基于小句复合体的中文机器阅读理解研究*

王瑞琦	罗智勇*	刘祥	韩瑞昉	李舒馨
北京语言大学	北京语言大学	北京语言大学	北京语言大学	北京语言大学
1159925366@qq.com	luo_zy@blcu.edu.cn	SmileTM@qq.com	15321103341@163.com	971787780@qq.com

摘要

机器阅读理解任务要求机器根据篇章文本回答相关问题。本文以抽取式机器阅读理解为例，重点考察当问题的线索要素与答案在篇章文本中跨越多个标点句时的阅读理解问题。本文将小句复合体结构自动分析任务与机器阅读理解任务融合，利用小句复合体中跨标点句话头-话体共享关系，来化简机器阅读理解任务的难度；并设计与实现了基于小句复合体的机器阅读理解模型。实验结果表明：在问题线索要素与答案跨越多个标点句时，答案抽取的精确匹配率（EM）相对于基准模型提升了3.49%，模型整体的精确匹配率提升了3.26%。

关键词： 机器阅读理解 ； 跨标点句问答 ； 小句复合体

Machine Reading Comprehension Based on Clause Complex

Wang Ruiqi	Luo Zhiyong*	Liu Xiang	Han Ruifang	Li Shuxin
北京语言大学	北京语言大学	北京语言大学	北京语言大学	北京语言大学
1159925366@qq.com	luo_zy@blcu.edu.cn	SmileTM@qq.com	15321103341@163.com	971787780@qq.com

Abstract

The machine reading comprehension task requires the machine to answer relevant questions according to the context. Taking extractive machine reading comprehension as an example, this paper focuses on the reading comprehension problems when the clue elements and the answer elements of question span multiple punctuation sentences in the text. In this paper, the automatic analysis task of clause complex structure is integrated with the machine reading comprehension task, and the difficulty of machine reading comprehension task is reduced by using the naming structure relationship of clause complex; and the machine reading comprehension model based on clause complex is designed and implemented. The experimental results show that: when the question clue elements and the answer elements span multiple punctuation sentences, the EM of the answer extraction increases by 3.49 percentage points with the benchmark model, and the EM of the whole model increases by 3.26 percentage points.

Keywords: machine reading comprehension , Cross-punctuation Q&A , clause complex

第二十届中国计算语言学大会论文集，第723页-第735页，呼和浩特，中国，2021年8月13日至15日。

1 引言

机器阅读理解（MRC）任务与人类阅读理解任务相似，是指计算机根据指定篇章文本回答相关问题的过程。近年来，随着深度学习技术，特别是词向量表示、预训练语言模型方法的发展，机器阅读理解模型的性能得到巨大提升，甚至在个别机器阅读理解数据集评测任务中逼近或超越了人类的水平（顾迎捷et al. (2020)）。但在涉及到远距离、深层次的语义关系时，现有的深度学习方法仍然没有取得实质性的突破。

在机器阅读理解任务中，这一现象主要体现在：中文机器阅读理解任务的篇章文本（context）长度较长，经常包含多个标点句；且问题（question）对应的答案与回答此问题需要的线索要素在篇章文本中跨越多个标点句，这种情况给机器阅读理解任务的答案抽取带来较大困难。具体示例如图1所示。

Context：**徐珂**，原名昌，字仲可，浙江杭县（今属杭州市）人。光绪举人。后任商务印书馆编辑。参加南社。**1901年**在上海担任了《外交报》、《东方杂志》的编辑。
Question：**1901年**，**徐珂**担任什么职务?
Answer：1901年在上海担任了《外交报》、《东方杂志》的编辑

话头-话体换行缩进图示：
c1:**徐珂**，
c2: 原名昌，
c3: 字仲可，
c4: 浙江杭县（今属杭州市）人。
c5: 光绪举人。
c6: 后任商务印书馆编辑。
c7: 参加南社。
c8: **1901年**在上海担任了《外交报》、《东方杂志》的编辑。

Figure 1: CMRC2018数据集中跨标点句远距离问答样例

图1中，我们使用换行缩进形式直观地表示了篇章文本中话头-话体共享关系(宋柔 (2017))：此样例中篇章文本（context）一共有8个标点句（用换行表示），其中，c2至c8等7个标点句均共享标点句c1中的话头“徐珂”（用缩进表示）。Question1中的线索要素包括：“1901年”（位于c8）、“徐珂”（位于c1，“**加粗**”显示），问题要素是“担任什么职务”（用“波浪线”标记），该问题对应的答案是“1901年在上海担任了《外交报》、《东方杂志》的编辑”（位于c8,用“下划线”标记）。线索要素与问题答案之间在篇章文本中跨越8个标点句，属于远距离关联关系。在本例中，要想让模型准确地抽取出答案，必须在标点句c1中，找到标点句c8缺失的话头“徐珂”，并将c8补充完整，再进行问答。

我们针对于这一跨标点句、远距离关联的现象，在Cui et al. (2018)发表的CMRC2018阅读理解数据集上进行了标注，结果如表1所示：其中，涉及跨标点句问答的问题占全部的训练数据的67.89%；而且，基于的BERT基线系统在跨标点句问答的问题上的精确匹配率（EM）为61.02%，相比于其他问题上的EM值72.02%，降低了11.00%，一定程度上说明了跨标点句问答给答案的抽取带来困难。

