硕士学位论文

面向篇章理解的事件时序关系抽取技术研究

RESEARCH ON TEMPORAL RELATION EXTRACTION TOWARDS DISCOURSE-LEVEL COMPREHENSION

栗扬帆

哈尔滨工业大学

2022 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

学术硕士学位论文

面向篇章理解的事件时序关系抽取技术研究

硕士研究生：栗扬帆
导　　师：张宇 教授
申请学位：工学学位硕士
学　　科：计算机科学与技术
所在单位：计算学部
答辩日期：2022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

RESEARCH ON TEMPORAL RELATION EXTRACTION TOWARDS DISCOURSE-LEVEL COMPREHENSION

Candidate: Li Yangfan
Supervisor: Prof. Zhang Yu
Academic Degree Applied for: Master of Engineering
Speciality: Computer Science and Technology
Affiliation: Faculty of Computing
Date of Defence: June, 2022
Degree-Conferring-Institution: Harbin Institute of Technology

摘　要

事件时序关系抽取是自动文本分析中的一个关键问题，能够执行此任务的系统在时间感知摘要、事件时间线构建和事件预测等应用中具有重要的意义。时序关系抽取任务要求模型能够正确的捕获并理解自然语言文本中提及的时间信息。当前阶段自然语言处理任务中广泛使用预训练语言模型作为编码器，然而这些模型所采用的自监督预训练任务往往并不能感知文本中与时间有关的信号。

本文探究面向篇章理解的事件时序关系抽取技术，即在预训练语言模型编码的基础上，挖掘文档中各层级的时序信息，以进行事件间时序关系的预测。本文的主要研究内容包括如下三个部分：

（1）基于句法结构的事件时序关系抽取方法。已有的研究表明两个事件触发词之间的最短依存路径中的词对于事件之间时序关系的正确识别至关重要，但已有的方法往往需要引入对最短依存路径的额外编码模块。本文设计了使用依存句法分析的结果直接限制预训练语言模型中注意力分布的方法，在不引入额外参数的前提下对句子中的词进行软性的筛选。实验结果验证了此方法对于依赖局部句子信息确定时序关系的事件对的有效性。

（2）基于信息增强的事件时序关系抽取方法。当前阶段，事件时序关系抽取任务研究的限制因素之一是数据集的规模。本文首先尝试通过同义词替换的方式扩充已有数据集，但实验结果显示此方法所扩充的数据样例复杂度有限，模型在其上训练的收益不高。之后，本文通过训练额外时序常识知识编码的方式，补充模型在有限的数据集上难以学习到的时序常识知识，实验结果显示了此方法对于依赖世界知识确定时序关系的事件对的有效性。最后，为了更充分地利用每一条训练数据，我们在模型中引入了事件相对时间预测子任务和时序关系对比学习子任务以引导模型更好地捕获文本中的时序信号，实验结果显示这两个子任务都能为事件时序关系抽取模型带来进一步的性能提升。

（3）基于篇章结构的事件时序关系抽取方法。良好的事件时序关系抽取模型需要整合篇章中各层级的信息，同时在预测每一组事件对的时序关系时，统筹考虑关联事件的时序关系情况，做出满足时序一致性要求的预测。为此，本文为每篇文档构建事件、句子、文档片段、文档的层级图以整合文档中各层次的时序信息，并在事件之间连接对应时序关系的边，最后使用图注意力神经网络在此有向带权图之上学习文档中所有事件之间时序关系的相互依赖。

最终的实验表明，相比于只在句子层级进行时序关系抽取的模型和已有的在篇章层级进行时序关系抽取的模型，此方法能够取得一定的性能提升。

关键词：事件时序关系抽取；事件关系识别；预训练语言模型；注意力机制；图注意力神经网络

Abstract

Event temporal relation extraction is a key issue in automatic text analysis. Systems capable of this task are of great significance in applications such as time-aware summarization, event timeline construction, and event prediction. Temporal relation extraction requires the model to correctly capture and understand the temporal information mentioned in natural language texts. Pretrained language models are widely used as encoders in natural language processing currently while the self-supervised pre-training tasks adopted by these models are often unable to perceive time-related signals in text.

Based on pre-training language model, this paper explores temporal relation extraction towards discourse-level comprehension by mining the temporal information of various level in the document. The main research of this paper includes the following three parts:

(1) A method of temporal relation extraction based on syntactic structure. Existing studies have shown that words in the shortest dependency path between triggers of event are crucial for the correct identification of the temporal relation, but existing methods often introduce additional modules to encode the shortest dependency path. This paper designs a method to directly limit the attention distribution in the pretrained language model based on the results of dependency parsing which filters the words in the sentence without additional parameters softly. Experiment verified the effectiveness of this method for event pairs that rely on local sentence information to determine their temporal relation.

(2) A method of temporal relation extraction based on information augmentation. At the current stage, one of the limiting factors for temporal relation extraction is the scale of the dataset. In this paper, we first try to expand existing dataset by synonym replacement, but experiment show that the complexity of samples expanded by this method is limited, and training on them is not profitable. After that, this paper supplements the temporal commonsense knowledge that is hard for the model to learn on limited dataset by training additional temporal commonsense embeddings. Experiments results show the effectiveness of this method for event pairs that rely on world knowledge to

determine their temporal relation. Finally, to make full use of each training sample, we introduced the event relative time prediction subtask and the temporal relation contrastive learning subtask into the model to guide it to better capture the temporal signals. Experiments show that both subtasks can bring further performance improvements to the temporal relation extraction model.

(3) A method of temporal relation extraction based on the discourse structure. A good event temporal relation extraction model needs to integrate the information of each level in the document and consider the temporal relation of related events when predicting the temporal relation of each group of event pairs to meet the requirements of temporal consistency. To this end, this paper constructs a hierarchical graph of events, sentences, document segments, and documents for each document to integrate the temporal information each level. Then connects the edges corresponding to temporal relation between events. Finally, inter-dependencies on the temporal relation among all events in this document are learned on top of this directed weighted graph by graph attention neural network. Comparing to the models that only extract temporal relation at the sentence-level and the existing models that extract temporal relation at the discourse-level, experiments show that this method can achieve a certain performance improvement.

Keywords: temporal relation extraction, event relation identification, pretrained language model, attention mechanism, graph attention neural network

目录

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

随着以 5G 为代表的通信技术的发展，越来越多的线下场景被搬移到线上进行，这些场景产生了海量的信息流。同时，医疗记录、实验日志、新闻报道等事件记录性质的文本材料越来越多地以电子的形式存储和分发。另一方面，移动互联网的不断发展使得越来越多的个体在互联网上投入了越来越多的时间，来自时政、金融、日常生活等各个领域的新消息被人们从不同的视角、不同的方面进行记叙、阐述、评论。如何对这些海量的无结构文本数据进行自动处理，梳理事件发展的脉络，挖掘事物发展的规律，就显得越来越重要。

为了应对上述问题，知识抽取技术应运而生。知识抽取是自然语言处理要解决的核心任务之一，它能从大规模的生语料中自动抽取出所关心领域的知识，包括实体的属性知识和事件的经验知识。事件抽取是知识抽取的一个分支,它关注两个方面的内容,一是从生文本中识别并抽取出事件及其属性，二是识别这些事件之间的关系。典型的事件间关系包括时序关系、因果关系、共指关系和子事件关系等。时序关系抽取即是专注事件之间时序关系的自动识别的事件关系抽取任务。

具体地说，时序关系抽取任务从生文本中识别出一系列所关注的与时间相关的对象，对这些对象两两之间做时间顺序的分类。时序关系抽取任务关注的与时间相关的对象包括两种，一种是动词性质的事件，识别过程的核心是找到这个动词性事件的触发词，它一般是一个动词，以下称之为事件；第二种是时间表达，即 time expression，在文献资料中一般简称为 Timex，它是一系列表述时间节点的文本片段，诸如“今天”、“五年前”、“现在”等。另外，一些研究中还关注事件和 Timex 与文本创建时间（document creation time，简称为 DCT）之间的时序关系。

1.1.2 课题研究的目的和意义

时序关系抽取是自动文本分析中的一个关键问题，能够执行此任务的系统在时间感知摘要、时间信息提取、事件时间线构建和事件预测等领域具有广泛的适用性。对篇章内容的深层次理解不能仅停留在捕获文档中的词共现规律，需要梳理文档中实体之间的联系、事件之间的关联，并能据此进行推理、预测。时序关系抽取任务需要模型能够正确地捕获并理解自然语言文本中提及的时间信息，简单的文本匹配无法解决这一问题，这就对文本中时序相关信息的深层次理解和外部时序常识知识的利用提出了更高的要求。

当前阶段自然语言处理任务中广泛使用预训练语言模型作为编码器，然而，这些模型采用的自监督预训练方式往往并不能感知文本中与时间信息相关的信号，事件时序关系抽取任务能从中获取的收益有限。因此，有必要针对事件时序关系抽取任务设计额外的监督任务和与之匹配的模型结构，并设法整合合适的外部知识,以使模型能够有效地感知文本中与时序相关的信息，提高事件时序关系抽取任务的性能。

另一方面,近年来知识图谱在越来越多的具体业务场景中获得广泛应用，面向各个垂域的知识图谱被学界和业界构建开来。然而，传统的知识图谱以实体为节点进行构建，只能反映实体之间的静态连接关系，不能反映实体变化、事物演化的规律。事件时序关系抽取任务将有助于从大量生文本中自动抽取、梳理事件之间的时间顺序,从而为事物发展演化规律的挖掘提供素材。

1.2 国内外研究现状

事件时序关系抽取技术的相关研究最早可以追溯到上世纪 90 年代，并随着 2003 年 TimeML 标注规范[1]的提出和语义评测 SemEval-2013 中的 TempEval-3 子任务的发布[2]而开始受到越来越多的关注。

同其他自然语言处理任务一样，在早期的研究中，事件时序关系抽取任务主要依赖于基于规则和统计的方法。但由于事件时序关系抽取任务本身的复杂性，在这一阶段中，相关模型所取得性能较为有限。而随着近十年来机器学习技术，特别是深度学习技术的蓬勃发展，自然语言处理领域的诸多任务都取得了令人欣喜的进展，事件时序关系抽取任务也有望取得突破。在过去数年内，事件时序关系抽取任务受到越来越多的研究者关注，一系列相关的数据集和新模型被不断地提出。

1.2.1 事件时序关系抽取任务数据集的构建

时序关系抽取任务面临的第一个重要问题是数据标注的困难性。

早期的 TimeBank 数据集[3]仅标注出了文档中具有显著时序关系的事件对之间的时序关系，这导致在此数据集之上训练的模型具有严重的偏置，模型倾向于消极地不判断输入的事件对的时序关系。为此，文档中任何一组事件对之间的时序关系都需要被标注出来，但是这样的标注过程意味着待标注的事件对的数量相对文档中事件数量的时间复杂度是$o\left(n^2\right)$，并不可行。作为折中，以 TimeBank-Dense[4]（一般简称为 TB-Dense）为代表的一系列数据集标注出了同句和邻句内所有事件对之间的时序关系。然而，这样的标注策略意味着模型只需要挖掘局部的句法特征就足够做出很好的预测，这限制了模型的能力。为此，在 TimeBank-Dense 的基础上，Naik 等人[5]分别通过人工和自动的方式标注了采样出的部分跨句的事件对之间的时序关系，构造了 TDDiscourse-Man 和 TDDiscourse-Auto 数据集，这两个数据集对模型建模篇章层级时序信息的能力提出了更高的要求。

时序数据集标注的另一个严重问题是低 IAA（inter-annotator agreement），标注数据中即存在的严重噪声限制了在其上训练的模型的性能。Ning 等人[6]关注到低 IAA 的主要来源是事件的结束时间点的模糊性以及不在同一叙事轴上的事件的不可比性，他们在 TimeBank-Dense 的基础上设计了仅比较主轴上的事件的开始点的 MATRES 数据集。

此外，部分研究者认为不同类型的事件关系的联合训练能够提高模型对事件及其关系的建模能力，Mostafazadeh 等人[7]和 O’ Gorman 等人[8]分别提出了将时序关系与因果关系联合标注的 CaTeRS 数据集和将时序关系、因果关系、共指关系联合标注的 RED 数据集，但是这两个数据集的规模相对较小。

1.2.2 基于规则和统计的事件时序关系抽取方法

早期时序关系抽取任务的模型主要使用基于规则的方法和基于统计的方法，研究者们开发了一系列能为后续研究提供帮助的模型。典型地，Chambers 等人[9]开发的 CAEVO 系统可以视为早期基于规则和统计方法的集大成者，该系统使用串联的多个基于规则的分类器，组成最终的时序关系分类系统。由于此系统的前几个分类器具有很高的准确率，它们在后续的研究中被广泛用作数据增强的标注工具[10][11]。

1.2.3 挖掘句子级别信息的事件时序关系抽取方法

随着深度学习技术的兴起，研究者们提出了更多建模能力更强、性能更优的时序关系抽取模型。

一些研究者关注将句法结构信息引入时序关系预测中，例如，Cheng 等人[12]抽取句子中位于 SDP（Shortest Dependency Path，即最短依存路径）上的词，使用 Word2Vec[13][14]转化为编码并输入到 BiLSTM[15]中，降低了模型要建模的输入的长度,在占比较高的时序关系类别上取得了明显的性能提升，证明了句法信息对时序关系预测的重要性。Meng 等人[16]则首先使用 LSTM 在事件触发词周围的固定窗口上建模得到事件表达，然后再将两个事件到各自 LCA（Least Common Ancestor，即最近祖先节点）的依存路径分别输入到两个单向 LSTM 上，此外他们还考虑了 token 级别的词性特征。

使用句子的线性结构信息同样是有效的，Ballestero[17]等人直接在 RoBERTa 预训练语言模型[18]的编码之上抽取第一个事件之前、第一个事件、两个事件之间、第二个事件和第二个事件之后共五个片段，分别进行池化聚合之后，再拼接得到两个事件对的组合表达，用以预测两个事件之间的时序关系。

上述方法都未在输入阶段即告知编码器事件的位置，而是在反向传播阶段模型才间接告知编码器事件的位置，这事实上限制了编码器对时序关系所涉及的事件对的编码能力。为此，Wang 等人[19]修改了 BERT 预训练语言模型[20]所使用的注意力机制，在其中添加了各个 token 相对事件触发词的距离的编码，强化了编码器对事件的感知能力，获得了明显的的性能提升。

除了事件表达，在深度学习模型中获取好的 Timex 的表达同样重要。Goyol 等人[10]注意到基于规则的方法已经能够以很高的准确率区分 Timex 之间的时序关系，甚至比通常的时序关系分类范式训练得到的深度学习方法更好，为了让深度学习模型同样获得优秀的区分 Timex 的时序关系的能力，他们采用基于规则的方法构建大量的时序关系已知的 Timex 对，再在其上训练得到基于 BiLSTM 的字符级别 Timex 编码器，在时序关系抽取模型的编码结果之上拼接使用此编码器得到的 Timex 的编码，获取了更好的事件间时序关系抽取的性能。

1.2.4 应用数据增强技术的事件时序关系抽取方法

由于时序关系抽取任务数据的稀缺性，同很多其他自然语言处理任务一样，借助额外的外部语料和外部知识也是当前时序关系抽取任务的一个重要研究方向。

针对时序关系抽取任务所提出的各个数据集采用的标注范式各不相同，一部分采用稠密标注，即标注出所有事件对之间的时序关系；一部分则采用稀疏标注，仅标注出那些时序关系显著的事件之间的时序关系，为此，Ning等人[21]的工作探讨了如何在模型训练时同时使用这两种不同标注范式的数据集。

由于时序关系抽取任务的数据集依然有限，并且数据集的规模普遍不大，部分研究者进行了将时序关系抽取任务与其他相关任务进行联合训练的一系列研究。Han 等人[22]设计了将事件抽取任务与时序关系抽取任务进行联合训练的方法，使得事件的预测与时序关系的预测共享同一个编码表示，此方法避免了传统的 pipeline 做法的错误累积，联合的两个任务能互相纠正对方的错误编码，获得了一定的性能提升。时序关系抽取任务与因果关系抽取任务天然地存在联系，因为原因总是发生在结果之前，Ning 等人[23]将两者进行联合训练，证明了两个任务能够帮助彼此更好地完成关系类别的预测。Wang等人[24]则关注到子事件关系预测任务中，父子事件之间的时序关系相互之间的约束性，他们将子事件关系抽取任务与时序关系抽取任务进行联合训练，也取得了不错的效果。

