 为基础，该模型不仅添加了命名实体识别辅助任务，而且从 NELL 知识库中有效地引入了外部知识。可以看到，FSNER-net 在开发集上的效果已经远远超过了传统的 QANet、DocQA 包括 BERT 等方法，说明了我们方法的优越性。另外相比于强大预训练语言模型 BERT-large-wwm，提升效果也很显著，在开发集上 EM 值提升 1.58，F1 值提升 1.36。这也说明即使是强大的预训练语言模型，可能能借助优秀的深度学习架构、适合的 mask 机制，已经在大量的无标注的文本中学习到了大量的外部知识，但是借助我们提出的结合实体感知增强和外部知识的机器阅读理解方法还能够进一步提升机器阅读理解的效果，说明联合命名实体识别的辅助任务增强了机器阅读理解的实体感知能力，使得机器阅读理解更多地关注实体相关的信息，这将更有助于实体相关的外部知识的引入。

4.5.3.2 模型消融实验

为了更好的分析 FSNER-net 模型各个部分的作用，我们还进行了模型消融实验，来查看模型的每个部分的贡献，实验结果表 4-2 所示：

从模型消融的实验结果可以看出 FSNER-net 模型各个部分都起到了作用。可以看到，NELL 知识库的引入极大地增强了机器阅读理解的性能，相比于引入 NELL 外部知识，F1 值大约提升 1.36 个点，这说明了 NELL 知识库中的实体信息对于机器阅读理解是有很大帮助的，也说明了我们显示地给机器阅读理解引入外部知识方法的有效性；另外，即使已经引入了外部知识，联合命名实体识别的辅助任务进一步提高了机器阅读理解的效果，这说明了联合命

名实体识别的辅助任务的确有助于增强机器阅读理解的实体感知能力，使得机器阅读理解模型能够更多的关注实体信息，从而更有利于机器阅读理解模型融合实体相关的外部知识信息。

表 4-2 FSNER-NET 模型消融实验结果表

模型	Dev	
	EM	F1
FSNER-net$_{\text{bert-large-wwm}}$+NELL	73.17	75.05
w/o NELL	71.59	73.69
w/o NER task	72.31	74.42
w/o self-attention	72.85	74.95

另外，FSNER-NET 的中的 self-attention 也能有少许提升，这很可能使因为 self-attention 能够进一步得到融合文本的交互信息表示，也说明传统机器阅读理解架构的有效性。总之，我们提出的 FSNER-net，不仅给机器阅读理解有效地引入了外部知识，而且联合命名实体识别的辅助任务提高了机器阅读理解的实体感知能力，进一步提高了机器阅读理解模型融合外部知识的能力，最终在效果上远超传统的方法。

4.6 本章小结

本章提出了结合实体感知增强和外部知识的机器阅读理解模型 FSNER-net，既通过有效的方法融合了外部知识，又创造性地联合命名实体识别任务提高了机器阅读理解模型的实体感知能力，从而进一步提高了机器阅读理解模型引入实体相关的外部知识的能力，通过实验对比，最终在效果上远超目前其他的方法。下一步我们的研究将尝试使用更大的预训练语言模型 XLNET[39]、ALBERT[52]等为基础来设计模型，或者在预训练语言模型预训练的过程中直接添加实体相关的任务如命名实体识别任务，来直接增强预训练语言模型的实体感知能力。

结　论

人类阅读理解和机器阅读理解一个很大的差异是，人类很善于利用除了文本之外的一些外部知识，来辅助自己理解获取答案。然而当前的很多机器阅读理解方法更多的是在文本匹配层面，仅仅是根据阅读理解所提供的文本和问题来寻找答案。但是现实世界中的机器阅读理解任务很复杂，仅仅根据所提供的文本和问题，无法获得问题的答案，需要借助一些常识性的外部知识信息。

本文以机器阅读理解引入外部知识为切入点，通过检索机器阅读理解任务相关的外部知识信息，然后设计对应的方法将其加入到机器阅读理解的获取问题答案的过程中，从而提高了机器阅读理解获取问题的答案的性能。主要包含以下三项研究工作：