目前针对中文机器阅读理解任务的研究方法多为模型结构的更改或增加实体信息等外部知识，却很少关注数据中普遍存在的跨标点句问答问题。本文应用小句复合体理论降低中文机器阅读理解任务中跨标点句问答问题答案抽取的难度，提高模型的性能。小句复合体理论基于逻辑语义关系和成分共享关系研究汉语中跨标点句的句间关系，本文应用该理论对阅读理解数据的篇章文本进行分析，使用标点句间的话头-话体共享信息构建远距离标点句之间的联系。篇章文本中的标点句补全缺失的话头话体成分，转化为自足的话题结构后，问题的线索要素与答案要素将处于同一话头自足句或者同一小句复合体结构内部，便于答案的抽取。标点句之间的远

基金项目：本文得到国家自然科学基金（62076037）资助

	数量	占比	BERT-baseline验证集的F1	BERT-baseline验证集的EM
全部问题	10142	100%	84.08%	64.55%
跨标点句问答问题	6885	67.89%	82.53%	61.02%
非跨标点句问题	3257	32.11%	87.36%	72.02%

Table 1: CMRC2018阅读理解数据集跨标点句问答统计情况

距离成分共享关系可以保证标点句语义的完整性，为模型提供额外的语义信息，提高模型的语义表示能力，对于机器阅读理解等自然语言处理任务具有基础性的意义。

本文的主要贡献在于：以抽取式机器阅读理解任务为例，重新审视跨标点句阅读理解问题；提出将小句复合体结构自动分析任务与机器阅读理解任务融合的策略，利用小句复合体中话头-话体远距离共享关系，为模型提供句级别的结构化语义信息，降低远距离答案抽取的难度；提出了基于小句复合体的机器阅读理解模型，并验证了小句复合体话头-话体共享关系在机器阅读理解任务中的作用效果；另外，本文在CMRC2018阅读理解数据集上的实验结果表明：小句复合体结构自动分析任务对机器阅读理解任务中的远距离跨标点句问答问题有明显的效果，与基准模型相比，基于小句复合体的机器阅读理解模型的整体精确匹配率（EM）提升3.26%，其中跨标点句问答问题的EM提升3.49%。

本文第2节介绍相关研究；第3节介绍相关概念和基于小句复合体的机器阅读理解任务建模；第4节介绍基于小句复合体的机器阅读理解模型设计；第5节介绍实验结果及分析，第6节为总结与展望。

2 相关研究

机器阅读理解任务的起源可以追溯到20世纪70年代，但是由于数据集规模过小和传统的基于规则的方法的局限性，当时的机器阅读理解系统性能较差，不能满足实际应用的需要。Lehnert (1977)提出基于脚本和计划的框架QUALM，专注于语用问题，以及故事的上下文背景对回答问题的影响。Hirschman et al. (1999)提出一个包含60个故事的数据集，并提出Deep Read系统使用基于规则的词袋模型进行浅层语言处理，加入词干提取、指代消解等帮助理解文本。Riloff and Thelen (2000)提出基于词汇和语义对应的QUARC系统。这些基于规则的方法，准确率最高只有30%-40%。

机器学习兴起后，阅读理解任务被定义为有监督问题。MCTest（Garcia-Diaz (2012)）和ProcessBank数据集（Berant et al. (2014)）的提出，促进了该任务的发展。MCTest提出了滑动窗口法计算篇章与问题、答案之间的信息重叠度，还提出将答案转化为语句然后做文本蕴含的方法。基于检索技术的阅读理解模型，通过关键词匹配在文章中搜索答案，存在局限性，匹配度高的结果有时并不是问题的答案。此阶段，机器学习模型对机器阅读理解任务带来的提升有限，主要由于模型使用语义角色标注系统等语言工具提取特征，这些工具多用单一领域的语料训练，难以泛化；而且数据集过小不足以支撑模型的训练。

2015年以后，深度学习飞速发展，提出了很多大规模数据集和神经网络模型，模型的效率与质量大幅度提升，在一些数据集上甚至可以达到了人类平均水平。Hermann et al. (2015)提出的“Attentive Reader”基础模型，成为了今后许多研究的基础，该模型将篇章和问题使用双向RNN分别表示后，利用attention机制在篇章中寻找问题相关的信息，最后根据相关程度给出答案的预测。18年BERT提出时也提供了阅读理解问答的模型架构。各种大规模预训练语言模型的提出推动了机器阅读理解任务的发展。

目前主流的语言表示模型如Peters et al. (2018)提出的ELMo、Devlin et al. (2018)提出的BERT等仅发掘了如character embedding或word embedding等对上下文敏感的特征，缺乏对结构化语言学信息的考虑。对于上下文语义表示和学习不足的问题，解决方法多为添加额外的语言学知识。Zhang et al. (2019b)提出：Semantics-aware BERT模型，将BERT与语义角色标注任务结合，用谓词-论元信息来提升阅读理解模型的语言表示能力，提高问题的准确率。该融合模型在机器阅读理解任务上应用的有效性，表明显式的上下文语义信息可以与预训练

语言模型的语言表示融合来提高模型在机器阅读理解任务上的性能。Zhang et al. (2019a)提出的ERNIE，用知识图谱来增强语言表示，该模型在BERT的基础上，加入了实体、短语等语义知识。这两种方法均应用额外的语义信息增强模型的表示，提高模型的性能，证明了结合必要的外部知识对提升模型性能的有效性。但语义角色标注和实体信息不能处理机器阅读理解任务中跨标点句问答的问题。