如何高效低成本地获取时序关系标注数据以扩充标注语料也受到研究人员的关注。Ning 等人[25]设计了先通过语义角色标注器在生语料上标注出事件，然后应用使用基于规则抽取出的特征在现有数据集 TimeBank-Dense[4]上训练好的模型在事件之上标注时序关系得到大规模标注数据，进而在其上训练得到事件时间感知的预训练语言模型，可以预见其将在时间知识相关的任务上发挥重要作用。Zhao等人[11]则使用Distant Supervision的策略，使用CAEVO[8]的前两个准确率最高的基于规则的分类器，从大量生语料中构建出大量的有监督时序关系数据，考虑到深度学习模型的强大学习能力，他们还使用了将规则标注器所参考的直接证据进行 mask 的策略避免模型直接学习到标注规则，在他们的数据上执行训练的模型在 zero-shot、few-shot 和 transfer 设置下，都取得了让人印象深刻的结果。

迁移学习是利用相关语料的重要方法，Ballestero 等人[17]设计了通过 Schedule Learning[26]的策略，分别将事件关系数据集 ACE relation extraction[27]、TimeBank 数据集[3]和在生语料上使用训练得到的最优模型标注的 silver 数据集与 MATRES 数据集[6]进行联合训练，在 MATRES 数据集上取得了明显的性能提升，并且其结果显示使用 silver 数据集的方法依然有通过增加迭代轮次来进一步提高性能的潜力。

Ning 等人[28]还留意到在真实文本中，两个动词对应的事件所发生的时序关系是存在一定的偏置的，例如，我们总是先“打开电脑”，然后开始“编写代码”，相反的情况很少发生，为了利用这样的外部先验常识知识，他们统计了动词对之间呈现不同时序关系的频率，构建了 TemProb 知识库。对此知识库的不同应用方式都能带来时序关系抽取任务性能的一定提升，例如，他们的另一项工作[29]使用孪生网络在 TemProb 上训练额外的事件编码，将之拼接到 LSTM 所建模的句子编码之上，带来了明显的性能改进。

1.2.5 挖掘篇章级别信息的事件时序关系抽取方法

时序关系抽取任务的一个难点还在于时序关系天然的拓扑结构。给定一篇文档中所有事件对的时序关系之后，以事件作为节点，以事件之间的关系作为有向图中带标签的边，我们将得到一张时序关系图。时序关系图必须满足如下的三条性质：

1、唯一性：对于任意一组事件对 A 和 B，事件 A 相对于事件 B 只能有一种时序关系。也就是说在时序关系图中，A 节点只能有一条指向 B 节点的边。

2、对称性：对于任意一组事件对 A 和 B，事件 A 相对于事件 B 与事件 B 相对于事件 A 的关系是相关的。例如，若事件 A 相对于事件 B 是 before 关系，那么事件 B 相对于事件 A 必须是 after 关系。也就是说在时序关系图中，A 节点指向 B 节点的边和 B 节点指向 A 节点的边的标签是对称的。

3、传递性：对于三个事件 A、B、C，事件 A 与事件 C 之间的时序关系事实上是被事件 A 与事件 B 和事件 A 与事件 C 之间的时序关系限定的。例如，若事件 A 相对于事件 B 和事件 B 相对于事件 C 都是 before 关系，那么事件 A 相对于事件 C 的关系也必然是 before。也就是说在时序关系图中，A 节点指向 C 节点的边的标签可以借由 A 节点指向 B 节点的边的标签和 B 节点指向 C 节点的边的标签来推定。

上述的三条限制我们统称为时序关系的一致性限制。一致性限制意味着对当前事件对的预测不只需要考虑事件对所在句子层级的局部信息，还需要

考虑文档层级的全局信息。而同时编码整个文档的全部信息又容易受到自然语言处理任务中广泛存在的长程依赖问题的影响，Wen 等研究者[30]采用了 Longformer[31]代替常用的 BERT 预训练语言模型，以编码事件对周围更长的文本片段，从而在时序相关任务中尽可能捕获全局信息。

为了满足时序关系的一致性限制，一个符合直觉的方法是将上述的三条限制分别对应到 ILP（Integer Linear Programming，整数线性规划）中的若干个不等式限制条件，Han 等人[32]即采用此方法，结合 SSVM[33]（Structured Support Vector Machine），进行满足一致性要求的时序关系推理预测。然而，此方法事实上依赖于在模型预测结果之上执行 ILP 的求解，来避免不符合一致性限制的时序关系，深度学习模型依然只能获取局部信息，也并没有感知到明确的时序关系一致性要求，这限制了深度学习模型的建模能力。Wang 等人[24]的研究中设计了分别对应于时序关系的对称性、传递性的 loss 函数项，以此让深度学习模型感知时序关系的一致性要求。另一部分研究者关注在根据模型预测的时序关系标签构建时序图的过程中进行剪枝[34][35]，以得到满足时序一致性要求的时序关系图。考虑到 Timex 之间的时序关系可以使用基于规则的方式得到准确率很高的时序关系分类结果，Meng 等人[16]首先将 Timex 之间的标签加入时序图中，然后对不同的标签类别分别逐个添加边到时序图中，遇到冲突时，则贪心地剔除那些置信度更低的预测标签。

进行满足时序关系一致性限制的时序关系预测的就要求模型必须整合输入文档在篇章层级的全局信息，而非只感知句子层级的局部信息。为了融入篇章层级的全局信息，国内王俊等人[36]在文档内的所有事件对之间应用注意力机制，得到上下文信息增强的事件表达，实验结果显示了此方法的明显性能提升。考虑到事件时序关系的拓扑结构，部分研究者尝试在模型输入中即反映这种结构性。Cheng 等人[37]使用序列结构，根据事件在文本中的出现顺序构建出一条事件链，并在其中插入 DCT，使用 GRU[38]对此事件链进行建模，由此预测这条链上任意的事件对之间的时序关系。在与时序关系抽取任务紧密相关的事件时间抽取任务上，Wen 等人[30]还使用图结构来传播文本信息进而捕获文档中的全局信息，具体地，他们首先依据事件时序关系和事件共享参数关系构建出两张图，然后使用 GAT（Graph Attention Network）[39]在其上设计有序的信息传播流程以获取包含全局事件信息的编码，在其研究的任务上获得了明显的性能提升。此外，Meng 等人[40]从人的标注过程中获得启发，通过在判断逐个输入的句子中事件对的时序关系时应用简化的神经图灵机[41]的读写机制，来模仿人在判断事件时序关系时，会不断回看文档的上文信息

的行为，来帮助对当前事件对时序关系的判断，同时当前事件对的判断结果也有可能修正此前由于未看到后续内容而未掌握全部信息导致的错误判断。

1.3 主要研究内容

我们的研究将以预训练语言模型为基础，结合外部时序知识库和时序辅助任务，探索时序关系抽取任务中事件的高质量表达方式，并使用句子的句法结构先验挖掘句子层级的局部时序信息，同时在篇章层级构建文档的层级图结构，使用图神经网络挖掘篇章层级的全局时序信息。

本文的研究内容主要分为以下三个部分：

（1）基于句法结构的事件时序关系抽取方法。本部分依据句法结构挖掘句子层级的局部时序信息。已有的研究表明，位于最短依存路径上的词对于判断事件之间的相互关系最为重要。有鉴于此，我们首先使用循环神经网络分别在预训练语言模型编码后的整个句子和最短依存路径上进行额外建模，验证事件时序关系抽取任务中最短依存路径上的词的重要性。之后依据相对最短依存路径的距离，给句子中的词赋予不同的重要性程度，通过设计损失函数使得预训练语言模型将更多的注意力分布到重要性程度更高的词之上，强化模型的句法结构感知能力，从而使得预训练语言模型以事件为中心更好地整合句子层级的局部时序信息。

（2）基于信息增强的事件时序关系抽取方法。本部分设法更充分地利用每一条样例的时序标签所携带的信息。现有的时序关系抽取数据集的规模往往有限，我们设计了一系列方案来应对此问题。首先，我们设计了基于同义词替换的数据增强方法来扩充数据集。其次，我们在时序常识知识库之上训练时序常识知识编码并拼接到原有的事件编码之上，使得我们获取的事件表达具有时序常识知识信息。最后，为了更充分地利用每一条训练样例，使得模型更充分地挖掘文本中的时序信号，我们分别引入了事件相对时间预测子任务和时序关系对比学习子任务来提供额外的监督信号，提高事件时序关系抽取模型的性能。

（3）基于篇章结构的事件时序关系抽取方法。本部分在上述研究的基础上，进行篇章层级的时序关系抽取研究。对于判断事件时序关系至关重要的信息可能并不与事件出现在同一句中，同时，模型必须统筹篇章中的所有事件对之间的时序关系，才能给出满足时序一致性要求的全局预测。为此，在使用预训练语言模型对文档进行编码之后，我们构建文档的层级结构图，并在文档中的事件节点之间连接对应时序关系的时序边，最后使用图神经网络

在此有向带权图之上进行建模，以整合文档的全局信息，并在预测每一条时序关系时考虑关联事件的时序关系，进一步提高事件时序关系抽取模型的性能。

1.4 本文章节安排

本文分别从基于句法结构的事件时序关系抽取方法、基于信息增强的事件时序关系抽取方法、基于篇章结构的事件时序关系抽取方法三个方面进行论述、研究和分析。

本文的主要内容及其结构安排如下：

第一章为本文的绪论。我们首先介绍课题的背景、研究目的及意义，阐述课题的研究价值和应用价值。随后对事件时序关系抽取技术的当前研究进展进行归类、总结和分析。之后以当下事件时序关系抽取任务的研究为基础，确定本文的主要研究内容。

第二章主要介绍基于句法结构的事件时序关系抽取方法。首先介绍本研究中将使用的预训练语言模型，然后介绍依据依存句法分析结果强化预训练语言模型的句子层级时序信息挖掘能力的方法并进行实验验证，最后对实验结果进行总结和分析。

第三章主要介绍基于信息增强的事件时序关系抽取方法。首先介绍应用于事件时序关系抽取任务的基于同义词替换的数据增强方法，然后介绍通过训练时序常识知识编码将时序常识知识融入事件向量表达的方法，随后介绍事件相对时间预测子任务和时序关系对比学习子任务两个辅助任务，最后设计实验并通过实验结果进行验证。

第四章主要介绍基于篇章结构的事件时序关系抽取方法。首先介绍图注意力神经网络，然后介绍将待进行事件时序关系抽取的文档映射为篇章层级时序图的方法，随后介绍在此有向带权图之上应用图注意力神经网络进行篇章层级事件时序关系抽取的模型，最后通过实验验证此篇章层级事件时序关系抽取模型的有效性。

整篇论文的结构框架如图 1-1 所示。

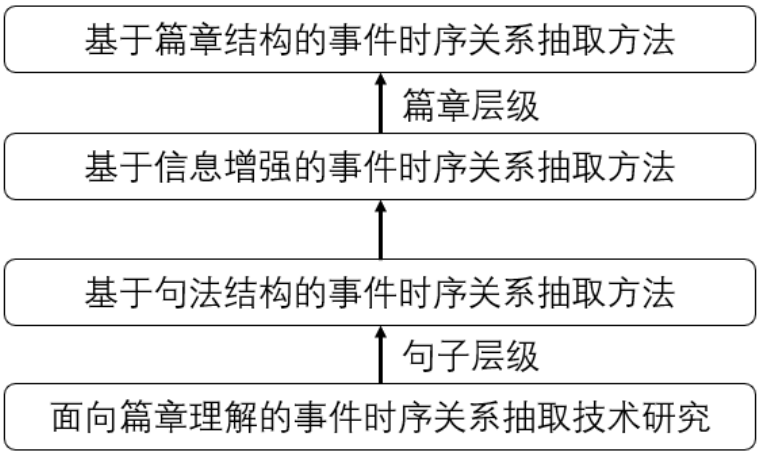

图 1-1 论文结构图示

第 2 章 基于句法结构的事件时序关系抽取方法

2.1 引言

在当前阶段的自然语言处理研究中，广泛使用预训练语言模型作为编码器，以获取字、词、句、文档等不同层级的文本元素的向量表达。通过在大规模生文本上设计合适的自监督训练任务，预训练语言模型能够学习到通用的各层次的语言知识。大量的研究和实验验证了预训练语言模型所学习到的这些知识在下游任务中的有效性。在我们的研究中，同样使用预训练语言模型作为编码器。

另一方面，在预训练语言模型开始被广泛应用之前，一些研究表明了两个事件触发词之间的最短依存路径中的词对于事件间关系的正确识别至关重要，后续的研究也表明句法信息对于句子级别的事件时序关系抽取任务非常有效。这些研究中的典型方法是在使用传统词向量将输入文本向量化之后，借由依存句法分析筛选出最短依存路径上的词，使用循环神经网络对这些词进行额外的编码。我们首先采用与之类似的方法，但将输入循环神经网络的传统词向量替换为预训练语言模型编码得到的向量，以此验证在使用预训练语言模型作为编码器的前提下，最短路径上的词对于判断事件之间的时序关系依然更为重要。

上述的方法结构简单而有效，但是一方面由于使用了另外的循环神经网络对最短依存路径进行编码而引入了额外的参数，另一方面预训练语言模型的参数经过了自监督预训练，而循环神经网络的参数则由随机初始化得到，两者之间训练过程中的异步可能限制了模型整体的学习能力。更重要的是，上述方法依据是否位于最短依存路径上对句子中的词进行硬性的筛选，可能剔除了某些不位于最短依存路径上但是同样包含时序信息的词。为此，我们设计了使用依存句法分析的结果直接限制预训练语言模型中的注意力分布的方法，在不引入额外参数的情况下对句子中的词进行软性的筛选，使得预训练语言模型能够直接感知句子中的句法结构信息并在句法结构的引导下以事件为中心整合句子中的时序信息，实验结果验证了此方法在事件时序关系抽取任务中的有效性。

本章的内容安排如下：第 2.1 节为本引言；第 2.2 节介绍本研究中所采用的预训练语言模型；第 2.3 节介绍本研究中所采用的循环神经网络；第 2.4

节介绍基于句法结构的事件时序关系抽取方法；第 2.5 节介绍本研究所采用的数据集，并给出本章所设计的方法的实验设置情况，同时对实验结果进行分析；第 2.6 节对本章内容进行小结。

2.2 预训练语言模型

预训练语言模型是仅几年来在自然语言处理领域乃至整个深度学习领域研究者们所取得最令人欣喜的成果之一，其被广泛验证的有效性不仅大幅度提高了几乎所有自然语言处理任务的性能指标，而且其所代表的“预训练-微调”范式启发了其他领域的研究[42]。此外，基于预训练语言模型获取到的高质量的文本表示，使得语言与图像相结合的任务，例如视觉常识知识问答[43]、基于文本的图像生成[44][45]等任务的研究成为可能。

早期的统计自然语言处理研究阶段，语言模型专指 n 元语言模型。随着深度学习被引入自然语言处理研究中，以 Word2Vec[13][14]、GloVe[46]为代表的通过各种方式训练得到的静态词向量，可以视为是早期广义的语言模型。随后，借助于计算机计算能力的显著提高，更多更为复杂、参数量更大的神经网络结构被引入到语言模型的预训练过程中。这其中两个代表性的工作分别是基于双向 LSTM 结构进行预训练的 ELMo 模型[47]和基于 Transformer 结构[48]进行预训练的 BERT 模型[20]，其中后者在当前的预训练语言模型结构中占据主流。这些被充分预训练的语言模型被广泛应用于各种下游的自然语言处理任务中。

2.2.1 Transformer 结构

在 Transformer 结构在以自然语言处理为代表的序列建模任务中成为主流之前，研究者们广泛采用以 GRU[38]、LSTM[15]为代表的循环神经网络。然而这些循环神经网络存在一系列固有的缺陷，首先，它们受到长程依赖问题的困扰，其次，循环神经网络所采用的逐个时刻串行计算的策略天然地不利于进行并行计算,这显著限制了采用循环神经网络的深度学习模型的计算速度。

Transformer 结构于 2017 年被 Ashish 等人[48]提出，并首先在机器翻译任务出显示出其强大的序列建模能力，并迅速被其他自然语言任务，特别是绝大多数语言模型预训练任务所采用。Transformer 结构不仅具有强大的序列建模能力，同时其采用的注意力机制也具有更好的可解释性，如图 2-1 所

示，经过充分训练之后，句子中的词的注意力权重分布被对齐到与此词语义紧密相关的词之上。

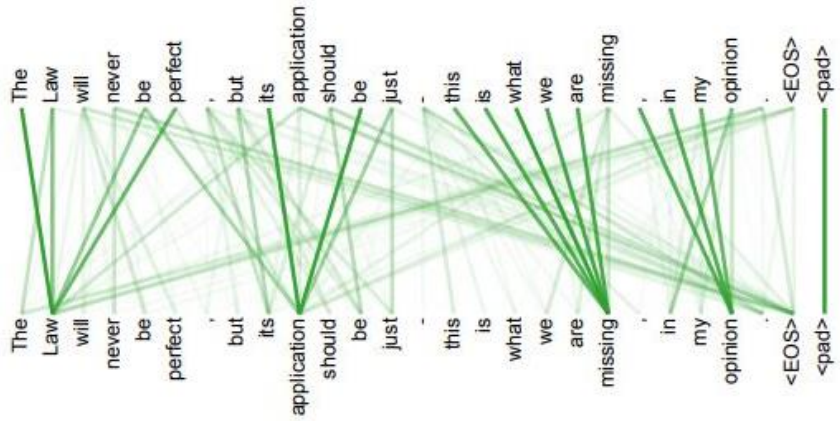

图 2-1 Transformer 结构获取的注意力分布图示[18]