（1）基于预训练语言模型的隐式的引入外部知识的方法。由于预训练语言模型借助优秀的深度学习架构，能够很好的利用大量的无标注的文本，而这些大量的无标注的文本中已经包含很多的外部知识，因此，直接使用预训练语言模型构建机器阅读理解模型来隐式地引入外部知识，相比于的传统的机器阅读理解方法，在需要外部知识的机器阅读理解上效果提升显著，说明预训练语言模型的预训练过程中已经隐式地引入了外部知识，证明了我们的假设。

（2）基于 Attention 机制的显式的引入外部知识的方法。针对当前很多外部知识库如 NELL、WordNet 等都包含丰富的知识信息，使用适当的方法检索出相关的知识之后，借鉴传统的机器阅读理解中注意力机制设计的启发，利用注意力机制设计知识融合模块，显示地将这些外部知识融合到现有的机器阅读理解模型当中，在需要外部知识的机器阅读理解上效果有进一步提升，而且使用不同的知识库提升幅度也不一样，说明了外部知识库对与阅读理解还是有很大帮助的，而且选择合适的知识库也很重要。总之，证明了我们提出的基于 Attention 机制的显式的引入外部知识的方法的有效性。

（3）结合实体感知增强的引入外部知识的方法。当前的以预训练语言模型为基础的机器阅读理解模型对文本进行分词之后会将部分实体拆开，而很多检索到的外部知识都是实体级别的，这会影响机器阅读理解融合实体相关的外部知识，针对这个问题，我们提出一个结合实体感知增强和外部知识的阅读理解，该方法添加了命名实体识别的辅助任务，该任务和机器阅读理

解任务一起联合训练，增强了机器阅读理解模型的实体感知能力，从而进一步提高机器阅读理解获取问题答案的能力，最后在需要外部知识的机器阅读理解数据集上发现进一步提高了机器阅读理解的效果，而且实体感知增强也进一步增强了模型结合实体相关的外部知识的能力，这些证明了我们提出的方法的有效性。

尽管本文对融合外部知识的机器阅读理解进行了一些研究而且取得了一些成果，然后由于结合外部知识的复杂性，仍存在改进和完善的空间。

（1）由于资源限制，当前融合外部知识的方法都是基于预训练语言模型 Bert，没有使用更优秀的预训练语言模型，后续工作会考虑使用更优秀的预训练语言模型 XLNET、ALBERT 等来进一步进行实验。

（2）由于需要外部知识的机器阅读理解任务需要的外部知识很多都是实体相关的知识，即使采用注意力机制方法显示地引入，仍然还是只能够引入有限的知识，部分文本可能仍然缺乏知识，后续工作将考虑直接在预训练语言模型的预训练过程中直接添加实体识别相关的辅助任务进行预训练，以便能够更好的利用大量的无标注的文本，直接从这些预料中学习到实体相关的外部知识，从而进一步提高机器阅读理解的效果。

参考文献

[1] 李济洪, 杨杏丽, 王瑞波等. 基于规则的中文阅读理解问题回答技术研究[J]. 中文信息学报, 2009, 23(4): 3–10.

[2] 毛先领, 李晓明. 问答系统研究综述[J]. 计算机科学与探索, 2012, 6(3): 193–207.

[3] 郝晓燕, 李济洪, 由丽萍等. 中文阅读理解语料库构建技术研究[J]. 中文信息学报, 2007, 21(6): 29–35.

[4] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teachingmachines to read and comprehend. In Advances in Neural InformationProcessing Systems 28, pages 1693–1701. Curran Associates, Inc.

[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392. Association for Computational Linguistics.

[6] Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. arXiv e-prints,arXiv:1704.05179.

[7] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611. Association for Computational Linguistics.

[8] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao,Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches, pages 96–105.

[9] Shuohang Wang and Jing Jiang. 2017. Machine comprehension using match-lstm and answer pointer. In International Conference on Learning Representations (ICLR).

[10] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In

International Conference on Learning Representations (ICLR).

[11] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-overattention neural networks for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 593–602. Association for Computational Linguistics.

[12] Caiming Xiong, Victor Zhong, and Richard Socher.2017. Dynamic coattention networks for question answering. In International Conference on Learning Representations (ICLR).

[13] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 189–198. Association for Computational Linguistics.