现有基于CMRC2018数据集的研究方法多为对于分词、或者模型结构的更改。在该数据集上取得较好结果的的MacBERT（Cui et al. (2020)）和RoBERTa-wwm-ext-large模型（Liu et al. (2019)），使用的改进方法都是针对预训练策略的更改，同样没有考虑篇章文本中存在的远距离问答的问题。而小句复合体结构分析可以为模型提供句间的语义信息，用话头-话体的共享关系来增强标点句间的语义完整性和相关性，简化答案抽取的难度，从而提升模型效果，故本文尝试将小句复合体结构分析与机器阅读理解任务融合，解决跨标点句问答问题。

小句复合体研究任务已经历十几年，定义、分类及内部理论体系已经成熟，在此基础上的话头识别工作有如下成果。起初仅对小句复合体语料中的堆栈类型数据进行单个标点句的话头结构分析，蒋玉茹and 宋柔 (2012a)使用穷举法找出全部候选话头，再采用语义泛化和编辑距离两种手段选出合适的话头，识别正确率为73.36%。蒋玉茹and 宋柔 (2012b)又采用相同方法研究堆栈类型标点句序列的话头结构识别。将各标点句的全部候选话头存储于树结构，选取概率最大的路径获得话头序列，最终正确率为64.99%。由于穷举法对系统执行效率和话题句识别的准确率存在限制，蒋玉茹and 宋柔 (2014)蒋玉茹利用标点句在篇章中的位置和话头的语法特征等信息减少生成的候选话头的数量，从而提高模型的识别效率和效果。Teng et al. (2018)提出的基于Attention-LSTM的神经网络模型在单个标点句的话头识别任务上的正确率达到81.74%。胡紫娟 (2020)在前面研究的基础上，增加了对新支、汇流、后置类型数据的分析，并且添加了句尾缺失成分（话体）的识别，总的正确率有93.24%。为小句复合体理论在实际任务中的应用打下基础。

3 基于小句复合体的机器阅读理解任务建模

3.1 机器阅读理解任务

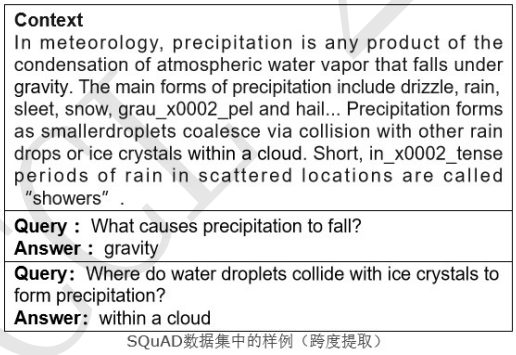

Figure 2: 跨度提取型阅读理解样例

机器阅读理解任务主要分为完形填空、多项选择、跨度提取和自由回答四种类型，另外还有会话式回答，多段式阅读理解等。本文涉及到的类型为跨度提取型阅读理解，如图2所示，该任务要求在原文中抽取一个片段作为答案。

3.2 跨标点句问答

本文使用线索要素、问题要素、答案要素来描述阅读理解任务中远距离跨标点句问答的情况:

线索要素：问题中明确给出的关键词，是寻找答案的限定条件。

问题要素：问题的提问方式，如when、where、how，what，why、who等。

答案要素：原文中的实体、短语、句子。阅读理解问题的答案，与问题要素相对应。

全部要素是否跨标点句：将问题中的问题要素替换成答案要素，并转化成陈述句，其中包含的线索要素与答案要素在原文中是否处于同一标点句。

跨标点问答分为两种类型：第一种答案要素较短，为词、短语、或者一个标点句，与线索要素距离较远而跨多个标点句（如图3中样例所示）。另一种类型答案要素很长，答案在篇章文本中跨标点句（如图11中样例所示）。

Context：**范廷颂**枢机，圣名保禄·若瑟，是越南罗马天主教枢机。1963年被**任为主教**；**1990年**被擢升为天主教河内总教区宗座署理；1994年被擢升为总主教，同年年底被擢升为枢机；2009年2月离世。范廷颂于1919年6月15日在越南宁平省天主教发艳教区出生；童年时接受良好教育后，被一位越南神父带到河内继续其学业。1994年3月23日，范廷颂被教宗若望保禄二世擢升为天主教河内总教区总主教并兼天主教谅山教区宗座署理。

Question1：**范廷颂**是什么时候被**任为主教**的?
Answer：1963年
Question2：**1990年**，**范廷颂**担任什么职务?
Answer：1990年被擢升为天主教河内总教区宗座署理

Figure 3: CMRC2018阅读理解数据集要素跨标点句样例

图3中，Question1的线索要素为：“范廷颂”和“任为主教”，问题要素是“什么时候”，答案要素是“1963年”。将问题中的问题要素替换为答案后，问句可以转化为陈述句“范廷颂是1963年被任为主教”，该陈述句在原文中对应的标点句序列是“范廷颂枢机，圣名保禄·若瑟，是越南罗马天主教枢机。1963年被任为主教；”，线索要素与答案要素跨越4个标点句。Question2的线索要素是“1990年”和“范廷颂”，问题要素是“担任什么职务”，答案要素“1990年被擢升为天主教河内总教区宗座署理”，全部要素在原文中跨越5个标点句。两问题均属于远距离跨标点句问答问题。