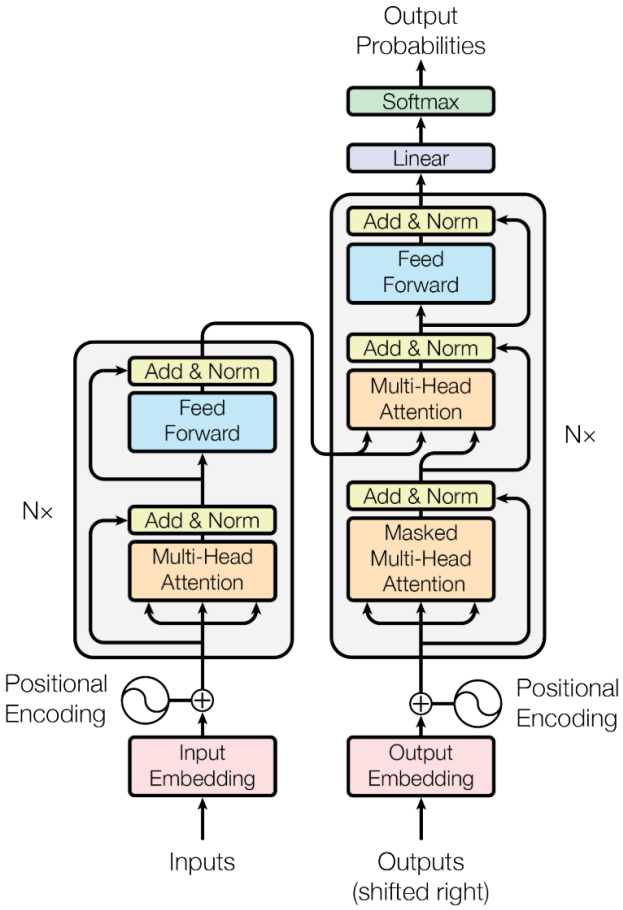

图 2-2 Transformer 模型结构图示[48]

如图 2-2 所示，一个典型的 Transformer 网络结构由若干个串联的 Transformer 层组成，具体所需的层数需要通过对任务所需的执行效率和任务本身的建模复杂程度进行综合折中确定。所有这些 Transformer 层往往采用相同的网络结构，包括一个多头注意力子层、一个前馈神经网络子层以及它们各自对应的残差连接结构和层正则化结构。

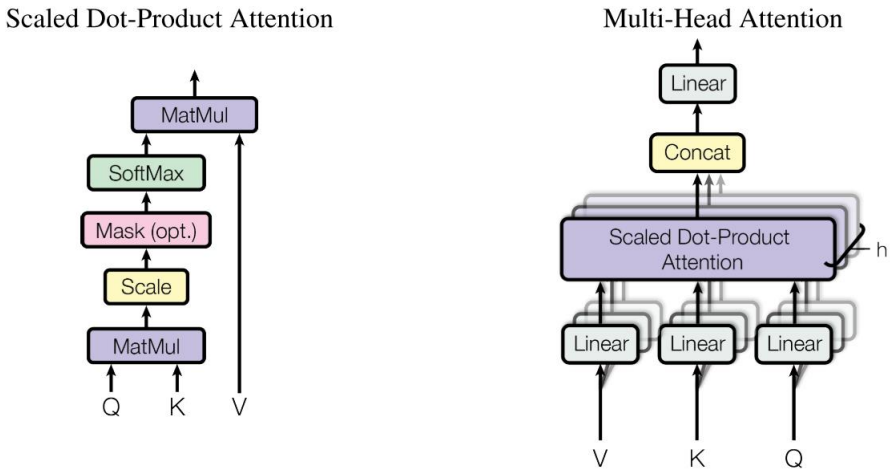

图 2-3 多头自注意力机制图示[48]

Transformer 结构强大的序列建模能力的核心在于其采用的多头自注意力机制，多头自注意力机制的结构如图 2-3 所示。对于原始的输入序列，Transformer 结构首先通过在此输入序列上的三个不同的线性变化，得到查询序列 Q、索引序列 K 和键值序列 V。之后对于序列中的每个元素，通过查询序列 Q 和索引序列 K 中每个位置的元素之间的点积得到两者之间的注意力得分，并以此注意力得分作为权值，通过对键值序列 V 进行加权求和，然后经过一次线性变换，得到新的输入序列表达。这个过程中计算得到的所有注意力得分构成一个矩阵，称之为 Attention map。此矩阵中的第 i 行第 j 列的元素表征了输入句子中的第 j 个词相对输入句子中的第 i 个词的注意力得分，第 i 行的全部元素就表征了句子中的其他词相对于第 i 个词的重要性程度。为了提高 Transformer 结构每层的表征能力，在每层内可以使用不同的线性变换矩阵得到不同的查询序列 Q、索引序列 K 和键值序列 V，从而得到多个注意力头对应的多个输入序列的新表达，再通过拼接得到最终的序列编码。多头自注意力机制的具体计算过程如公式（2-1）、（2-2）、（2-3）所示。

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \qquad (2\text{-}1)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{2-2}$$

$$MultiHead\left(Q,K,V\right) = Concat\left(head_1, head_2, ..., head_h\right)W^O \tag{2-3}$$

其中，h为注意力的头数，W_i^Q、W_i^K、W_i^V是第i个注意力头中Q、K、V各自对应的可训练权重变换矩阵，W^O是用于最终的输出序列的权重变换矩阵，$Concat$表示拼接操作。

Transformer 结构通过其前馈神经网络层引入非线性。具体地，其前馈神经网络层包含两次线性变换以及介于两次线性变换之间的一个进行非线性变换的激活函数。Transformer 结构的不同应用中所采用的激活函数各有不同，当采用 ReLU 激活函数时，其前馈神经网络子层如公式（2-4）所示。

$$FFN\left(x\right) = max\left(0, xW_1 + b_1\right)W_2 + b_2 \tag{2-4}$$

其中，W_1、b_1和W_2、b_2分别是两次线性变换的权重矩阵和偏置。

最后，上述的多头自注意力子层和前馈神经网络子层分别被残差连接和层正则化包围。残差连接允许梯度更为方便地从神经网络的输出端传递到底层，它对于训练多层的神经网络至关重要；而层正则化则提高了模型的泛化性能，避免模型过拟合于训练数据。残差连接和层正则化可以被形式化地表达如公式（2-5）所示。

$$SubLayerOutput = LayerNorm\left(x + \left(SubLayer\left(x\right)\right)\right) \tag{2-5}$$

其中，$SubLayer$表示多头自注意力子层或前馈神经网络子层，$LayerNorm$表示层正则化操作。

上述部分介绍了 Transformer 结构中的编码器，事实上，如图 2-2 所示，Transformer 结构中的解码器与编码器有所不同。不同于编码器仅包含一个多头注意力子层，解码器包含两个多头注意力子层。其中第一个多头注意力子层采用了掩码的机制，它通过将注意力矩阵 Attention map 乘以一个左下三角的 MASK 矩阵，以使得每个输入元素的注意力仅分布于在其前面的那些元素之上，从而避免每个元素能够直接看到其后面的元素，导致在自回归解码时标签泄漏。第二个注意力子层并非计算自注意力，而是计算输入序列相对输出序列的注意力分布。

通过以上的介绍可知，由于所采用的多头注意力子层和前馈神经网络子层，Transformer 结构事实上丢失了序列本身的位置信息。为此，如图 2-2 所

示，Transformer 结构中引入了位置编码，在序列被输入 Transformer 结构之前，首先与位置编码进行求和，以此引入位置信息。以 Ashish 等人[48]所实现的基于正余弦函数的绝对位置编码方案为例，对于序列中第 pos 个位置的向量中的第 $2i$ 维和第 $2i+1$ 维，其位置编码分别如公式（2-6）和公式（2-7）所示。

$$PE(pos,2i)=sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{2-6}$$

$$PE(pos,2i+1)=cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{2-7}$$

2.2.2 RoBERTa 预训练语言模型

正如 RoBERTa 预训练语言模型[18]（A Robustly Optimized BERT Pretraining Approach）的名字所示，其由 BERT 预训练语言模型（Bidirectional Encoder Representation from Transformers）进行强化训练而来。我们首先介绍 BERT 预训练语言模型的模型结构及其所采用的预训练任务，随后再介绍 RoBERTa 预训练语言模型在其基础上所做的改进。

BERT 预训练语言模型是近年来不断涌现的各种预训练语言模型中最具有代表性的一个。如图 2-4 所示，BERT 预训练语言模型使用 Transformer 结构作为其网络结构。事实上，BERT 只使用如图 2-2 所示的 Transformer 结构左侧的编码器部分，这使得 BERT 预训练语言模型本质上是一个自编码语言模型（Auto-Encoder Language Model），其对输入文本进行双向建模的能力使得其更擅长自然语言理解任务（Natural Language Understanding）。

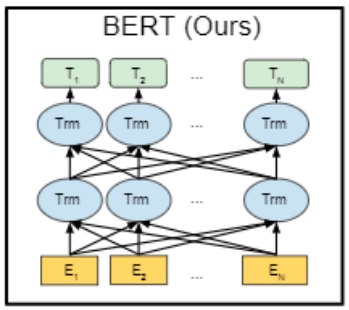

图 2-4 BERT 预训练语言模型结构图示[18]

BERT 预训练语言模型包括输入层、编码层和输出层三个部分。其中，输入层对词向量、段向量和位置向量进行加和，这之中，词向量是用于编码词的向量表征，段向量用于区分词所在的句子，位置向量用于编码词的位置信息。不同于 Ashish 等人[48]采用的基于公式的位置编码方案，BERT 通过随机初始化并在预训练过程中同步训练得到位置编码。BERT 的编码层采用图 2-2 所示的 Transformer 结构中位于左侧的编码器部分。BERT 的输出层给出最终编码得到的句子向量表达，可以应用于各种形式的下游自然语言处理任务。

BERT 的预训练任务包括两个，分别是掩码预测 MLM（Masked Language Model）任务和 NSP（Next Sentence Prediction）任务。在预训练过程中，BERT 的输入是带有[MASK]标签的两个句子，这两个句子来自同一篇文档中连续的两句或来自两篇不同的文档之中，[MASK]标签通过将原本句子中的词进行替换得到。具体地，句子中 15%的词被选中作为预训练目标，这之中，80%的词被替换为[MASK]标签，10%的词被替换为其他的词，10%的词则不进行替换保留原样。掩码预测任务即预测这些被选中作为预训练目标的词原本是哪一个词，而 NSP 任务则预测这两个句子是来自同一篇文档还是来自不同的文档。

相对 BERT 预训练语言模型，RoBERTa 预训练语言模型采用与之完全相同的模型结构，但在预训练过程中做了如下几方面改进。其一，RoBERTa 采用了更大规模的无监督生语料和更大的 batch size；其二，为了更充分地利用训练语料，RoBERTa 的[MASK]标签在训练过程中动态地生成；其三，RoBERTa 预训练时输入的序列长度更长；其四，RoBERTa 删除了 NSP 预训练任务，实验结果显示，删除 NSP 任务之后，模型反而能在大多数下游任务中取得一定程度的性能提升。

一般认为，RoBERTa 预训练语言模型的预训练过程更为充分，在我们的后续研究中，我们将使用 base 版本的 RoBERTa 预训练语言模型开展研究。

2.3 LSTM 循环神经网络

LSTM（Long Short-Term Memory）是循环神经网络的一个重要变体，其结构如图 2-5 所示，它通过引入一个长期记忆向量的方式，能够较好地应对一般的循环神经网络的梯度爆炸和梯度消失的问题。

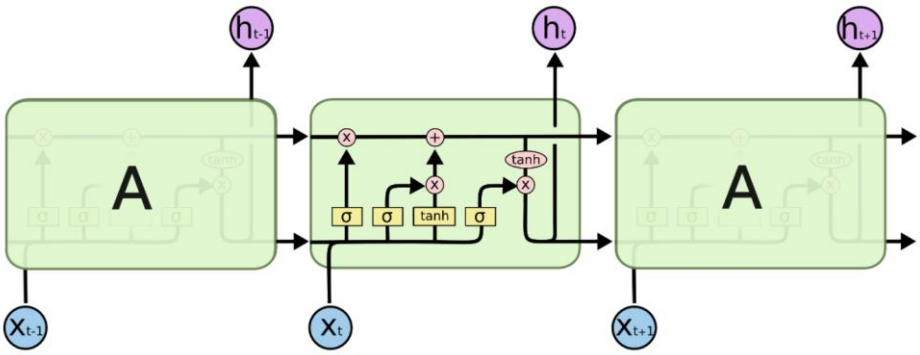

图 2-5 LSTM 模型结构图示

BiLSTM 结构中引入了一个充当长期记忆的内部状态向量 c_t。为了维护词向量随时间的更新，还引入了三个门来控制信息的筛选和流动，分别是输入门 i_t、遗忘门 f_t和输出门 o_t，他们的计算方式分别如公式（2-8）、（2-9）和（2-10）所示。

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{2-8}$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{2-9}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{2-10}$$

其中，h_t是第$t-1$时刻的隐藏状态向量，x_t是第t时刻的输入，W、U和b分别是三个门各自对应的可训练的变换矩阵和偏置，σ是$sigmoid$函数。

内部状态向量c_t的维护方式如公式（2-11）和（2-12）所示，其中$\odot$表示向量逐点积运算。

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{2-11}$$

$$\tilde{c}_t = tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{2-12}$$

t时刻的输出的计算方式则如公式（2-13）所示。

$$h_t = o_t \odot tanh\left(c_t\right) \tag{2-13}$$

2.4 在事件时序关系抽取任务中引入句法结构信息

2.4.1 事件表达

获取好的事件表达是进行事件时序关系抽取的第一步。遵照现阶段自然语言处理任务的研究范式，我们使用预训练语言模型来获取事件的向量化表达。具体地，针对待识别时序关系的两个事件 A、B 及其所在的文本片段 S，我们构造如公式（2-14）所示的 RoBERTa 预训练语言模型的输入，以获取文本片段 S 的编码结果，其中[CLS]和[SEP]是预训练模型中的特殊 token 标记。当事件 A、B 位于相同的句子时，文本片段 S 只包含此句子；当事件 A、B 位于不同的句子时，文本片段 S 为这两个句子的拼接。

$$F_S = RoBERTa\left(\left[[CLS],S,[SEP]\right]\right) \tag{2-14}$$

由于预训练语言模型编码的最小粒度为子词，事件 A、B 将被映射到数量不定的子词之上，我们需要将这些子词的向量进行聚合来得到事件 A、B 的表达。由于对应的子词数量不定，不能使用卷积的方式进行。以事件 A 为例，假设其事件触发词对应句子 S 中从第 i 到第 j 个子词，我们将对比如下的几种聚合方式：

1）如公式（2-15）所示，贪心地取第一个子词的向量编码来作为事件 A 的表达。

$$F_A = S_{[i]} \tag{2-15}$$

2）如公式（2-16）所示，逐维度取这些子词向量编码的最大值来作为事件 A 的表达。

$$F_A = max\left(S_{[i:j]}\right) \tag{2-16}$$

3）如公式（2-17）所示，逐维度取这些子词向量编码的平均值来作为事件 A 的表达。

$$F_A = average\left(S_{[i:j]}\right) \tag{2-17}$$

4）如公式（2-18）、（2-19）所示，应用注意力的机制，定义一个虚拟的查询向量 h，其与这些子词向量编码的点积经过 $softmax$ 函数概率化之后，得到各自的权值 a_x，之后使用它们的加权和 F_A 来作为事件 A 的向量表达。

$$a_x = \frac{exp\left(\frac{S_x^T h}{\sqrt{d_h}}\right)}{\sum_{l=i}^{j} exp\left(\frac{S_l^T h}{\sqrt{d_h}}\right)} \tag{2-18}$$

$$F_A = \sum_{l=i}^{j} a_l S_l \tag{2-19}$$

如公式（2-20）所示，得到事件 A、B 的向量化表达 F_A 和 F_B 之后，我们将这两个向量以及它们的和与逐点积进行拼接，输入到一个多层感知机（即公式（2-20）中的 MLP）之中进行事件时序类别的预测。

$$P_{AB}^r = MLP\left(cat\left(F_A, F_B, F_A + F_B, F_A \odot F_B\right)\right) \tag{2-20}$$