[14] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In International Conference on Learning Representations (ICLR).

[15] Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1499–1509. Association for Computational Linguistics.

[16] Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, Technical report, OpenAI.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv e-prints, arXiv:1810.04805.

[18] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. arXiv e-prints, arXiv:1810.12885.

[19] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. arXive-prints, arXiv:1803.05457.

[20] Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. Semeval2018 task 11: Machine comprehension using commonsense knowledge. In Proceedings of The 12th International Workshop on Semantic Evaluation, pages 747–757. Association for Computational Linguistics.

[21] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2381–2391. Association for Computational Linguistics.

[22] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. arXiv e-prints, arXiv:1811.00937.

[23] Teng Long, Emmanuel Bengio, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. World knowledge for reading comprehension: Rare entity prediction with hierarchical lstms using external descriptions. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 825–834, 2017.

[24] Bishan Yang and Tom Mitchell. Leveraging knowledge bases in lstms for improving machine reading. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1436–1446, 2017.

[25] Todor Mihaylov and Anette Frank. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 821–832, 2018.

[26] Yibo Sun, Daya Guo, Duyu Tang, Nan Duan, Zhao Yan, Xiaocheng Feng, and Bing Qin. Knowledge based machine reading comprehension. arXiv preprint arXiv:1809.04267, 2018.

[27] Chao Wang and Hui Jiang. Exploring machine reading comprehension with explicit knowledge. arXiv preprint arXiv:1809.03449, 2018.

[28] Dirk Weissenborn, Toma´s Ko ˇ cisk ˇ y, and Chris Dyer. `2017. Dynamic integration of background knowledge in neural NLU systems. arXiv e-prints, arXiv:1706.02596.

[29] Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018.Commonsense for generative multi-hop question answering tasks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages

4220–4230. Association for Computational Linguistics.

[30] Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 821–832. Association for Computational Linguistics.

[31] Xiaoman Pan, Kai Sun, Dian Yu, Heng Ji, and Dong Yu. 2019. Improving question answering with external knowledge. arXiv e-prints, arXiv:1902.00993.

[32] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2406–2417. Association for Computational Linguistics.

[33] Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. 2018. Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. In Proceedings of The 12th International Workshop on Semantic Evaluation, pages 758–762. Association for Computational Linguistics.

[34] Yang, An, et al. "Enhancing Pre-Trained Language Representations with Rich Knowledge for Machine Reading Comprehension." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[35] CUI Y, LIU T, CHEN Z 等. Consensus Attention-based Neural Networks for Chinese Reading Comprehension[J]. 2016:1-10.

[36] 李济洪, 王瑞波, 王凯华等. 基于最大熵模型的中文阅读理解问题回答技术研究[J]. 中文信息学报, 2008, 22(6): 55–62.

[37] Yang, Bishan, and Tom Mitchell. "Leveraging knowledge bases in lstms for improving machine reading." arXiv preprint arXiv:1902.09091 (2019).

[38] Duan, Xingyi, et al. "Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension." China National Conference on Chinese Computational Linguistics. Springer, Cham, 2019.

[39] Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." Advances in neural information processing systems. 2019.

[40] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[41] Bengio, Yoshua, et al. "A neural probabilistic language model." Journal of

machine learning research 3.Feb (2003): 1137-1155.

[42] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[43] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[44] Abnar, Samira, et al. "Incremental Reading for Question Answering." arXiv preprint arXiv:1901.04936 (2019).

[45] Hermann, Karl Moritz, et al. "Teaching machines to read and comprehend." Advances in neural information processing systems. 2015.

[46] Kadlec, Rudolf, et al. "Text Understanding with the Attention Sum Reader Network." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.

[47] Cui, Yiming, et al. "Attention-over-Attention Neural Networks for Reading Comprehension." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.

[48] Seo, Minjoon, et al. "BI-DIRECTIONAL ATTENTION FLOW FOR MACHINE COMPREHENSION." arXiv preprint arXiv:1611.01603 (2016).

[49] Wang, W. R-NET: machine reading comprehension with self-matching networks. Natural Language Computer Group, Microsoft Reserach. Asia, Beijing. China, Technical Report 5, 2017.

[50] 郭剑毅, et al. "基于层叠条件随机场的旅游领域命名实体识别." 中文信息学报 23.5 (2009): 47-53.