3.3 小句复合体理论

标点句：本文的标点句是指被逗号、分号、句号、问号、叹号所分隔出的词语序列。如图4中样例共有13个标点句。

话头、话体：在微观话题角度，话语的出发点叫做话头（Naming），话体（Telling）是对话头的说明。

话头结构：话头话体间关系构造的多个标点句之间的结构称为话头结构。换行缩进标注体系是使用空格表明话头结构的方式。

话头自足句：如果一个标点句本身不缺少话头话体，也不作为其他标点句的话头，那么称之为话头自足句（NT小句，NTC）。小句复合体结构自动分析的任务的目标就是就是分析标点句缺失的话头话体，并补全为话头自足句。

小句复合体：是话头共享关系和逻辑关系都不可分割的最小标点句序列。主要有堆栈、汇流、新支、后置四种类型。本文将阅读理解的篇章文本看作一个整体，分析各标点句之间的话头话体共享关系。

c1:**范廷颂**枢机，
c2: 圣名保禄·若瑟，
c3: 是越南罗马天主教枢机。
c4: 1963年被**任为主教**；
c5: **1990年**被擢升为天主教河内总教区宗座署理；
c6: 1994年被擢升为总主教，
c7: 同年年底被擢升为枢机；
c8: 2009年2月离世。
c9:范廷颂于1919年6月15日在越南宁平省天主教发艳教区出生；
c10: 童年时接受良好教育后，
c11: 被一位越南神父带到河内继续其学业。

Figure 4: 用换行缩进表示话头话体共享关系

将图3中的标点句使用小句复合体理论进行分析，并用换行缩进格式表示其话头结构，如图4所示，标点句c1至c8的8个标点句处于同一小句复合体结构，c9至c11的3个标点句处于另一小句复合体结构，这种关系的划分基于话头-话体共享关系与逻辑语义关系（本课题不研究逻辑语义关系）。该样例中，标点句c1成分完整，但被后面的标点句共享话头，因此c1至c8处于一个小句复合体。其中，c2至c6和c8共享c1中的话头“范廷颂”，c7共享的话头来自标点句c1和c8，为“范廷颂1994年”。c9不共享其他标点句中的话头，且自身不缺少成分，但被c10、c11共享话头“范廷颂”，故c9至c11这3个标点句被划分在一个小句复合体结构。

3.4 基于小句复合体的机器阅读理解研究

将阅读理解的篇章文本使用小句复合体结构分析，转换为换行缩进模式，并补全为话头自足句，线索要素和与答案要素处于同一标点句或同一小句复合体结构。根据线索要素与问题要素在篇章文本中抽取答案，答案要素短的样例在一个话头自足句中提取即可；答案要素跨越多个标点句的样例在一个小句复合体结构内抽取答案，同一小句复合体内的标点句由于共享话头或话体被组织到一起，更容易把跨标点句的答案要素提取完全。

c1:范廷颂枢机，
c2:范廷颂圣名保禄·若瑟，
c3:范廷颂是越南罗马天主教枢机。
c4:**范廷颂**1963年被任为主教；
c5:**范廷颂1990年**被擢升为天主教河内总教区宗座署理；

Figure 5: 话题自足句

将图4中以换行缩进模式表示的部分标点句补全缺失成分，转化为话头自足句（即NT小句）如图5所示。从图中可以看出，标点句c4“范廷颂1963年被任为主教”包含了Question1的全部线索要素与答案要素，c5“范廷颂1990年被擢升为天主教河内总教区宗座署理”包含了Question2的全部线索要素与答案要素。直接在一个标点句完成问答，化简了答案抽取的难度。

3.5 机器阅读理解任务的机器学习问题描述

本文使用的CMRC2018阅读理解数据集的类型是跨度提取，该类型任务可定义为：将机器阅读理解任务看做一个三元组$< C, Q, A >$，给定长度为n篇章上下文$C = \{t_1, t_2, \ldots, t_n\}$以及问题Q，要求在原文C中提取一个子序列$a = \{t_i, t_{i+1}, \ldots, t_{i+k}\}$作为正确答案，通过最大化条件概率$P = (a \mid C, Q)$来获取答案A。图6为基于BERT的机器阅读理解模型图。

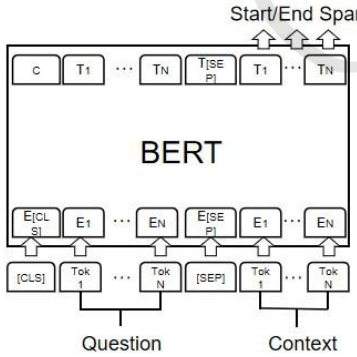

Figure 6: 机器阅读理解模型

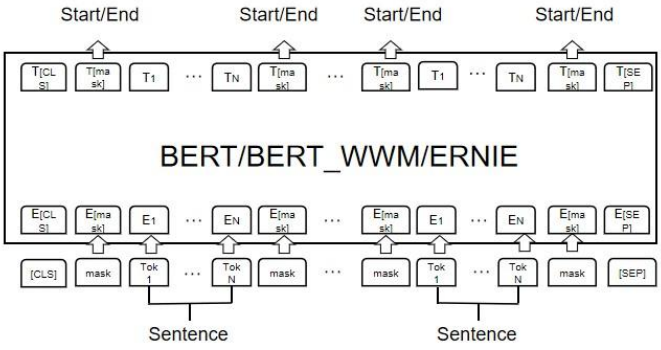

Figure 7: 小句复合体结构自动分析模型

3.6 小句复合体结构自动分析任务的机器学习问题描述

小句复合体结构自动分析任务可以定义为：已知小句复合体$C_1, \ldots, C_n$，求对应的话头自足句$Z_1, \ldots, Z_n$的过程。在每个标点句首尾插入[MASK]，并在[MASK]处预测话头的位置