基于事件表达的事件时序关系抽取模型的结构如图 2-6 所示。

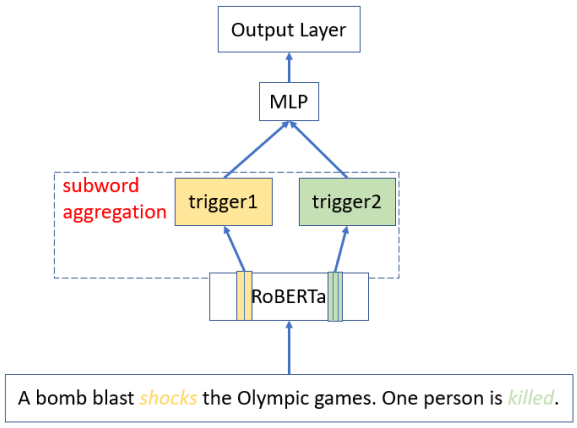

图 2-6 基于事件表达的事件时序关系抽取模型的结构图

我们使用如公式（2-21）所示的交叉熵损失函数和反向传播算法对模型中的参数进行优化。

$$L\left(R \middle| A, B, \theta\right) = -log \frac{P_{AB}^R}{\sum_{r \in \{Relation\_set\}} P_{AB}^r} \tag{2-21}$$

其中 $Relation\_set$ 为数据集中定义的事件时序关系类别集合，R 是事件 A 与事件 B 之间真实的时序类别标签。

2.4.2 使用循环神经网络进行额外句法信息编码

容易发现，上一节中实现的方法事实上只有 RoBERTa 预训练语言模型能看到事件所在的上下文，后续模块的输入只有事件 A、B 的编码，而丢失了上下文信息。我们希望后续模块也可以看到事件 A、B 所处的句子中那些比较重要的上下文信息。同时，已有的大量研究表明，在关系抽取任务中，两个关系对象的最短依存路径上的词对于确定两者之间的关系最为重要，因此我们使用句法信息来增强事件时序关系抽取模型对句子级别局部信息的建模能力。

具体地，我们依据句子依存句法分析的结果，抽取出事件 A、B 的事件触发词在依存句法树上的最短依存路径上的那些词，将这些词的编码构成的向量序列使用 BiLSTM 进行建模。当事件 A、B 位于不同的句子时，我们假设这两个句子的根节点连接到一个虚拟的公共根节点上，从而也可以抽取出最短依存路径。

我们从文本片段 S 的编码结果 F_S 中，采用上一节中描述的方法应用注意力机制进行聚合得到最短依存路径上的每个词对应的向量编码，输入到单层的 BiLSTM 中，得到两个事件触发词的前向和后向各两个向量表达，采用如公式（2-20）所描述的方式输入到多层感知机中进行分类。此方法对应的模型结构图如图 2-7 所示，我们称之为 SDP-BiLSTM。此外，为了验证最短依存路径上的词的重要性，作为对比，我们还设计了将整个句子中的所有词都输入到 BiLSTM 中的模型，我们称之为 Sent-BiLSTM。

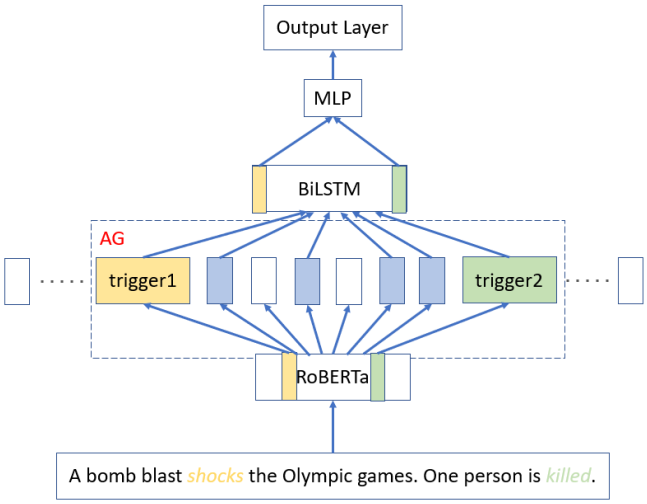

图 2-7SDP-BiLSTM 模型结构图示

2.4.3 使用句法结构信息限制注意力分布

考虑到上小节中依据是否位于最短依存路径对句子中的词进行过滤的方法本质上是一种硬编码，即规则地过滤掉不在最短依存路径上的词，而保留位于最短依存路径上的那些词，这有可能遗漏某些对于事件时序关系抽取任务至关重要的信息。

以此句子为例：At four o'clock I was playing table tennis, at five o'clock I was having dinner，判断“playing table tennis”和“having dinner”两个事件的时序关系最关键的证据是来自其各自发生的时间“At four o'clock”和“at five o'clock”，但是这两个文本片段中的词都不位于最短依存路径上。上小节中实现的 SDP-BiLSTM 模型将过滤掉这些信息，这可能将导致模型不能正确地进行此类样例的时序类别分类。

此外，由于 SDP-BiLSTM 模型的所使用的 BiLSTM 网络结构，引入了额外的待训练参数，而这些待训练参数在模型微调阶段只能通过随机的方式来进行初始化，而 RoBERTa 预训练语言模型中的参数则经过了预训练过程。这两类参数之间异步的训练过程可能限制了模型的学习能力，使得模型能够取得的性能有限。

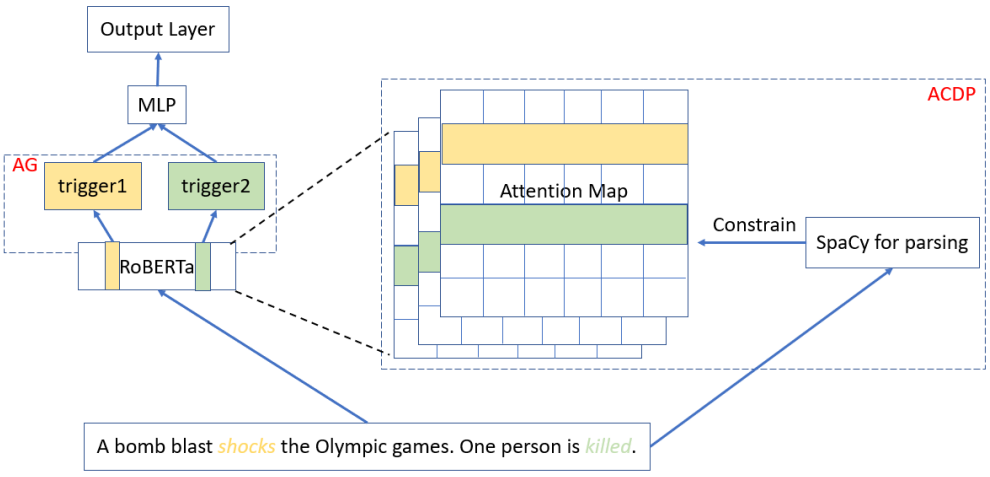

图 2-8 ACDP 模型结构图示

为此，我们设计了一种软编码的方案，一方面让模型额外地关注那些位于最短依存路径上的词，另一方面让模型依照不同的重要性程度关注那些不在最短依存路径上的词：我们认为，如果一个词到最短依存路径的距离越近，那么这个词就具有越高的重要性。此方法对应的模型结构图如图 2-8 所示，

我们将之称为 ACDP（Attention Constrained by Dependency Parsing）模型。

具体地，对于句子中的第t个词，我们首先计算它到事件触发词 A 与 B 之间的最短依存路径的距离d_t，之后，我们如公式（2-22）所示计算出依据相对最短依存路径的距离得到的句子中每个词的重要性得分s_t^{dep}。我们将之视为是依据依存句法分析得到的每个词的重要程度的概率分布。

$$s_t^{dep} = \frac{\exp(-d_t)}{\sum_{j=1..N} \exp(-d_j)} \tag{2-22}$$

其中，N为当前输入中词的总数。

另一方面，我们认为预训练语言模型中的 Transformer 结构中的注意力矩阵可以视为是由模型给出的其他词相对于某一个词的重要程度的概率分布。具体地，在本任务中，取事件触发词 A 和事件触发词 B 在注意力矩阵中所对应的两行s_t^{modA}和s_t^{modB}，我们认为这两行所对应的概率分布应该与上述的依据依存句法分析得到的每个词的重要程度的概率分布s_t^{dep}尽可能地接近。为此，我们在如公式（2-21）所示的交叉熵损失函数的基础上，增加如公式（2-23）所示的 KL 散度作为额外的损失函数项，以拉近概率分布s_t^{modA}和s_t^{modB}相对概率分布s_t^{dep}的距离。最终的损失函数如公式（2-24）所示，其中α为调节两个损失函数之间的权重的超参数。

$$L_{KL}\left(R|A,B,q\right) = -\sum_{i=1..N} s_i^{dep} log \frac{s_i^{dep}}{s_i^{mod}} \tag{2-23}$$

$$L\left(R|A,B,q\right) = -log \frac{P_{AB}^R}{\sum_{r\in\{Relation\_set\}} P_{AB}^r} - \alpha\left(\sum_i s_i^{dep} log \frac{s_i^{dep}}{s_i^{modA}} + \sum_i s_i^{dep} log \frac{s_i^{dep}}{s_i^{modB}}\right) \tag{2-24}$$

通过上述方式限制预训练语言模型的注意力分布，ACDP 模型得以以事件为中心软性地融入句子的句法结构信息。

2.5 实验验证与分析

2.5.1 实验数据集

本研究采用 MATRES 数据集[6]和 TDDiscourse 数据集[5]作为实验数据。其中，MATRES 数据集只比较位于主轴上的事件的开始时间，此标注策略保证了

标注结果的 IAA 较高，数据集中的噪声较少，因而在事件时序关系抽取任务中被广泛采用；而 TDDiscourse 数据集则包含了跨句的事件对之间的时序关系样例，能更好的地表征事件时序关系抽取模型的篇章层级建模能力，与我们预期的研究目标相匹配。TDDiscourse 数据集包含 TDD-Man 和 TDD-Auto 两个子集，分别采用纯人工标注和辅助自动标注的方式构建。

MATRES 数据集包含四种类别的标签：before、after、equal 和 vague，其中 vague 标签的含义是仅参考已掌握的信息，此事件对的时序关系存在多种可能性，不能高置信度地分类为前三种类型中的某一个。MATRES 数据集由 TimeBank、Aquaint、Platinum 三个文档集合组成，一般使用 Platinum 作为测试集，TimeBank 和 Aquaint 则合并用于模型的训练和开发。MATRES 数据集的基本统计信息如表 2-1 所示。

表 2-1 MATRES 数据集基本信息

文档集合	before	after	equal	vague	总计
timebank.txt	3229	2044	208	855	6336
aquaint.txt	3233	2263	232	676	6404
platinum.txt	424	269	31	113	818
总计	6886	4576	471	1644	13577

MATRES 数据集采用微平均 F1 值作为其评价指标，由于 vague 标签的特殊性，其微平均 F1 值的计算方式与常规方法稍有不同。设各变量的含义如图 2-9 的混淆矩阵中所示，则其微平均 F1 的计算公式如公式（2-25）、（2-26）、（2-27）所示。

$$Precision = \frac{C_{b,b} + C_{a,a} + C_{e,e}}{S_1} \tag{2-25}$$

$$Recall = \frac{C_{b,b} + C_{a,a} + C_{e,e}}{S_2} \tag{2-26}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2-27}$$

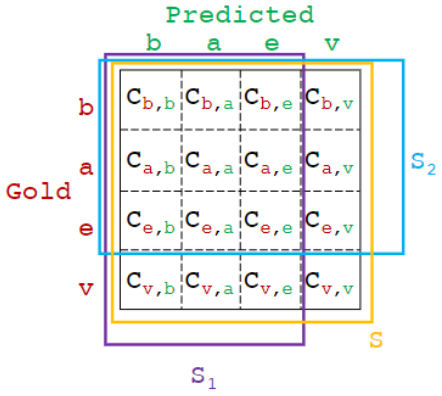

图 2-9 MATRES 数据集的微平均 f1 值的计算方式[29]

TDDiscourse 数据集包含五种类型的标签：before（b）、after（a）、simultaneous（s）、includes（i）、is included（ii）。TDDiscourse 数据集由 TDD-Man、TDD-Auto 两个子集构成，分别给出了官方的训练集、开发集、测试集划分。TDDiscourse 数据集的基本统计信息如表 2-2 所示。

表 2-2 TDDiscourse 数据集基本信息

数据集	数据拆分	b	a	s	i	ii	总计
TDD-Man	训练集	983	613	64	1300	1016	3976
	开发集	172	253	29	124	72	650
	测试集	401	188	46	572	293	1500
	总计	1556	1054	139	1996	1381	6126
TDD-Auto	训练集	9873	11136	4515	3289	3794	32607
	开发集	327	440	504	66	98	1435
	测试集	1377	1190	689	487	515	4258
	总计	11577	12766	5708	3842	4407	38300

TDDiscourse 数据集采用常规的微平均 F1 值作为评价指标。

考虑到 MATRES 和 TDDiscourse 数据集以篇章为单位进行标注，而这些篇章中的绝大多数的长度超过了预训练语言模型单次输入的上限，我们使用 NLTK 工具包[49]对这些篇章进行分句操作。为了自动地获取这些句子的句法结构信息，我们统一使用开源的 SpaCy 工具包对这些句子进行依存句法分析。

2.5.2 实验结果与分析

我们首先在 MATRES 数据集和 TDDiscourse 数据集上对比不同的事件表达方式的效果。在实验中，我们设置 batch size 为 128，梯度累计步骤为 4，

epoch 为 10，选用 base 版本的 RoBERTa 预训练语言模型，我们使用待学习率预热的 Adam 优化器在开发集上进行学习率的调优。实验结果如表 2-3 所示。为了提高实验结果的可靠性，我们使用五个不同的随机数种子进行初始化，汇报这五轮实验结果的平均值及其标准差。在后续实验中，如无特殊说明，我们统一采用此实验设置。

表 2-3 不同事件表达方式的实验结果

预训练语言模型	子词聚合方式	MATRES	TDD-Man	TDD-Auto
RoBERTa	first subword	78.90±1.16	43.96±2.46	66.66±1.46
	max	79.73±0.35	—	—
	average	79.86±0.29	—	—
	attention	80.18±0.42	45.48±1.61	67.22±1.48

从实验结果可以看出，在事件时序关系抽取任务中，相比于只使用事件触发词的第一个子词来表征事件的基线模型，三种不同的聚合事件的所有子词的方式都能够带来一定的性能提升，这说明了事件触发词的每个子词对于表达事件的全部信息都是有帮助的。我们认为这是因为事件触发词往往是一个动词，而动词的时态信息对于判别事件的时序关系是至关重要的。在英语中，时态信息由词的后缀体现，只使用事件触发词的第一个子词表征整个事件可能会丢失其时态信息。事实上，TDDiscourse 数据集给出了测试集中的一部分样例所对应的时序推理类型，在 TDD-Man 和 TDD-Auto 中各标注出了时态指示（Tense Indicator）类型 13 个和 51 个。应用注意力机制聚合事件触发词的全部子词来表征事件的模型正确分类了其中的 11 个和 43 个，而只使用事件触发词的第一个子词来表征事件的模型只正确分类了 9 个和 36 个。

最终，应用注意力机制的子词聚合方式取得了最佳的实验结果。我们认为这是因为此参数化的方法具有更强大的表征能力，能够自适应地抽取出所有子词中与事件时序关系相关的信息。在后续实验中，如无特别说明，我们统一使用注意力机制来聚合所有子词以获取词的向量表达。

接下来，我们继续在这两个数据集上进行实验，以验证句子的句法结构信息对于事件时序关系抽取任务的有效性。在 ACDP 模型的实验中，对于要限制的预训练语言模型中注意力矩阵的层数，我们在{最后一层、后两层、后三层、后六层、所有层}中进行调参；对于公式（2-24）中的α，我们在{0.01、0.05、0.1、0.2、0.5、1.0、10}中进行调参，其余超参数设置与之前的实验设置相同。实验结果显示，限制后两层且当α为 0.5 时取得最优的实验结果。

在 MATRES 数据集和 TDDiscourse 数据集上的实验结果如表 2-4 所示。其中 BERT-base Transformer 和 RoBERTa-base 分别使用 BERT 预训练语言模型和 RoBERTa 预训练语言模型，在句子层级进行逐对的事件时序关系预测[50]。

表 2-4 不同依存句法分析的应用方式的实验结果

模型		MATRES	TDD-Man	TDD-Auto
BERT-base Transformer		77.2	37.5	62.3
RoBERTa-base		78.9	37.1	61.6
RoBERTa +	SDP+BiLSTM	80.69±0.18	45.25±1.39	68.28±1.35
	Sent+BiLSTM	80.56±0.63	—	—
	ACDP	81.04±0.58	47.86±1.40	69.32±0.75

从在 MATRES 数据集上的实验可以看出，在事件时序关系抽取任务中，相比于将整个句子使用 BiLSTM 进行额外编码的 Sent+BiLSTM 模型，只将最短依存路径上的词进行额外编码的 SDP+BiLSTM 模型的实验结果更佳，这两个实验结果的对比验证了依存句法分析结果对事件时序关系抽取任务的有效性。