[51] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[52] Lan, Zhenzhong, et al. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." International Conference on Learning Representations. 2019.

攻读硕士学位期间发表的论文及其它成果

（一） 发表的学术论文

[1] 乐远，张宇，刘挺. 结合实体感知增强和外部知识的阅读理解. The 26th China Conference on Information Retrieval, CCIR 2020. 已投出

（二）参与的科研项目及获奖情况

[1] 2019YFF0303003，面向冬奥场景的多语种智能问答关键技术研究，国家科技部重点研发计划

[2] 61976068，面向智能客服的问题语义分析相关技术研究，国家自然科学基金项目

哈尔滨工业大学学位论文原创性声明及使用授权说明

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《融合外部知识的机器阅读理解方法研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：  日期：2020 年 06 月 27 日

学位论文使用授权说明

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：  日期：2020 年 06 月 27 日

导师签名：  日期：2020 年 06 月 27 日

致　谢

光阴似箭，日月如梭，转眼间两年的研究生求学生涯已经接近尾声，我也将步入人生的下一阶段。这一路走来，收获良多，回首以往，是大家一路上的支持与帮助，鼓舞着我砥砺前行。

衷心感谢我的导师张宇教授，正是由于当初初次见面一个点头，让我有机会能够在优秀的社会计算与信息检索研究中心实验室进行学习。在这两年的时光里，无论是科研上还是在生活中，张老师都给予了我悉心的指导与帮助。在科研上，张老师是我的导师，从进入实验室学习开始，无论是竞赛、科研项目，还是硕士论文选题、开题、中期直到最后的结题，张老师都给了我宝贵的建议与帮助。特别是学会看数据、找准问题的科研思路让我受益匪浅，这些将是我今后学习、工作的宝贵财富。在生活上，张老师更像是我的父亲，大到每逢节假日的安全叮嘱，小到生活作息规律的提醒，让我感受到长辈的亲切关怀，每次都感觉心里暖暖的。能够在研究生期间作为你的学生，是我莫大的幸运。

感谢社会计算与信息检索研究中心的主任刘挺老师，正是你当初对我的一次邮件提醒，才让我没有错失进入赛尔实验室进行学习的机会。你的宏大格局、高瞻远瞩、宽广胸襟让我深感佩服，而且你对实验室每个同学的关心也让人感到浓浓暖意。硕士找工作期间，每次提到我是来自于你的实验室，几乎总是一路绿灯，作为你实验室的学生，真心感到自豪。

感谢社会计算与信息检索研究中心的秦兵老师、车万翔老师、刘铭老师、赵研研老师、张伟男老师、丁效老师、冯骁骋老师，感谢老师们一直关心着我的学业与前途，为我排忧解难，照亮我前行的道路。

感谢实验室的博士师兄尹庆宇、郭茂盛、齐乐，在刚进入实验室的时候给予了我诸多帮助，让我很顺利的进入到科研的道路上。

感谢实验室的兄弟姐妹，施琦、颜欣、李威宇、妥明翔、尹志博、将润宇、孙月晴、栗扬帆，学习和生活之中都收到了你们的帮助。感谢实验室的同窗伙伴，廖阔、王必聪、陈昱宇、蔡碧波、胡啸、吴洋、冯夏冲、王帅、石乾坤、冯掌印、茅佳峰、孙卓、陆鑫、刘元兴，我们共同成长，一起进步，两年的研究生时光，因为有你们的陪伴，才更加丰富多彩。感谢实验室的师弟师妹们，在与你们的交流中我也不断成长，祝愿你们前程似锦。

感谢我的室友刘荣鑫、张思齐、周瑞亮、温凯明、饶仲文、张良仁、苏林，我们朝夕相处，感谢你们一路相伴。

感谢我的健身伙伴，妥明翔、吴浩楠、张涵、刘强、陈昱宇、卢坤，因为你们，让我的闲暇时光更加丰富多彩。

最后感谢我的父母，是你们一路上一直支持着我，见证着我的成长，一直是我最坚强的后盾，欲报之德，昊天罔极。

寥寥数语，纸短情长，感谢所有关心和帮助过我的人，感谢一路有你！