（start，end），即预测共享的成分，补全共享的成分便能得到相应的话头自足句。不添加[MASK]则使用T_1, T_n位置的向量预测。图7为基于BERT的小句复合体结构自动分析模型图。

4 基于小句复合体的机器阅读理解模型

4.1 融合模型一：$BERT_{NTC}$

$BERT_{NTC}$模型将小句复合体结构自动分析任务作为预训练任务，先在中文小句复合体数据集上进行预训练，训练后将模型保存，之后在机器阅读理解数据集上对模型进行微调。此方法用于初步验证小句复合体结构自动分析任务对于机器阅读理解任务是否有作用。模型的结果于第五章实验结果部分展示并分析。

4.2 融合模型二：BERT_NTC/MRC

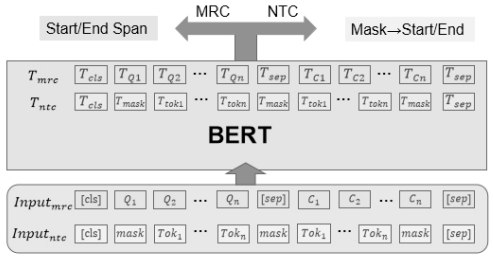

Figure 8: 融合模型二

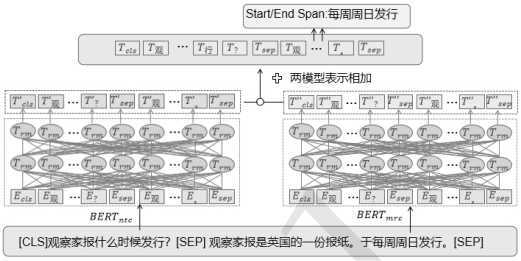

Figure 9: 融合模型三

BERT_NTC/MRC模型如图8所示，为机器阅读理解任务与小句复合体结构自动分析任务融合的第二种方法。本模型中采用多任务学习的方式同时训练小句复合体结构分析任务与机器阅读理解任务，两任务共享一个BERT模型的参数。训练时，对于一个批次的数据，如果是机器阅读理解类型的输入，经过BERT预训练语言模型编码，获取上下文语义信息及话头-话体共享信息后，进入MRC的输出层，进行答案片段Start/End span的预测；如果是小句复合体结构分析类型的数据，获得上下文表示后进入NTC的输出层，通过MASK位置或者标点句的首位位置的向量预测缺失的话头话体的位置Start/End。

4.3 融合模型三：BERT_NTC_add_MRC

图9中，BERT_NTC_add_MRC模型为机器阅读理解任务与小句复合体结构自动分析任务融合的第三种方法。该模型可以分为2个模块，分别使用独立的预训练语言模型。左侧模型$BERT_{ntc}$进行话头-话体结构信息抽取，右侧$BERT_{mrc}$抽取上下文及问题信息。训练时，$BERT_{ntc}$模型训练好然后保存参数，使该模型拥有表达话头-话体信息的能力；再训练$BERT_{mrc}$模型。

4.4 融合模型三的各层设计

BERT_NTC_add_MRC模型分为三层：编码层、信息交互层和答案预测层。

编码层:该层的功能是将阅读理解任务的输入$Input = [CLS] + Context + [SEP] + Question + [SEP]$转换成计算机可以理解的向量$E_{mrc}$，$E_{mrc}$向量由词向量(Token Embedding)、位置向量(Position Embedding)、句向量(Segment Embedding) 拼接得到。

信息交互层:在此层E_{mrc}通过两个BERT预训练语言模型获得包含上下文信息、问题信息和话头-话体共享信息的词向量表示。

$BERT_{ntc}$模型的输入格式为：$E_{ntc} = E_{[CLS]} + E_{Context} + E_{[SEP]}$

$BERT_{mrc}$模型的输入格式为：$E_{mrc} = E_{[CLS]} + E_{Context} + E_{[SEP]} + E_{Question} + E_{[SEP]}$

经过BERT编码后，分别取两模型最后一层隐藏层的输出T_{ntc}与T_{mrc}，将二者相加。

$T = T_{mrc} + T_{ntc}$，把T接全连接层得到：$O^{NX2} = FC(T^{NXD})$。其中D为隐藏层大小768，经全连接层FC将维度转换为2，获得每个字作为start和end的logit值。

答案预测层：本层对start/end logits对进行相加计算，经softmax选择概率最高的一组，经处理得到最后的答案。

基于小句复合体的阅读理解模型答案预测总计算公式如下所示：
$start/end\ \ logit = softmax(FC(BERT_{ntc}(E_{ntc}) + BERT_{mrc}(E_{mrc})))$。

5 实验

5.1 数据集

小句复合体语料：北京语言大学中文小句复合体标注语料（简称小句复合体语料），包括百科全书、政府工作报告、新闻、小说4个领域，其中共有小句复合体9256 个，标点句37635个。该语料主要用于训练小句复合体结构自动分析模型（NTC模型）。

机器阅读理解语料：CMRC2018阅读理解数据集用以训练机器阅读理解模型。该数据集的篇章文本来自于维基百科，问题由人工撰写，属于片段抽取式阅读理解任务。其中，训练集有篇章2403篇，问题10142个。验证集有篇章848篇，问题3219个。在对CMRC2018阅读理解数据集的研究中发现了存在大量不严谨的地方，如答案长度提取不一致、问题答案不对应、答案位置错误等。经过核对，数据集中约有20%的样例存在此问题，现已全部修改。本文主要使用的是经过纠正后的数据集。