相比于硬性地抽取最短依存路径上的词的 SDP+BiLSTM 模型，软性的 ACDP 模型在三个数据集上一致地取得了最佳的实验结果，我们认为一方面这与我们的假设一致，即不在最短依存路径上的词同样可能提供对于判别时序关系至关重要的信息，软性的应用方式使得模型能以事件为中心更好地利用这些词的信息；另一方面，这种方法还避免了引入额外的参数，减低了模型的参数量，同时模型的推理速度也更快。

在 TDD-Man 和 TDD-Auto 中各标注出了 27 个和 100 个单句信息（SS，Single Sentence）时序推理类型的样例。ACDP 模型正确分类了其中的 12 个和 66 个，而 SDP+BILSTM 的模型只正确分类了 8 个和 61 个。对 SS 时序推理类型更好的预测结果说明 ACDP 模型通过根据依存句法分析的结果限制预训练语言模型的注意力分布，能够更好地挖掘句子级别的局部事件时序信息。

2.6 本章小结

当前使用预训练语言模型的事件时序关系抽取方法普遍使用事件触发词的第一个子词表征整个事件，我们提出通过注意力机制聚合事件触发词的全部子词的方式来获取更好的事件表达，实现结果显示此事件表达方法能够取

得更好的事件时序关系抽取效果，特别是对于依赖时态信息确定时序关系的事件对有一定的性能提升。

早期的事件时序关系抽取技术的研究注意到了两个事件触发词的最短依存路径上的词对于确定这两个事件的时序关系的重要性，我们首先通过使用 BiLSTM 对预训练语言模型的编码结果中最短依存路径上的词进行额外编码的方式，验证了应用预训练语言模型时，最短依存路径上的词依然有其重要性。之后，我们设计了依据句子中的词到最短依存路径的距离对句子中的词的重要性程度进行打分，并据此对基于 Transformer 结构的预训练语言模型的注意力矩阵进行软性限制的方式，在不增加额外参数量的情况下，取得了更好的事件时序关系抽取任务的性能。此方法对于依赖句子局部信息确定时序关系的事件对之间的时序分类的准确率显著提高。

第 3 章 基于信息增强的事件时序关系抽取方法

3.1 引言

当前阶段，事件时序关系抽取任务研究的限制因素之一数据集的规模。事件时序关系的标注需要首先筛选出富含事件的生文本片段，再在其上标注出事件，随后标注出存在时序关系的事件对之间的时序关系。一方面，不同标注规范对于事件、事件时序关系类别的定义存在差异，另一方面，事件之间的时序关系往往模糊，存在不同的合理解释。受限于如上所述的事件时序关系标注的多阶段性、困难性和歧义性，已有的事件时序关系数据集的规模往往有限。在这些有限的数据集上训练的模型性能受限，同时其可靠性也得不到保证。

因此，我们首先尝试扩充已有的数据集的规模，具体地，考虑到将句子中的事件触发词替换为其同义词，两个事件之间的时序关系一般不会改变，我们在事件时序关系抽取任务中应用了基于同义词替换的数据增强方法。其次，模型在有限的数据集上可能学习不到现实世界中复杂的时序常识知识，我们通过训练额外时序常识知识编码的方式补充模型的常识信息。最后，为了更充分地利用每一条训练数据，我们在模型中增加了两个与事件时序相关的子任务，以引导模型更好地捕获文本中的时序信号。

本章的主要内容如下：第 3.1 节为本引言；第 3.2 节介绍应用于事件时序关系抽取任务的基于同义词替换的数据增强方法；第 3.3 节介绍额外时序常识编码的训练方式及其应用方式;第 3.4 节介绍事件相对时间预测子任务；第 3.5 节介绍时序关系对比学习子任务；第 3.6 节对本章所介绍方法进行实验验证和分析；第 3.7 节总结本章的内容。

3.2 基于同义词替换的数据增强方法

当前阶段，事件时序关系抽取任务研究的一个显著问题是数据集规模往往有限，不少研究者提出了在此任务上进行数据增强的方案，我们同样给出了对此的一些思考。关系抽取任务中广泛使用的远程监督方法认为实体之间的关系由这两个实体本身决定，如果数据库中标注出两个实体之间呈现某种关系，那么任何存在这两个实体的共现的自然语言文本，都能够部分地体现这种关系。但我们认为不同于名词性的实体，事件的关系更多地由两个事件

所处的上下文决定。将数据集样例中的两个动词性事件单个或同时替换为其同义词，并尽可能地保留其时态、人称等方面的词形信息，此样例的时序标签类别不变。

具体地，我们采用同义词替换的数据增强方法，通过将事件触发词替换为其动词性的同义词，低成本地构造大量的训练数据。一方面，此方法能够简易地成倍扩充训练数据；另一方面，这些样本暗示了模型应当更加关注上下文信息，即尽管事件触发词不同，但是只要它们位于相同的上下文语境之中，那么事件的时序关系类别就应当是一致的。这将让模型感知到事件对的时序关系由其上下文决定，而不是过拟合于事件触发词对本身的分类上。

在实现过程中，我们使用 NLTK 工具包获取句子中事件触发词的人称和时态标注，使用 NLTK 工具包中整合的 WordNet 获取事件触发词的若干个同义词，使用开源的 python 包 pattern 将同义词变换为与事件触发词在句中一致的人称和时态形式。为了与其他实验的结果进行公平的对比，我们仅以此方法扩展模型的训练集，而保留原来的开发集和测试集不变。

同时，在之前的实验过程中，我们留意到训练得到的模型对于占比较少的 equal 和 vague 类别的预测表现普遍不佳，换言之，数据集中类别的不均衡也是在其上训练的模型性能的限制因素之一。为此，我们设计了通过上述提出的数据增强方案，在 MATRES 数据集上只对占比较少的小类别 equal 和 vague 进行扩展，而保留占比较多的 before 和 after 类别不变，使得各个类别的样例数量比例相当的实验，以此应对类别不均衡的问题。

3.3 额外时序常识知识编码

此外，我们添加了应用 TemProb 知识库[28]的子模块，其模型结构如图 3-1 所示。

TemProb 知识库的作者留意到在真实文本中，两个动词对应的事件所发生的时序关系是存在一定的偏置的。例如，我们总是先“打开电脑”，然后开始“编写代码”。为了挖掘这样的先验常识知识，他们构建了 TemProb 知识库，这个知识库中统计了在大规模语料库中每组动词对呈现每种时序关系的频率。

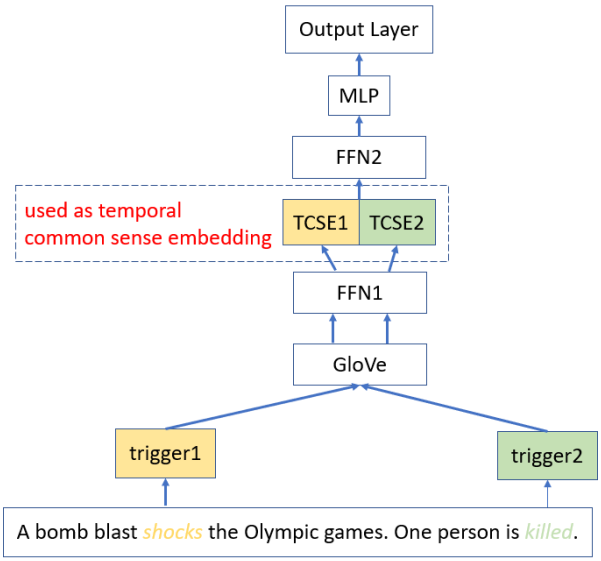

图 3-1 应用 TemProb 知识库的子模块的模型结构图示

我们在此时序常识知识库 TemProb 之上训练时序常识知识编码。具体地，我们首先使用 GloVe 词向量将词对进行向量化得到W_A和W_B，然后使用双塔结构来预测此词对在 TemProb 知识库中高频的时序关系类别r。我们首先使用一个如公式（3-1）所示的非线性层得到其降维后的表达$\tilde{W}_A$，然后如公式（3-2）所示，拼接后使用另一个非线性层进行时序类别预测。

$$\tilde{W}_A = \sigma\left(LinearLayer1\left(W_A\right)\right) \tag{3-1}$$

$$P_r = \sigma\left(LinearLayer2\left(cat\left(\tilde{W}_A, \tilde{W}_B\right)\right)\right) \tag{3-2}$$

其中，线性层 *LinearLayer1* 和 *LinearLayer2* 对输入向量执行一次线性变换，σ 为 *ReLU* 非线性激活函数，*cat* 表示向量拼接操作。

当此子模块进行充分训练之后，我们可以认为$\tilde{W}_A$编码了事件的时序常识知识。在应用时，我们将$\tilde{W}_A$作为时序常识知识编码拼接到预训练语言模型给出的事件编码结果F_A之上，使得主模型具有时序常识知识的感知能力。

我们同样使用公式(2-21)所示的交叉熵损失函数对此子模块进行优化。

3.4 事件相对时间预测子任务

在之前的研究中，我们将事件时序关系抽取任务视为是以事件对为对象的分类任务，这种研究范式有一些潜在的问题。首先，在文档层面，对于此文档的所有事件所构成的全部事件对，若孤立地将这些事件对进行时序类别的分类，可能导致最终得到文档的时序关系图中存在不一致性。其次，假设

文档中事件的数量与文档的长度正相关，则文档中待分类的事件对的数量相对文档长度的复杂度为$o\left(n^2\right)$，这意味着随着文档长度的增加，模型需要预测的事件对的数量将快速增长，加重了模型的预测负担。最后，这与我们人类理解文档中事件的先后顺序的过程并不一致，人类在理解文档中的事件时，更自然的过程是将每一个事件对应到一个具体的事件发生时间，然后通过比较这些事件的具体发生时间对事件发生的先后顺序进行排序。

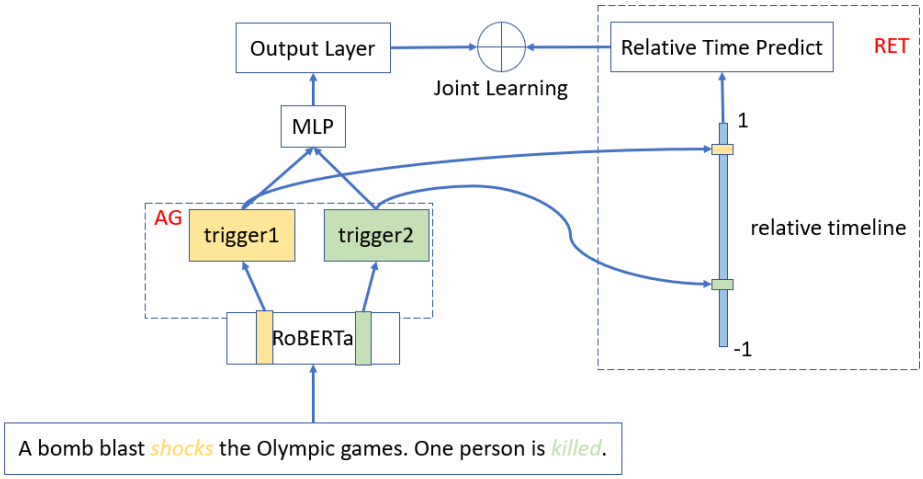

图 3-2 引入事件相对时间预测子任务的模型结构图示

基于上述原因，我们引入了事件相对时间预测子任务作为模型训练的辅助任务，来提高事件时序关系抽取模型的性能。此模块的模型结构如图 3-2 所示。具体地，在我们依照公式（2-14）到（2-19）所示的方法获取事件 A 和事件 B 各自的向量表达F_A和F_B后，如公式（3-3）所示，通过一组最终输出维度为 1 的两层前馈神经网络（即公式（3-3）中的FFN），我们得到投影在-1 至 1 之间的事件 A 和事件 B 的相对发生时间t_A和t_B。

$$t_X = tanh\left(FFN_2\left(ReLU\left(FFN_1\left(F_X\right)\right)\right)\right) \tag{3-3}$$

之后，我们使用如公式（3-4）所示的损失函数来优化事件相对时间预测子任务。

$$\begin{aligned} L_R = & I\left[r_{AB} = before\right] max\left(0,1-\left(t_B - t_A\right)\right) \\ & + I\left[r_{AB} = after\right] max\left(0,1-\left(t_A - t_B\right)\right) \\ & + I\left[r_{AB} = equal\right] \left|t_A - t_B\right| \end{aligned} \tag{3-4}$$

其中r_{AB}为事件 A 与 B 之间真实的时序关系标签，$I[\bullet]$为指示函数，若其输入的判别条件为真，则输出值为 1，否则输出值为 0。公式（3-4）的含义是，当事件 A 与事件 B 之间的时序关系为 equal 时，它们的相对发生时间t_A和t_B之间的距离应该被优化到尽可能小；当事件 A 与事件 B 之间的时序关系为 before 时，事件 B 的相对发生时间t_B应被优化到比事件 A 的相对发生时间t_A大 1；反之，当事件 A 与事件 B 之间的时序关系为 after 时，事件 B 的相对发生时间t_B应被优化到比事件 A 的相对发生时间t_A小 1；在事件 A 与事件 B 之间的时序关系为 vague 时，则不做处理。

通过此种方式，我们希望模型能够通过公式（3-4）所示的额外限制，更充分地利用事件对的时序类别标签。同时，事件相对时间预测子任务将使得模型能够捕获到事件时序本身的线性性，也即事件的发生时间事实上可以被映射到现实时空中时间线上的一个点，所有事件之间都可以通过比较两者在此时间线上的相对位置来确定它们之间的时序关系。

最终模型优化所使用的损失函数如公式（3-5）所示。它是公式（2-21）所示的交叉熵损失函数与公式（3-4）所示的损失函数的加权和，其中β为用于调节两者的权值。

$$L\left(R\middle|A,B,\theta\right)=-log\frac{P_{AB}^{\mathrm{R}}}{\sum_{r\in\{Relation\_set\}}P_{AB}^{r}}+\beta\begin{pmatrix}I\left[r_{(A,B)}=before\right]max\left(0,1-\left(t_B-t_A\right)\right)\\+I\left[r_{(A,B)}=after\right]max\left(0,1-\left(t_A-t_B\right)\right)\\+I\left[r_{(A,B)}=equal\right]\left|t_A-t_B\right|\end{pmatrix} \quad (3\text{-}5)$$

3.5 时序关系对比学习子任务

对比学习是自监督学习的一种，其目的在于从缺乏标注标签的生数据中学习更好的特征表达。对比学习最早大规模应用于图像领域，它通过设计自监督的预训练任务，使得在缺乏大规模有标注数据集的情况下，学习图像信号的先验知识分布成为可能[51][52][53]。

研究者通过对比学习的方式学习得到一个编码器，使得当输入的两个实例相似时，此编码器输出的编码向量的相似度较高，而当输入的两个实例不相似时，此编码器输出的编码向量的相似度较低。在实践中，一般使用余弦相似度作为编码向量的相似度度量函数。此外，编码器输出端一般连接一个L_2正则化函数，将输出的编码向量的模长放缩到 1。

事实上，对比学习的核心设计点在于如何构造相似的实例和不相似的实例。在图像领域，一般通过对同一个图像进行裁剪、平移、对称、旋转等操作构造此图像的相似实例或称正样本，而通过在其他图片中进行随机采样得到原始图像的不相似实例或称负样本。在实践中，为了尽可能加速训练过程，一般选择同一个批量的其他图像作为负样本。

针对事件时序关系抽取任务，在我们依照公式（2-14）到（2-19）所示的方法获取事件的向量表达 F_X 之后，我们将待进行时序分类的事件对 A、B 整体作为一个实例 R_{AB}， R_{AB} 的计算方式由公式（3-6）确定。

$$R_{AB} = cat\left(norm\left(F_A\right), norm\left(F_B\right)\right) \tag{3-6}$$

其中， $norm$ 为 L_2 正则化函数， cat 表示向量拼接操作。

设每个批量中所有具有时序标注的事件对构成集合 τ，我们将此批量中事件时序关系类别相同的事件对相互作为正样本，而将此批量中事件时序关系类别不同的事件对相互作为负样本。以事件对 R_{AB} 为考察对象，设事件对 R_{AB} 的所有正样本构成集合 τ_R^+，事件对 R_{AB} 的所有负样本构成集合 τ_R^-，则时序关系对比学习子任务对应的损失函数 L_C 如公式（3-7）所示。

$$L_C = \sum_{R\in\tau}\left[-\sum_{R_P\in t_R^+} log\frac{exp\left(\frac{cos\left(R,R_P\right)}{T}\right)}{|t_R^+|\sum_{R_N\in\tau_R^-} exp\left(\frac{cos\left(R,R_N\right)}{T}\right)}\right] \tag{3-7}$$

其中 T 为对比学习中使用的温度超参数， cos 表示余弦相似度函数。

最终的损失函数如公式（3-8）所示。它是公式（2-21）所示的交叉熵损失函数与公式（3-7）所示的损失函数的加权和，其中 γ 为用于调节两者的权值。

$$L\left(R\middle|A,B,\theta\right) = -log\frac{P_{AB}^{R}}{\sum_r P_{AB}^r} + \gamma\left(\sum_{R\in\tau}\left[-\sum_{R_P\in t_R^+} log\frac{exp\left(\frac{cos\left(R,R_P\right)}{T}\right)}{|t_R^+|\sum_{R_N\in\tau_R^-} exp\left(\frac{cos\left(R,R_N\right)}{T}\right)}\right]\right) \tag{3-8}$$

最终我们实现的引入时序关系对比学习子任务的模型的结构如图（3-3）所示。

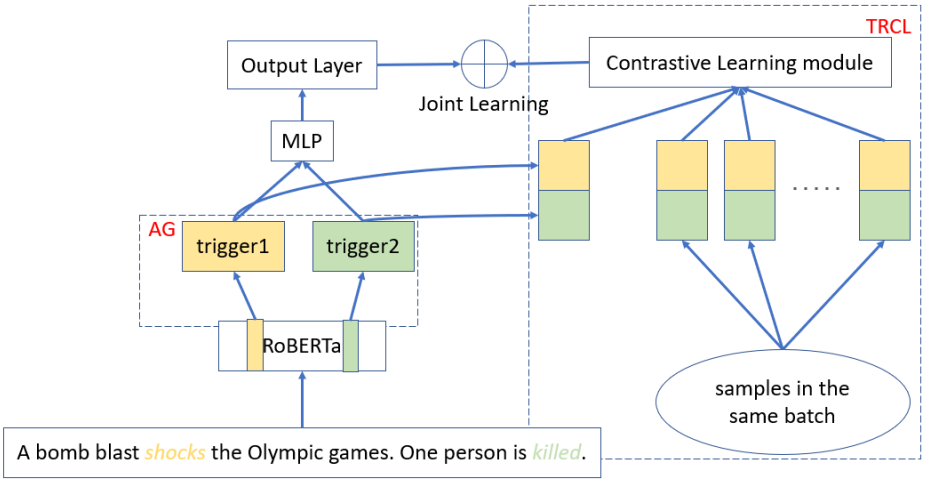

图 3-3 引入时序关系对比学习子任务的模型结构图示

3.6 实验验证与分析

我们在 MATRES 和 TDDiscourse 数据集上进行实验，以验证所采取的几种信息增强方法的有效性。对于公式（3-4）所示的事件相对事件预测子任务的损失函数项 L_R 的权值 β 和公式（3-7）所示的时序关系对比学习子任务的损失函数项 L_C 的权值 γ，我们在{0.01、0.05、0.1、0.2、0.5、1.0、10}中进行调参，最终的实验结果显示，β 值为 0.01、γ 值为 0.2 时，取得最优的实验结果。对于时序关系对比学习子任务的损失函数项 L_C 中的温度超参数 T，经过在开发集上调优后，我们使用 0.05。

表 3-1 是在 MATRES 数据集和 TDDiscourse 数据集上的实验结果。其中其中 BERT-base Transformer 和 RoBERTa-base 分别使用 BERT 预训练语言模型和 RoBERTa 预训练语言模型，在句子层级进行逐对的事件时序关系预测[50]；+DA（Data Augment）即 3.2 节所介绍的通过同义词替换的方式将所有类别都进行 4 倍扩充的方法；+Balanced DA 则为将 MATRES 数据集中的 equal 类别和 vague 类别分别进行 9 倍和 4 倍扩充，而 before 类别和 after 类别不进行扩充，以使得类别比例接近均衡的方法；+TemProb 则为 3.3 节所介绍的在 TemProb 知识库上训练额外的时序常识知识编码的方法；+RET（Relative Event Time）为 3.4 节所介绍的引入事件相对时间预测子任务的方法；+TRCL（Temporal Relation Contrastive Learning）为 3.5 节所介绍的引入事件时序关系对比学习子任务的方法。

表 3-1 不同信息增强方式的实验结果

模型		MATRES	TDD-Man	TDD-Auto
BERT-base Transformer		77.2	37.5	62.3
RoBERTa-base		78.9	37.1	61.6
RoBERTa	+ DA	80.50±1.19	—	—
	+ Balanced DA	80.34±0.66	—	—
	+ TemProb	80.97±0.26	45.22±2.94	68.59±1.11
	+ RET	81.42±0.60	46.10±2.07	69.47±0.79
	+ TRCL	81.18±0.73	46.27±1.47	70.94±1.21

从实验结果可以看出，在事件时序关系抽取任务中，使用同义词替换的方式扩充训练数据的方法能够带来稳定的性能提升。然而，在针对 MATRES 数据集的实验中我们也发现了一些待解决的问题。在几乎所有实验中，我们一致地发现由于如表 2-1 所示的数据不均衡因素的影响，对于 equal 和 vague 两个小类别的预测准确率很低，特别是对于 equal 类别，几乎没有模型能够正确的分类任何一个样本。因此，我们设计了通过数据增强的方式将 equal 和 vague 两个小类别的数据分别扩充 9 倍和 4 倍以与 before 和 after 类别数据量相当的实验（即表 3-1 中的 RoBERTa + Balanced DA 模型）。我们期望通过使用数据增强的方式将四个类别的样例数量扩展到相当的水平，能够训练得到在各个类别上准确率相当的模型。然后，在实验结果中虽然能够观察到相对基准模型整体 F1 值的提升，但是其训练得到的模型对于 equal 和 vague 类别的预测准确率并没有明显的改善，与 before 和 after 类别的准确率依然有很大的差距。这说明简单的同义词替换策略所构造的训练样例的复杂程度有限，单纯的基于同义词替换的数量扩增并不能使得模型学习到正确识别 equal 和 vague 类别的特征。此外，与第二章中应用注意力机制进行子词聚合的模型的结果对照，基于同义词替换的数据增强方案带来的性能提升只有 0.32，但是却需要花费 4 倍的训练时间，从模型训练的角度来说，此种方法的性价比较低。综上所述，我们认为在事件时序关系抽取任务中，基于简单的同义词替换策略的数据增强方案的效果欠佳，能够带来的模型性能的提升较为有限，同时对于解决类别之间的不均衡问题也没有明显改善，需要探索更为有效的数据增强方案。

同时，通过拼接在 TemProb 知识库上训练得到的时序常识知识编码，模型也能够取得一定的性能提升。在 TDD-Man 和 TDD-Auto 数据集中各标注出

了 18 个和 1 个世界知识（WK，World Knowledge）时序推理类型的样例，应用注意力机制聚合事件触发词的全部子词来表征事件的模型正确分类了其中的 10 个和 0 个，而 RoBERTa + TemProb 模型则能够正确分类其中的 13 个和 1 个，这一结果验证了我们预期的时序常识知识编码的效果。

另外，在事件时序关系抽取任务中，结合事件相对时间预测子任务也能够带来明显的性能提升。结合之前的分析，我们认为此子任务引导模型直接比较所有事件在时间线上的相对位置来判断事件之间的时序关系，对于在时间线上关联多个事件的时序关系预测任务的提升可能会更加明显。而在实验结果中，值得注意的是，在 TDD-Man 和 TDD-Auto 数据集的测试集中标注出的 63 个和 10 个需要链式推理（CR，Chain Reason）的样例中，基线模型分别正确分类了 34 个和 4 个，而结合事件相对时间预测子任务的模型则能够正确分类 43 个和 8 个，这说明事件相对时间预测子任务对于涉及多个关联事件、需要链式推理才能正确分类的样例能够提供明显的帮助，从而验证了事件相对时间预测子任务对于需要综合整个时间线的信息才能正确分类的样例的有效性。

最后，在事件时序关系抽取任务中，结合时序关系对比学习子任务也能够带来明显的性能提升。我们认为时序关系对比学习子任务能够更充分地挖掘事件对的时序类别标签所携带的信息，引导模型捕获具有相同时序类别标签的事件对之间的共性和具有不同时序类别标签的事件对之间的差异性，使得模型学习到更好的应用于事件时序关系抽取任务中的事件特征向量表达。在样本规模有限的事件时序关系抽取任务中，引入时序关系对比学习子任务能够有效地提升模型的学习效果。

3.7 本章小结

由于现阶段事件时序关系抽取任务的相关数据集的规模往往有限，我们设计了一系列的信息增强方式，来更好地利用每一条珍贵的训练数据。

通过对事件触发词进行时态、人称形式不变的同义词替换，可以低成本地扩充训练数据，但是此方法扩充出的训练数据所能提供的样本复杂度有限，所能带来的性能提升并不让人满意。

通过在时序常识知识库 TemProb 上训练事件时序常识知识编码并整合入模型中，可以使得模型感知到在规模有限的数据集上学习不到的时序常识知识，实验结果表明此方法对于需要世界知识的事件对的时序类别预测具有一定的帮助。

为了能够更有效地利用每一条训练数据，借由事件相对时间预测子任务和事件时序对比学习子任务两个辅助任务，模型能够学习到更好的适用于事件时序关系抽取任务的事件向量表达，实验结果显示这两个辅助任务都能够提升事件时序关系抽取模型的性能。

第 4 章 基于篇章结构的事件时序关系抽取方法

4.1 引言

在上述章节的事件时序关系抽取模型中，我们以事件对所在的两个句子作为模型的输入，来判断这两个事件对之间的时序关系。一方面，这种处理范式仅考虑了句子级别的局部信息，而对判断当前事件对的时序关系有帮助的信息可能位于文档中其他位置；另一方面，这忽略了事件时序关系的一致性要求。相比于其他事件关系抽取任务，对事件间时序关系的预测必须统筹考虑整个文档的全局信息，并参考关联事件对间的时序关系以避免不满足事件时序一致性的预测出现。

为了整合文档的全局信息,并使得模型在判断当前事件对的时序关系时，能感知到关联事件的时序关系情况以做出满足时序一致性的预测，我们设计了基于篇章结构的篇章层级事件时序关系抽取模型。首先，我们以篇章为单位使用预训练语言模型进行编码，对于超过预训练语言模型编码长度限制的文档，我们切分出尽可能长的连续文档片段逐次编码，并构建文档的层级结构图。此外，我们在文档中的事件节点之间添加时序边，将每篇文档映射为一幅图，之后在此图之上使用图注意力神经网络进行篇章层级的事件时序关系建模，最终输出每个待预测的事件对的时序关系类别。实验结果显示此方法能够在篇章层级事件时序关系抽取数据集 TDDiscourse 的两个子集上一致地取得更好的实验结果。

本章的主要内容安排如下：第 4.1 节为本引言；第 4.2 节介绍图神经网络特别是图注意力神经网络；第 4.3 节介绍应用图注意力神经网络的篇章层级事件时序关系抽取模型;第 4.4 节对本章所介绍方法进行实验验证和分析；第 4.5 节为本章内容的一个小结。

4.2 图注意力神经网络

随着越来越多大规模数据集的构建以及计算机运算能力的显著提高，深度学习技术席卷了人工智能研究的绝大多数领域，并普遍地证明了其强大的建模能力。

然而，虽然传统深度学习擅长于在欧式空间中提取数据的潜在特征，但是在一部分实际应用场景中，真实数据是从非欧式空间中产生的。典型地，

在电子商务场景中，商品、品牌、店铺、客户等参与要素之间的交互关系是任意的，它们相互之间构成一张无规则的图，每个节点拥有数量不定的邻接节点，这些邻接节点的数据类型也各不相同。传统的卷积神经网络、循环神经网络等深度学习模型往往要求输入数据具有特定的有规则的结构，不适用于此类结构不规则的数据场景的建模。

近年来，随着研究人员发现越来越多的应用场景可以抽象出具有图结构的数据，图神经网络受到越来越多的关注。图神经网络用于学习图结构中节点、边、子图、全图等各层级元素的表示。图神经网络一般由串联的若干个参数不共享但是运算规则相同的层组成。在每一层中，以上一层输出的节点的向量表达作为本层的输入，通过聚合当前节点的邻接节点的信息，学习新的当前层的节点的嵌入向量，并传递给下一层。

我们以图卷积神经网络 GCN[54]（Graph Convolution Netword）为例来介绍图神经网络。图卷积神经网络是将卷积神经网络直接从图像领域推广到图结构数据的结果，也是最基础、最经典、结构最清晰的图神经网络之一。类比卷积神经网络，图卷积神经网络采用卷积操作，聚合当前节点的邻接节点的信息。对于第 l 层的第 i 个节点 $h_i^{(l)}$，设 $\mathcal{N}(i)$ 为此节点的邻接节点所构成的集合，则经过第 l 层的图卷积操作后，此节点的新的向量表达 $h_i^{(l+1)}$ 如公式（4-1）所示。

$$h_i^{(l+1)} = \sigma\left(b^{(l)} + \sum_{j \in \mathcal{N}(i)} \frac{W^{(l)} h_j^{(l)}}{|\mathcal{N}(i)|} \right) \tag{4-1}$$

其中，$W^{(l)}$ 和 $b^{(l)}$ 分别为待学习的第 l 层的权重矩阵和偏置，σ 为非线性激活函数，例如 $ReLU$ 函数。可以看到，图卷积神经网络实际执行的操作即是首先对当前节点的每一个邻接节点进行相同的线性变换，再执行一次平均池化操作，之后通过一个非线性激活函数引入非线性，得到最终的编码输出。

图卷积神经网络简单高效，但是存在一系列问题。首先图卷积神经网络忽略了边的类别信息，而在事件时序关系抽取任务中，边的类别即事件之间的时序关系类别是我们的主要研究对象，不能忽略。其次，由公式（4-1）可以看出，一层的图卷积神经网络使得每个节点聚合了其一阶邻接节点的信息，而为了使每个节点聚合更远的节点的信息（或者说扩大其感受野），就需要串联更多层的图卷积神经网络，但是受限于过渡平滑[55]的问题，深度过大的图卷积神经网络往往难以收敛。有鉴于此，在我们的研究中，使用图注意力神经网络来建模事件时序关系抽取任务中的文档篇章结构。

图注意力神经网络 GAT[39]（Graph Attention Network）的优点在于其能较好地体现图中边的标签信息，其采用的多头注意力机制亦保证了其强大的建模能力，同时 GAT 不易出现过渡平滑的现象，使得构建更深的图神经网络成为可能。相比于图卷积神经网络采用的平均池化策略，图注意力神经网络采用注意力机制来确定每个邻接节点的权重。对于第l层的第i个节点$h_i^{(l)}$，设$\mathcal{N}(i)$为此节点的邻接节点所构成的集合，则$\mathcal{N}(i)$中的第 j 个节点所对应的权值$\alpha_{ij}^{(l)}$如公式（4-2）所示。

$$\alpha_{ij}^{(l)} = softmax\left(\sigma\left(a^{(l)T}\left[W_a^{(l)}h_i^{(l)};W_a^{(l)}h_j^{(l)};e_{ij}\right]\right)\right) \tag{4-2}$$

其中，$W_a^{(l)}$为待学习的权重矩阵，$a^{(l)}$为待学习的虚拟查询向量，σ为非线性激活函数，例如$ReLU$函数。$e^{(l)}$为第i个节点和第 j 个节点之间的边的类型嵌入向量，通过在计算邻接节点的注意力分数时拼接当前节点与此邻接节点之间的边的嵌入向量，使得模型能够学习图结构中边的类型信息。

最终，以公式（4-2）的计算结果$\alpha_{ij}^{(l)}$作为权值，第i个节点的新的表达$h_i^{(l+1)}$通过如公式（4-3）所示的加权和运算得到。

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)}\alpha_{ij}Wh_j\right) \tag{4-3}$$

事实上，Transformer 结构中采用的多头注意力机制也可以应用在图注意力神经网络中，以学习在不同子语义空间中的注意力权重，丰富图注意力神经网络所能学习到的语义层次，强化其表征能力。应用多头注意力机制的图注意力神经网络的计算规则如公式（4-4）和公式（4-5）所示。

$$\alpha_{ij}^{k(l)} = softmax\left(\sigma\left(a^{k(l)T}\left[W_a^{k(l)}h_i^{(l)};W_a^{k(l)}h_j^{(l)};e_{ij}\right]\right)\right) \tag{4-4}$$

$$h_i^{(l+1)} = \|_{k=1}^{K}\sigma\left(\sum_{j\in\mathcal{N}(i)}\alpha_{ij}^{k}W^{k}h_j\right) \tag{4-5}$$