5.2 评估指标

本文采用的评估指标有F1、EM、AVERAGE。对于每个训练样例，预测的文本为prediction，长度为prediction_len，答案文本为answer，长度为answer_len，它们之间最长重合部分为lcs，lcs_len为重合文本的长度。

精确率（Precision）为正确预测为阅读理解答案的部分占全部预测比例。定义为：$Precision = lcs\_len/prediction\_len$。

召回率（Recall）为正确预测为阅读理解答案的部分占全部真实答案比例。定义为：$Recall = lcs\_len/answer\_len$。

模糊匹配率（F1）为精确率和召回率的调和平均数，两个值都很高时才高，可以综合体现预测的水平。定义为：$F1 = (2 * precision * recall)/(precision + recall)$。

精确匹配率（EM）是完全匹配的体现，当prediction=answer记为1，不等时记为0。

对于全部训练样例，count为训练样例的个数，整体的$F1 = \sum(f1)/count$，整体的EM为$EM = \sum(em)/count$。

$AVERAGE = (F1 + EM)/2$，为EM与F1的平均值。

5.3 基线模型水平

	AVERAGE	F1	EM
Bert(tf)	74.318%	84.082%	64.554%
Bert(tf+new_traindata)	78.147%	86.831%	69.463%

Table 2: CMRC2018上机器阅读理解任务基线模型结果

如表2所示为BERT基线模型在CMRC2018阅读理解数据集上的结果，EM为64.55%，更改数据集中不严谨的情况后的EM提升至69.46%。

话头	个数	Accuracy	F1
No	6181	98.94%	98.94%
堆栈	1166	70.95%	78.48%
新支	114	32.14%	47.34%
后置	38	60.00%	64.78%
汇流段	25	53.85%	55.80%
新支	7524	93.24%	94.69%

Table 3: 小句复合体语料库上小句复合体自动分析结果

在如表3所示，在小句复合体数据集中，小句复合体结构分析模型对与所有类型话头总Accuracy为93.24%，F1为94.69%。

5.4 实验设置

模型参数：实验中涉及的模型均使用pytorch搭建，BERT、RoBERTa、RoBERTa_wwm_ext三个预训练语言模型均为base版本，12层的Transformer。Batch size设置为8，Epoch为2，学习率（learning_rate）为3e-5。

5.5 实验结果

在修改后的CMEC2018阅读理解数据集上，对两种基于小句复合体的机器阅读理解模型的效果进行验证，结果如表4所示，其中第一行结果为BERT基线系统的结果，F1为86.83%，EM为69.46%。

$BERT_{NTC}$模型，将小句复合体结构自动分析任务作为预训练任务，在修改后CMRC2018阅读理解数据集上的结果与基准模型相比F1值提升1.11%，EM值提升1.49%。该结果初步证明小句复合体结构自动分析可以为机器阅读理解任务带来一定帮助。

多任务学习模型BERT_NTC/MRC，与基线模型相比，F1提升1.91%，EM提升2.21%。虽然小句复合体结构分析任务与机器阅读理解任务具有一定的相似性，但毕竟属于两个不同的任务。其中，MRC是根据篇章和问题在原文中找到相应的答案片段，而NTC结构分析要对每个缺失成分的标点句在原文中找到相应的话头话体成分。从结果上看，这样的模型融合方式中，小句复合体结构自动分析任务虽然给机器阅读理解任务带来了一定的提升，但提升的效果并不显著。

BERT_NTC_add_MRC模型，将CMRC2018阅读理解数据集篇章文本部分的输入经过两个模型的最后一层隐藏层的表示相加，再进行答案片段的预测。这个过程使篇章文本的表示即包含词级别的信息和篇章上下文的语义信息，也包含基于话头话体共享关系的结构化语义信息。此模型的EM达到72.72%，与基线模型相比提高了3.26%。本文提出的三种小句复合体结构自动分析任务与机器阅读理解任务融合的方法，均可以给机器阅读理解任务带来提升。

更换预训练语言模型为Roberta后，EM为72.60%，添加小句复合体信息使EM提升1.24%。将语言模型换为Roberta_wwm_ext，融合模型三在CMRC2018阅读理解数据集上的EM达到74.96%，提高1.17%。实验结果表明，在不同的预训练语言模型上，小句复合体结构自动分析信息的融入，均能够给模型带来一定的提升，证明了小句复合体理论在实际任务中的应用价值。

	AVERAGE	F1	EM
Bert	78.15%	86.83%	69.46%
$BERT_{NTC}$	79.45%	87.94%	70.95%
BERT_NTC/MRC	79.98%	88.29%	71.67%
BERT_NTC_add_MRC	80.90%	89.09%	72.72%
RoBERTa	80.58%	88.60%	72.57%
RoBERTa_NTC_add_MRC	81.69%	89.56%	73.81%
RoBERTa_wwm_ext	81.45%	89.10%	73.79%
RoBERTa_wwm_ext_NTC_add_MRC	82.62%	90.28%	74.96%