其中，K为多头注意力的头数，$\|$表示向量拼接操作，其他参数的含义与上述相同。

4.3 基于图注意力神经网络的篇章层级事件时序关系抽取

仅依据两个事件所在句子的信息并不足以解决事件时序关系抽取任务，事件时序关系的预测必须在篇章层级才能得到妥善的解决。一方面，对于当前待预测事件对之间的时序关系至关重要的信息不一定位于两个事件所处的句子之内，而可能位于文档中距离两个事件较远的位置。特别是对TDDiscourse 数据集而言，在标注过程中即限制每个标注样例的事件对之间必须间隔多于五个句子，这意味着夹在两个事件之间的数个句子都可能提及了与这两个事件的时序相关的信息。另一方面，只对文档中的事件独立地两两成对预测其时序关系，最终得到的事件间的时序关系很可能违反事件时序一致性的要求。在预测每一组事件对的时序关系时，必须综合考虑文档中每个事件的时序情况，特别是与这两个事件关联的事件的时序关系情况，才能给出满足篇章内所有事件的时序一致性要求的预测。为此，我们提出能够综合篇章的全局信息并参考关联事件的时序情况进行篇章层级事件时序关系抽取的 HDTRG（Hierarchical Discourse-level Temporal Relation Graph）模型。接下来，我们具体介绍此模型的结构及其训练和预测方法。

在之前的研究中，我们使用 RoBERTa 预训练语言模型来编码句子层级的文本信息，但是将此方法过渡到篇章层级时会遇到 RoBERTa 预训练语言模型编码长度限制的问题。为此，我们首先使用 SpaCy 工具包对每一篇文档进行分句。随后将文档以句子为最小粒度切分为一系列连续的文档片段，使得每一个片段中包含的子词数量小于 510（空余两个子词用于特殊标记[CLS]和[SEP]），同时其内的句子都是完整的且片段之间没有交集。

之后，这些文档片段即可使用 RoBERTa 预训练语言模型进行编码。接下来，我们将属于同一篇文档的所有文档片段的编码结果拼接起来，即得到 RoBERTa 预训练语言模型对整个文档的编码结果。这样的处理方式能够得到文档的向量表达，但是属于同一篇文档的不同文档片段之间的信息没有进行交互。为此，我们构建了事件-句子-文档片段-文档四级层级结构，来实现文档中任意内容之间的信息交互。

对于事件 E，如公式（4-6）和公式（4-7）所示，我们采用注意力机制整合其事件触发词涉及的全部子词 $S_{[i:j]}$ 的信息，得到其最终的向量表达 F_E。

$$\alpha_i = \frac{exp\left(\frac{S_i^T h_E}{\sqrt{d_h}}\right)}{\sum_{l=i}^{j} exp\left(\frac{S_l^T h_E}{\sqrt{d_h}}\right)} \tag{4-6}$$

$$F_E = \sum_{l=i}^{j} \alpha_l S_l \tag{4-7}$$

其中 h_E 是待学习的针对事件的虚拟查询向量，它在所有事件之间共享参数，d_h 是子词向量的维度。

得到事件 A、B 的向量表达 F_A 和 F_B 之后，我们将这两个向量以及它们的和与逐点积进行拼接，输入到一个多层感知机（即公式（4-8）中的 MLP_{ptm}）之中进行事件时序类别的预测，得到基于预训练语言模型预测出的事件时序关系类别概率分布 P_{ptmAB}^{r}。

$$P_{ptmAB}^{r} = MLP_{ptm}\left(cat\left(F_A, F_B, F_A + F_B, F_A \odot F_B\right)\right) \tag{4-8}$$

对于句子 S，我们同样采用注意力机制整合其涉及的所有子词 $S_{[1:|S|]}$ 的信息，得到最终的向量表达 F_S。

$$\alpha_i = \frac{exp\left(\frac{S_i^T h_S}{\sqrt{d_h}}\right)}{\sum_{l=1}^{|S|} exp\left(\frac{S_l^T h_S}{\sqrt{d_h}}\right)} \tag{4-9}$$

$$F_E = \sum_{l=1}^{|S|} \alpha_l S_l \tag{4-10}$$

其中 h_S 是针对句子表示的待学习的虚拟查询向量，它在所有句子之间共享参数，d_h 是子词向量的维度。

对于文档片段 L，我们直接取 RoBERTa 预训练语言模型的输出中标记<s>所在位置的向量作为其最终的表达 F_L。

对于文档 D，我们使用一个随机初始化的可学习向量 F_D 作为其表达，它在所有的文档之间共享。事实上，我们可以认为这个表征文档的可学习向量对应了文档创建时间，也就是 DCT（Document Creation Time）。

我们所构建的事件-句子-文档片段-文档四级层级结构被映射为一幅图。图中的节点分别对应事件、句子、文档片段和文档，这些节点具有不同的含义，但是使用相同维度的向量进行表征。图中的边为有向边，分为两种类型，分别对应属于关系和包含关系。我们仅在相邻层级的元素之间连接边。具体地，我们在事件与其所属的句子、句子与其所属的文档片段、文档片段与其所属的文档之间连接属于类型的边，在文档与其包含的文档片段、文档片段与其包含的句子、句子与其包含的事件之间连接包含类型的边。

文档中所有事件之间的时序关系被映射到上述图中事件节点之间的边上。在本章节所进行的篇章层级事件时序关系抽取研究中，我们使用TDDiscourse数据集作为研究对象。此数据集中的五种时序关系类型：before、after、simultaneous、include、is_include分别对应事件之间的五种时序边的类型。每篇文档所标注的每一条事件时序关系被对应为图上的一条以其时序类别为标签的有向带权边。在标注时，对于一组事件A和B，TDDiscourse数据集仅标注出事件A相对于事件B的时序关系，而由于事件时序关系的对称性，我们可以推断出事件B相对于事件A的时序关系为其相反关系。为了将文档中的事件之间的时序关系尽可能全面地映射到图中，我们将这些依据对称性可以推断的时序关系也添加到图中。我们将构建得到的图称之为篇章层级事件时序图。

此外，对于篇章层级事件时序图中的所有节点，我们都添加了以其自身做为起点、以其自身做为终点的self loop边。最终，我们的文档层级时序图包含四种类型的节点：事件、句子、文档片段、文档，八种类型的边：属于、包含、对应五种时序类别的五种时序边以及self loop边。

我们使用图注意力神经网络在所构建的篇章层级事件时序图之上进行建模，以进行篇章层级的事件时序关系抽取，最终，图注意力神经网络输出每个事件X对应的编码结果G_X。

如公式（4-11）所示，我们同样将事件A、B的由图神经网络输出的向量表达G_A和G_B以及它们的和与逐点积进行拼接，输入到一个多层感知机MLP_{gat}之中进行事件时序类别的预测，得到基于图注意力神经网络预测出的事件时序关系类别概率分布P_{ptmAB}^{r}。

$$P_{gatAB}^{r} = MLP_{gat}\left(cat\left(G_A, G_B, G_A + G_B, G_A \odot G_B\right)\right) \tag{4-11}$$

为了使得图注意力神经网络学习到篇章中的事件之间时序关系的相互约束，给出满足事件时序一致性的预测，在训练阶段，对于训练集中的每一篇文档，我们将此文档的所有事件时序标注依据比例 $\delta:(1-\delta)$ 随机划分为两个不相交的子集 ω_D 和 $\omega_D^{'}$。子集 ω_D 中的事件时序标注用于构建上述的文档层级时序图中事件之间的时序边，同时用于微调 RoBERTa 预训练语言模型的参数；$\omega_D^{'}$ 中的事件时序标注则只用于训练图注意力神经网络的参数。一方面，这使得图注意力神经网络学习如何依据已知的部分事件对之间的时序关系，并综合篇章中不同层级的信息，推断未知的事件对之间的时序关系，另一方面，这使得图注意力神经网络学习做出与已有的事件时序关系相合的时序关系预测，避免与已有的时序关系发生冲突。

最终，在训练阶段，子集 ω_D 中的事件时序标注对应的损失函数如公式（4-12）所示，而子集 $\omega_D^{'}$ 中的事件时序标注对应的损失函数如公式（4-13）所示，其中 R 是事件 A 与事件 B 之间真实的时序类别标签。最终用于模型训练的损失函数为这两个损失函数的加和。训练阶段 HDTRG 模型的结构如图 4-1 所示。

$$L_{ptm}\left(R\middle|A,B,\theta\right)=\sum_{AB\in\omega_D}-log\frac{P_{ptm\,AB}^{\ R}}{\sum_{r\in\{Relation\_set\}}P_{ptm\,AB}^{\ r}} \tag{4-12}$$

$$L_{gat}\left(R\middle|A,B,\theta\right)=\sum_{AB\in\omega_D^{'}}-log\frac{P_{gat\,AB}^{\ R}}{\sum_{r\in\{Relation\_set\}}P_{gat\,AB}^{\ r}} \tag{4-13}$$

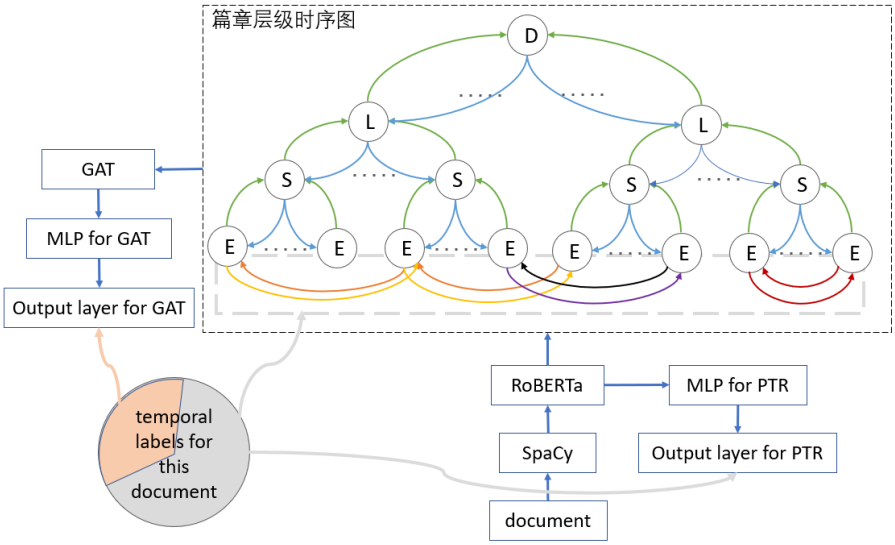

图 4-1 HDTRG 模型训练阶段的结构图示

而在测试阶段，由于没有任何一条输入文档中的事件对的时序标注，我们只能构建一副仅包含文档层级结构的图，此图中的事件节点之间没有连接任何边。为此，我们首先仅依据 RoBERTa 预训练语言模型给出的事件时序类别概率分布 $P_{ptm_{AB}}^{r}$，得到所有待预测的事件对之间的时序关系。之后，依据置信度排序，从这些预测结果中筛选出置信度较高的比例 δ 的事件时序预测。

我们采用 MC-Dropout[56]（Monte-Carlo Dropout）来衡量模型的置信度。衡量神经网络输出的置信度的最简单的方式是使用模型输出的 *softmax* 概率的高低，但是已有的研究显示模型预测的概率值高低与样本的分类结果的可靠性并不必然正相关，具有较高概率值的预测也可能并不具有较高的置信度。MC-Dropout 被证明可以近似地表征模型的置信度。具体地，在模型的预测阶段，MC-Dropout 启用神经网络中的 *dropout* 层，对同一个输入样本进行 K 次预测，从而得到 K 个预测结果。我们假设在这 K 个预测结果内，标签集合 R 中的标签 r 被预测了 N_r 次，则模型对此样本的不确定性如公式（4-14）所示。不确定性与置信度负相关，我们选择此不确定性较低的比例 δ 的事件对构建篇章层级图中的时序边。其含义事实上是这 K 个预测结果的信息熵。

$$Uncertanty = -\sum_{r \in R} \frac{N_r}{K} log\left(\frac{N_r}{K}\right) \tag{4-14}$$

接下来，我们以这些预测结果为根据，构建篇章层级事件时序图中的事件之间的边。之后，使用图注意力神经网络在此补充了事件之间的时序边的篇章层级事件时序图上预测剩余的待预测事件对之间的时序关系。测试阶段 HDTRG 模型的结构如图 4-2 所示。

此分两阶段的事件时序关系抽取策略在预测时，首先在第一阶段根据预训练语言模型给出的整合上下文信息的编码结果，做出依据局部信息足以高置信度地预测的事件对之间的时序关系类别；而在第二阶段，图注意力神经网络则综合预训练语言模型对上下文信息的编码结果和第一阶段预测出的每个事件的关联事件的时序情况，进行综合篇章层级信息的事件时序预测，确保在预测事件时序关系时，即整合了文档各层级的全局信息，又兼顾了关联事件的时序关系情况，从而做出满足事件时序一致性要求的最终预测。

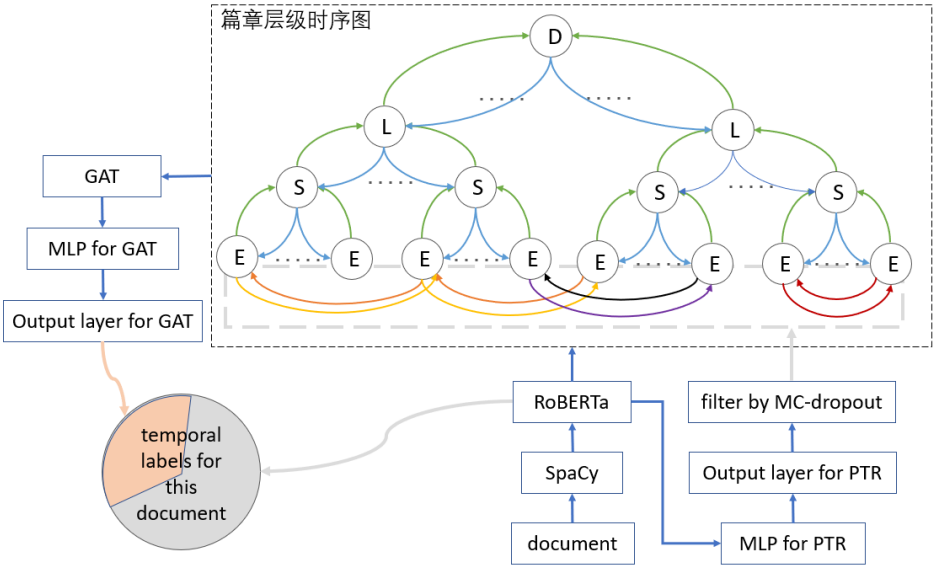

图 4-2 HDTRG 模型预测阶段的结构图示

4.4 实验验证与分析

TDDiscourse 数据集是针对篇章层级的事件时序关系抽取任务设计的数据集，我们在此数据集上进行实验来验证本章节所提出的基于篇章结构的事件时序关系抽取模型的性能。在实验中，我们设置 batch size 为 8，梯度累计步骤为 4，epoch 为 1000，选用 base 版本的 RoBERTa 预训练语言模型，在开发集上进行学习率的调优。对于标注样例的划分比例 δ，我们在{0.05、0.1、0.2、0.3、0.5}中进行调参，最终的实验结果显示，当 δ 为 0.2 时取得最优的实验结果。对于图注意力神经网络的层数，我们在{1、3、5}中进行调优，最终采用的图注意力神经网络的层数为 3。

在 TDDiscourse 数据集上的实验结果如表 4-1 所示。其中，Majority 模型[5]简单地将所有事件对之间的时序关系类别预测为比例最大的时序类别；CAEVO 模型[9]是已知的采用启发式规则所能构建的性能最优的事件时序关系抽取模型；SP 模型[57]则在传统词向量的基础上使用机器学习模型 SSVM（Structure SVM）进行时序分类；SP+ILP 模型[57]在 SP 模型预测的时序类别概率的基础上使用 ILP（整数线性规划）来强制模型做出符合事件时序一致性的预测；BiLSTM 模型[12]在传统词向量的基础上使用 BiLSTM 来编码句子中的信息；BERT-base Transformer 和 RoBERTa-base 分别使用 BERT 预训练语言模型和 RoBERTa 预训练语言模型，在句子层级进行逐对的事件时序关系预测[50]；UCGraph+BERT 模型[58]以文档中的事件作为节点，以事件之间的时序关

系作为边构建文档的时序图，在此时序图之上使用 R-GCN[59]（Relation Graph Convolutional Networks）学习同一篇文档中的所有时序关系之间的相互依赖；TIMERS 模型[50]为每篇文档构建依存句法图、语篇修辞关系图、时间参数图共三幅图，在这三幅图之上应用 GR-GCN（Gated Relation Graph Convolutional Network）来挖掘篇章层级的时序信息。

表 4-1 HDTRG 模型在 TDDiscourse 数据集上的实验结果

模型	TDD-Man	TDD-Auto
Majority	37.1	33.2
CAEVO(2014)	16.1	42.5
SP(2017)	22.7	43.2
SP+ILP(2017)	23.8	46.1
BiLSTM(2017)	24.3	51.8
BERT-base Transformer	37.5	62.3
RoBERTa-base	37.1	61.6
UCGraph+BERT(2021)	43.4	61.2
TIMERS(2021)	45.5	71.1
HDTRG	49.19±0.84	73.92±1.78