Table 4: 基于小句复合体的机器阅读理解模型验证集上的结果

5.6 小句复合体结构分析对远距离跨标点句问答的解决情况分析

为验证小句复合体结构自动分析任务对机器阅读理解任务中远距离跨标点句问答问题的影响，在全部的数据集中标注出线索要素与答案要素在原文是否跨标点句，为此定义了一个新的标签“tag”。其中，tag为0的样例为较为简单的阅读理解问题；tag为1的样例为跨标点句问答且能应用小句复合体结构信息化简答案抽取难度的问题；tag=2中也存在跨标点句问答问题，但是由于其他因素不能使用话头-话体共享信息化简任务难度，具体标注样例如图10所示。

tag=0：问题与答案的全部要素对应回全文处于同一标点句。

tag=1：全部要素不处于同一标点句，但是处于同一小句复合体内。
tag=2：除上述两种情况之外的情况：包括指代消解、推理等问题。

Context：**观察家报**The Observer）是英国的一份**报纸**。于每周周日**发行**。观察家报实际上是周一到周六发行的卫报的周日版。政治立场偏向自由主义和社会民主主义。观察家报创刊于1791年12月4日，是世界第一份在礼拜日发行的报纸。在1911年，William Waldorf Astor收购了观察家报，之后该报在政治立场上较为倾向保守党。1942年，**该报公开宣言了无党派倾向的编辑方针**，这在当时是较为少有的。
Question1：**观察家报**是哪国的**报纸**?
Answer：英国　　tag：0
Question2：**观察家报**什么时候**发行**?
Answer：每周周日发行　　tag：1
Question3：**观察家报**哪一年**公开宣言了无党派倾向的编辑方针**?
Answer：1942年　　tag：2

Figure 10: tag标注样例

图10中，Question1的线索要素是“观察家报”、“报纸”，问题要素为“哪国”，答案要素为“英国”，全部要素对应回原文处于同一标点句，故tag标签为0。Question2的线索要素为“观察家报”、“发行”，答案要素为“每周周日发行”，将问题转换为陈述句为：观察家报每周周日发行，对应回原文跨越2个标点句，且中间分隔符号为句号，但是按照小句复合体理论，由于这两个标点句共享话头“观察家报（The Observer）”，因此处于同一小句复合体内部，tag为1。Question3的线索要素为“观察家报”、“公开宣言了无党派倾向的编辑方针”，问题要素为“哪一年”，答案要素为“1942年”，虽然跨标点句，但“该报”指代线索要素“观察家报”，为指代消解的问题，小句复合体结构分析难以解决这种情况，故tag标签为2。

	数量	数据修改前EM	数据集修改后EM	增加ntc后EM
tag=0	759	71.15%	74.31%	76.39%
tag=1	1565	68.37%	72.27%	75.76%
tag=2	895	56.09%	60.45%	64.51%

Table 5: CMRC阅读理解数据集验证集上tag标签EM统计情况

对CMRC2018阅读理解数据集的验证集中tag标签的情况进行统计。结果如表5所示。数据集修改前后，验证集上tag=1的数据的EM都小于tag=0的EM值，tag=2的问题由于需要复杂推理等较难回答，EM值最低，拉低了整体水平。符合之前的认知，跨标点句的远距离问答会给机器阅读理解任务的答案抽取带来困难。Bert基础上的融合模型三经过训练后，tag=0样例的完全匹配率提升2.08%，tag=1的EM提升3.49%，tag=2的EM提升4.06%。小句复合体结构分析不仅对跨标点句问答有提升，对其他类型也有帮助。

Context：2013年，于乐摆脱伤病困扰恢复状态，成为深圳队锋线上颇具冲击力的福将，赛季中期状态火爆屡屡打入关键入球。
Question：2013年于乐状态有了什么变化?　　tag：1
标准答案列表：（1）摆脱伤病困扰恢复状态　　（2）摆脱伤病困扰恢复状态
（3）于乐摆脱伤病困扰恢复状态，成为深圳队锋线上颇具冲击力的福将，赛季中期状态火爆屡屡打入关键入球
不加小句复合体信息的答案预测：赛季中期状态火爆屡屡打入关键入球。
BERT_NTC_add_MRC训练后答案预测：于乐摆脱伤病困扰恢复状态，成为深圳队锋线上颇具冲击力的福将，赛季中期状态火爆屡屡打入关键入球。

Figure 11: 加入小句复合体信息后的答案预测

用实例具体分析基于小句复合体的机器阅读理解模型对跨标点句问答问题的解决情况。如图11所示，为tag标签为1的样例，对模型增加小句复合体结构信息前后，答案预测的结果进行分析。该样例中，前两个标点句单层汇流的关系，第三和第四个标点句共享第二个标点句中的话头“于乐”，故四个标点句处于同一话头结构内。不添加小句复合体结构信息时，预测答案较短导致错误；加入话头话体共享信息后，线索要素与答案要素虽然不处于同一标点句，但是处于同一小句复合体结构，模型预测出了完整答案。证明了小句复合体结构分析对于解决机器阅读理解任务中跨标点句问答的有效性。

5.7 错误样例分析

Question1：布雷西亚足球俱乐部因为签下了谁而得到广泛关注？　F=0 Answer：罗伯特·巴乔 Prediction：前世界足球先生罗伯特·巴乔
Question2：大莱龙铁路有什么重要作用？　F=1 Answer：是横贯山东省北部的铁路干线德龙烟铁路的重要组成部分，构成山东省北部沿海通道，并成为环渤海铁路网的南部干线。 Prediction：是横贯山东省北部的铁路干线德龙烟铁路的重要组成部分，构成山东省北部沿海通道 不加ntc时的Prediction：是横贯山东省北部的铁路干线德龙烟铁路的重要组成部分
Question3：布拉德利在1993-1996期间效力于哪个队？　F=2 Answer：费城76人队 Prediction：费城76人队（1993-1996）、新泽西网队（1996-1997）和达拉斯小牛队（1997-2005）
Context：凡氏下银汉鱼为辐鳍鱼纲银汉鱼目银汉鱼科的一种。本鱼分布印度西太平洋区，包括东非、红海、阿拉伯海、波斯湾、日本南部、台湾、中国沿海、越南、菲律宾、印尼、澳洲、马来西亚、孟加拉湾、印度沿岸、新喀里多尼亚、索罗门群岛等海域。 Question4：凡氏下银汉鱼在中国主要分布在什么地区？　F=3 Answer：台湾、中国沿海 Prediction：印度西太平洋区

Figure 12: 错误样例分析

CMRC2018阅读理解数据集的验证集问题数为3219，全部错误样例数量为878。对全部的错误样例进行分类，其中，F=0：按照正常逻辑，可以算作正常答案，共计438个；F=1：答案过长，预测缺失，共计77；F=2：答案很短，预测过长，共计227；F=3：由于复杂推理等原因，完全预测错误，共计130。

如图12所示的错误样例中，Question1，模型预测的答案多了人物的定语，被判定预测错误。然而，很多标准答案中也包含加定语的答案，所以此类型的错误样例可以算做正确。Question2的结果表明小句复合体结构分析对跨标点句问答有帮助，虽然该样例没有预测出全部的答案，但是与不加小句复合体信息时相比预测的长度更长了。当然，也存在部分样例由于添加了话头话体共享关系信息，预测了多余的答案。如Question3所示，对于包含多个时间的情况，经常预测错误，但是这类问题不是小句复合体结构分析可以解决的。Question4涉及指代消解，而且需要推理属于中国的地区，是小句复合体结构分析解决不了的问题。

总体而言，小句复合体结构分析提供的结构化语义信息可以解决部分跨标点句问答问题，给机器阅读理解任务带来帮助。

6 总结与展望

在中文机器阅读理解任务中，增加外部知识成为了提高模型表现的一种热门方向。小句复合体理论的话头-话体共享关系保证了标点句的语义连贯性，加强了远距离标点句之间的联系，缺失成分的补充同时化简了答案抽取的难度。更换不同的基础模型，模型效果均有不同程度的提升，也体现了小句复合体结构分析任务与机器阅读理解任务融合的有效性。

目前，在小句复合体结构自动分析任务中，实际需要预测的节点的准确率还有提升空间，虽然该任务上F1有94.6%，但预测的位置大多不缺少成分，预测的结果表现高于实际的预测水平，这对于模型融合的效果产生了影响。另外，本文中实验没有采取先对阅读理解数据进行话头-话体结构分析，补全缺失成分后，再作为机器阅读理解任务的输入，微调模型的方法。主要由于机器识别话头-话体的水平有限，且没有对机器阅读理解数据集进行小句复合体结构的人工标注。因此，提高小句复合体结构自动分析任务的准确率；对机器阅读理解数据集进行小句复合体话头-话体结构标注；探索小句复合体理论在其他自然语言处理任务中的应用是下一步的研究工作。

参考文献

蒋玉茹and 宋柔. 2012a. 基于广义话题理论的话题句识别. 中文信息学报, 26(5):114–120.

蒋玉茹and 宋柔. 2012b. Topic structure identification of pclause sequence based on generalized topic theory. Proceedings of the 2012 1st CCF Conference on Natural Language Processing and Chinese Computing., pages 85–96.

蒋玉茹and 宋柔. 2014. 基于细粒度特征的话题句识别方法. 计算机应用, 34(5):1345–1349.

顾迎捷, 桂小林, 李德福, 沈毅, and 廖东. 2020. 基于神经网络的机器阅读理解综述. 软件学报, (7):2095–2126.

宋柔. 2017. 小句复合体的理论研究和应用. http://2011.gdufs.edu.cn/info/1070/2085.htm.

胡紫娟. 2020. 汉语小句复合体话头结构分析. 北京语言大学硕士学位论文.

Jonathan Berant, Vivek Srikumar, P. Chen, A. V. Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In EMNLP.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. arXiv e-prints, page arXiv:1810.07366, October.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. arXiv e-prints, page arXiv:2004.13922, April.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv e-prints, page arXiv:1810.04805, October.

V Garcia-Diaz. 2012. Mctest: towards an improvement of match algorithms for models. Software Iet, 6(2):127–139.

K. M. Hermann, Tomá Koisk, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. 2015. Teaching machines to read and comprehend. In MIT Press.

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: A reading comprehension system. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 325–332, College Park, Maryland, USA, June. Association for Computational Linguistics.

W. Lehnert. 1977. The process of question answering. research report no. 88. Computer Programs, page 293.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv e-prints, page arXiv:1907.11692, July.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

E. Riloff and M. Thelen. 2000. Rule-based question answering system for reading comprehension tests. Proceedings of Workshop on Reading Comprehension Naacl/anlp.

M. Teng, Y. Zhang, Y. Jiang, and Y. Zhang. 2018. Research on construction method of chinese nt clause based on attention-lstm. Springer, Cham.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019a. ERNIE: Enhanced Language Representation with Informative Entities. arXiv e-prints, page arXiv:1905.07129, May.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2019b. Semantics-aware BERT for Language Understanding. arXiv e-prints, page arXiv:1909.02209, September.