从实验结果中可以看出，我们所提出的 HDTRG 模型通过构建篇章的层级图整合篇章中各层级的全局信息，通过构建事件时序图整合与当前事件相关的事件的时序关系情况，获得了最优的篇章层级事件时序关系抽取性能。

为了验证我们所提出的 HDTRG 模型各模块的作用，我们设计了消融实验，实验结果如表 4-2 所示。其中，HDG 模型只保留了 HDTRG 模型的篇章层级时序图中的篇章层级图，而舍弃了事件节点之间的所有时序边；相反，TRG 模型只保留了 HDTRG 模型中的事件节点以及事件节点之间的时序边，缺少篇章层级结构。

表 4-2 HDTRG 模型在 TDDiscourse 数据集上的消融实验结果

模型	TDD-Man	TDD-Auto
HDG	47.55±1.47	70.13±0.64
TRG	49.00±3.62	71.08±1.68
HDTRG	49.19±0.84	73.92±1.78

可以看出，HDTRG 模型所采用的篇章层级时序图中的篇章层级部分和事件时序部分都有一定的作用。HDTRG 模型相对于 HDG 模型的性能提升说明在预

测当前事件对的时序关系时，综合考虑当前事件的关联事件的时序情况能够提高对当前事件对预测的准确率。对应地，HDTRG 模型相对于 TRG 模型的性能提升说明整合文档中各层级的时序信息对于进行篇章层级的事件时序关系抽取不可或缺。

4.5 本章小结

事件时序关系抽取模型不只需要整合篇章中的全局时序信息，还需要在预测每一组事件对的时序关系时，统筹考虑关联事件的时序情况，做出满足事件时序一致性要求的预测。为了整合文档中的全局信息，我们提出的基于篇章结构的事件时序关系抽取模型 HDTRG 首先使用 RoBERTa 预训练语言模型每次编码尽可能长的文本片段，之后为每篇文档构建事件-句子-文档片段-文档的层级图。为了整合关联事件的时序关系，我们在所构建的图中的事件节点之间连接对应事件之间的时序关系的时序边,得到篇章层级事件时序图。最后，使用图注意力神经网络在此有向带权图之上进行篇章层级的整合全局时序信息的事件时序关系抽取。最终的实验表明，我们的模型可以有效地在不同层级的文档元素之间传递全局的时序信息，同时参考关联事件的时序关系情况，实现更好的篇章层级事件时序关系抽取性能。

结　论

理解文本中事件之间的时序关系对于文本时间线的构建、时间感知的文本摘要等任务具有显著的意义。然而，自然语言处理任务中广泛使用的编码方案，包括传统词向量和预训练语言模型的自监督训练中都缺乏与时间信号相关的任务，这使得以其为基础构建的模型在编码阶段不能很好地感知时序信号，限制了应用于下游事件时序关系抽取任务时的性能。

本文探究了面向篇章理解的事件时序关系抽取技术，即挖掘文档中不同层级的时序信息，进行事件间时序关系的预测。本文主要进行了如下三部分研究工作：

（1）基于句法结构的事件时序关系抽取方法。已有的研究显示位于最短依存路径上的词对于事件间关系的预测至关重要。我们首先通过使用 BiLSTM 对最短依存路径上的词进行额外编码，验证了时序关系抽取任务中最短依存路径上的词的重要性。之后设计了依据到最短依存路径的距离给句子中的词打分，并通过此分数限制预训练语言模型中 Transformer 结构的注意力分布的方法，在不引入额外参数的条件下，实现软编码式的以事件为中心的句法结构信息嵌入。实验结果显示了此方法对于依赖句子局部信息确定时序关系的事件对的有效性。

（2）基于信息增强的事件时序关系抽取方法。事件时序关系抽取任务的瓶颈之一在于相关数据集的规模普遍较小。为此，我们首先尝试通过同义词替换的方式扩充训练数据，然而实验结果显示此方法获取的额外训练数据复杂程度有限，模型在其上训练获取的收益不高。随后，我们在时序常识知识库 TemProb 之上训练时序常识知识编码，并拼接到原有的事件编码之上，实验结果显示了此方法对于依赖世界知识确定时序关系的事件对的有效性。最后，为了充分利用每一条训练数据，我们在模型中增加了事件相对时间预测子任务和时序关系对比学习子任务来辅助模型的训练，实验结果显示这两个子任务都能为事件时序关系抽取模型带来进一步的性能提升。

（3）基于篇章结构的事件时序关系抽取方法。良好的事件时序关系抽取模型必须保证对文档中所有事件之间时序关系的预测结果满足时序一致性的要求，这需要模型在预测时不只考虑当前句子中的局部信息，还需要整合文档中的全局信息，并在预测当前事件对时考虑关联事件的时序关系情况。为此，我们首先为每篇文档构建事件-句子-文档片段-文档的层级图以整合文

档中各层级的时序信息。之后在事件之间连接对应时序关系的边得到最终的篇章层级事件时序图，最后使用图注意力神经网络在此有向带权图上学习文档中所有事件之间的时序关系的相互限制，使得模型在预测每一条时序关系时考虑其关联事件的时序关系。实验结果显示，相比于只在句子层级进行时序关系抽取的模型和已有的在篇章层级进行时序关系抽取的模型，此方法能够获得一定的性能提升。

最后，受限于计算资源，我们的研究在 RoBERTa-base 的基础上开展，后续的工作将尝试使用规模更大、能够一次编码更长文本的预训练语言模型如 XLNet 作为编码器。此外，当前最新的研究尝试通过不使用传统的向量来表达事件以更好地在模型中引入时序一致性限制。例如，使用双曲空间中的点来表达事件，则双曲空间中点之间天然的树形层级结构能够反映事件之间的时序关系；使用概率框格来表达事件则能够确保模型对事件时序关系的预测满足对称性的要求。后续我们将探究在我们的模型中采用这类事件表达方式的有效性。

参考文献

[1] Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., & Radev, D. R. (2003a). TimeML: Robust specification of event and temporal expressions in text. In New directions in Question Answering, Vol. 3, pp. 28–34.

[2] UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., & Pustejovsky, J. (2013). Semeval-2013 Task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In Proceedings of the Joint Conference on Lexical and Computational Semantics and the International Workshop on Semantic Evaluation (*SEM-SemEval), Vol. 2, pp. 1–9. ACL.

[3] Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. SemEval-2015 Task 5: QATEMPEVAL - evaluating temporal information understanding with question answering. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pages 792–800.

[4] Taylor Cassidy, Bill McDowell, Nathanel Chambers,and Steven Bethard. 2014. An annotation frame-work for dense event ordering. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pages 501–506.

[5] Naik A, Breitfeller L, Rose C. TDDiscourse: A Dataset for Discourse-Level Temporal Ordering of Events[C]//20th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2019: 239.

[6] Ning Q, Wu H, Roth D. A Multi-Axis Annotation Scheme for Event Temporal Relations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1318-1328.

[7] Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende.2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In Proceedings of the 4th Workshop on Events: Definition, Detection, Coreference, and Representation. pages 51–61.

[8] Tim O'Gorman, Kristin Wright Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In Proceedings of the 2nd Work shop on Computing News Storylines

(CNS 2016). Association for Computational Linguistics, Austin, Texas, pages 47–56.

[9] Chambers N, Cassidy T, McDowell B, et al. Dense event ordering with a multi-pass architecture[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 273-284.

[10]Goyal T, Durrett G. Embedding Time Expressions for Deep Temporal Ordering Models[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4400-4406.

[11]Zhao X, Lin S T, Durrett G. Effective Distant Supervision for Temporal Relation Extraction[C]//Proceedings of the Second Workshop on Domain Adaptation for NLP. 2021: 195-203.

[12]Cheng F, Miyao Y. Classifying temporal relations by bidirectional lstm over dependency paths[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017: 1-6.

[13]Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[14]Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

[15]S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997.

[16]Meng Y, Rumshisky A, Romanov A. Temporal information extraction for question answering using syntactic dependencies in an lstm-based architecture[J]. arXiv preprint arXiv:1703.05851, 2017.

[17]Ballesteros M, Anubhai R, Wang S, et al. Severing the Edge between before and after: Neural Architectures for Temporal Ordering of Events[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 5412-5417.

[18]Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[19]Wang H, Tan M, Yu M, et al. Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1371-1377.

[20]Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.

[21]Ning Q, Yu Z, Fan C, et al. Exploiting Partially Annotated Data in Temporal Relation Extraction[C]//Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. 2018: 148-153.

[22]Han R, Ning Q, Peng N. Joint Event and Temporal Relation Extraction with Shared Representations and Structured Prediction[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 434-444.

[23]Ning Q, Feng Z, Wu H, et al. Joint Reasoning for Temporal and Causal Relations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2278-2288.

[24]Wang H, Chen M, Zhang H, et al. Joint Constrained Learning for Event-Event Relation Extraction[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 696-706.

[25]Zhou B, Ning Q, Khashabi D, et al. Temporal Commonsense Acquisition with Minimal Supervision[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7579-7589.

[26]Eliyahu Kiperwasser and Miguel Ballesteros. 2018.Scheduled multi-task learning: From syntax to translation. Transactions of the Association for Computational Linguistics, 6:225–240

[27]Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction (ace) program-tasks, data, and evaluation[C]//Lrec. 2004, 2(1): 837-840.

[28]Ning Q, Wu H, Peng H, et al. Improving Temporal Relation Extraction with a Globally Acquired Statistical Resource[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 841-851.

[29]Ning Q, Subramanian S, Roth D. An Improved Neural Baseline for Temporal

Relation Extraction[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 6203-6209.

[30]Wen H, Qu Y, Ji H, et al. Event Time Extraction and Propagation via Graph Attention Networks[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 62-73.

[31]Beltagy I, Peters M E, Cohan A. Longformer: The Long-Document Transformer[J].

[32]Han R, Hsu I H, Yang M, et al. Deep Structured Neural Network for Event Temporal Relation Extraction[C]//Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). 2019: 666-106.

[33]Thomas Finley and Thorsten Joachims. 2008. Training structural svms when exact inference is intractable. In Proceedings of the 25th International Conference on Machine Learning, ICML ’08, pages 304–311, New York, NY, USA. ACM

[34]Weiyi Sun. 2014. Time Well Tell: Temporal Reasoning in Clinical Narratives. PhD dissertation. Department of Informatics, University at Albany, SUNY.

[35]Inderjeet Mani, Ben Wellner, Marc Verhagen, and James Pustejovsky. 2007. Three approaches to learning tlinks in time-ml. Technical Report CS-07–268, Computer Science Department.

[36]王俊,史存会,张瑾,俞晓明,刘悦,程学旗.融合上下文信息的篇章级事件时序关系抽取方法[J/OL].计算机研究与发展:1-9[2021-07-08]

[37]Cheng F, Asahara M, Kobayashi I, et al. Dynamically Updating Event Representations for Temporal Relation Classification with Multi-category Learning[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020: 1352-1357.

[38]Cho K, Van Merriënboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. arXiv preprint arXiv:1409.1259, 2014.

[39]Petar Velickovic, Guillem Cucurull, Arantxa Casanova,Adriana Romero, Pietro Liò, and Yoshua Bengio.2018.Graph attention networks. In 6th

International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May3, 2018, Conference Track Proceedings. OpenReview.net.

[40]Meng Y, Rumshisky A. Context-aware neural model for temporal information extraction[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 527-536.

[41]Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. CoRR, abs/1410.5401.

[42]He K, Chen X, Xie S, et al. Masked Autoencoders Are Scalable Vision Learners[J].

[43]Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6077-6086.

[44]Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical Text-Conditional Image Generation with CLIP Latents[J]. arXiv e-prints, 2022: arXiv: 2204.06125.

[45]Saharia C, Chan W, Saxena S, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding[J].

[46]Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[47]Sarzynska Wawer J, Wawer A, Pawlak A, et al. Detecting formal thought disorder by deep contextualized word representations[J]. Psychiatry Research, 2021, 304: 114135.

[48]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[49]Loper E, Bird S. NLTK: The Natural Language Toolkit[C]//Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. 2002: 63-70.

[50]Mathur P, Jain R, Dernoncourt F, et al. TIMERS: Document-level Temporal Relation Extraction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2021: 524-533.

[51]He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.

[52]Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.

[53]Grill J B, Strub F, Altché F, et al. Bootstrap your own latent-a new approach to self-supervised learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 21271-21284.

[54]Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[55]Li Q, Han Z, Wu X M. Deeper insights into graph convolutional networks for semi-supervised learning[C]//Thirty-Second AAAI conference on artificial intelligence. 2018.

[56]Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]//international conference on machine learning. PMLR, 2016: 1050-1059.

[57]Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1027–1037, Copenhagen, Denmark. Association for Computational Linguistics.

[58]Liu J, Xu J, Chen Y, et al. Discourse-Level Event Temporal Ordering with Uncertainty-Guided Graph Completion[J].

[59]Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling Relational Data with Graph Convolutional Networks[J].

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] 栗扬帆,张宇.面向事件时序关系抽取任务的事件表达优化方法.CCL2022 已投出

（二）参与的科研项目及获奖情况

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《面向篇章理解的事件时序关系抽取技术研究》，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：栗扬帆　　日期：2022 年 06 月 16 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　栗扬帆　　日期：2022 年 06 月 16 日

导师签名：　　日期：2022 年 06 月 16 日

致　谢

夏季总是别离的季节，六年前的八月辞别父母，第一次离开他们的怀抱，踏足外面的世界。而今时节，窗外又是蛙声一片。不久之后，又将辞别在哈尔滨工业大学结识的每一位师长、朋友，步入职场。感慨良多。韶华易逝，青春如歌，在象牙塔中最后的两年研究生生涯终要奏起尾声。

感谢我的导师张宇教授，感谢张老师在三年前认可我加入社会计算与信息检索研究中心问答组。我性格散漫，做事拖延，张老师的敦促和殷切使我在懒散之时能够及时收回自己，投心学习之中，不断进步。在研究过程中，张老师为我的研究方向查阅了大量资料，在我研究卡壳、自我怀疑的时候坚定了继续研究的信心和决心。在生活中，张老师关心每一位学生的日常生活，在每一件小事中培养我们良好的生活习惯，注意安全，讲究卫生，尊敬师长，谈吐有度。在问答组的三年也刚好和疫情的三年重合，不论是假期还是在校期间，张老师都一再提醒我们疫情防护的细节，关心我们每一个人的身心健康情况。祝张老师广育英才，人生幸福美好。

感谢三年来实验室提供的舒适安静的学习环境和充沛广足的实验资源，让我能够顺利而安心地开展实验，验证自己的猜想，浅尝科研的乐趣。感谢课题组提供的宝贵项目经历，在参与项目过程中获取的经验不只在去年秋招阶段发挥了重要作用，也将内化为自己的经验，在之后的工作、生活中继续让自己受益。祝实验室在今后取得更加长足的发展。

感谢齐乐、施琦、赵雅欣、妥明翔、乐远、尹治博、蒋润宇等实验室的师兄，感谢他们对我在文献阅读和实验设计上的指导，他们对待学习和科研认真务实的态度使我受益良多。感谢实验室的孙月晴、侯鹏钰、严未圻、汤嘉琦、陈卓旎、张梓寒、崔涵、杨昕、宋岩奇等同学，和他们一起学习进步的每一天都倍感充实，一起放松娱乐也很快乐。祝他们在今后的日子里学业有成，事业顺遂。

感谢室友卢延悦、田一间、张开颜、张家乐、廖其鑫、宋锦文、张少锋，祝福我们今后都有光明的未来。特别感谢张开颜同学，感谢他与我一起准备秋招的面试笔试，一起讨论分享各自的学习研究，交换想法。

感谢从去年十一月到今天为止，每一位和我联系的故友新朋，谢谢他们在我最困厄的时候的倾听和陪伴。感谢曾和郭恺铭彻夜清谈时共枕的床榻，感谢曾和李国飞兴起畅谈时齐望的明月。特别感谢侯均杰近乎随叫随到的陪

伴。感谢王哲培、栗子茂、王玉慧、胡林慧等朋友在那段时光里对我的鼓励和肯定。祝福他们每一个人都能够拥抱幸福的人生。也感谢这几个月来阅读的每一本书的作者，他们用有深度也有温度的文字启发我反思和内省，亦安慰和平抚我的思绪。

最后感谢父母一直以来的支持和鼓励，他们永远是我最坚实的后盾。他们的辛勤让我在十数年的求学路途上没有后顾之忧，得以潜心学习，探索整个世界。祝福他们永远身体健康，万事如意。

我相信“爱的能力取决于审美的能力”，我相信“一切知识都是为了心动”，我相信一百八十亿年前那场物质和能量的邂逅的全部意义和价值正在于我们真诚的相信，而“努力工作及爱人”就是对此最好的验证。

余生，结蕙为纕，揽茝作冠，兰皋纵马，椒丘驰车。