

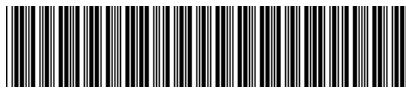


200233

发文日：

上海桂平路 435 号 上海专利商标事务所有限公司
施浩(021-34183200)

2025 年 03 月 22 日



申请号：202210112284.8

发文序号：2025032200024860

申请人：上海金融期货信息技术有限公司

发明创造名称：金融产品价格影响因素传导分析方法及系统

第一次审查意见通知书

1. 应申请人提出的实质审查请求，根据专利法第 35 条第 1 款的规定，国家知识产权局对上述发明专利申请进行实质审查。

根据专利法第 35 条第 2 款的规定，国家知识产权局决定自行对上述发明专利申请进行审查。

2. 申请人要求以其在：

申请人已经提交了经原受理机构证明的第一次提出的在先申请文件的副本。

申请人尚未提交经原受理机构证明的第一次提出的在先申请文件的副本，根据专利法第 30 条的规定视为未要求优先权要求。

3. 经审查，申请人于_____提交的修改文件，不符合专利法实施细则第 57 条第 1 款的规定，不予接受。

4. 审查针对的申请文件：

原始申请文件。 分案申请递交日提交的文件。 下列申请文件：

5. 本通知书是在未进行检索的情况下作出的。

本通知书是在进行了检索的情况下作出的。

本通知书引用下列对比文件(其编号在今后的审查过程中继续沿用)：

编号	文 件 号 或 名 称	公开日期 (或抵触申请的申请日)
1	“面向金融领域的事理图谱构建关键技术研究”，廖阔，《中国优秀硕士学位论文全文数据库信息科技辑》，第 2 期，I138-2530 页	2021-02-15
2	CN113887230A	2022-01-04

6. 审查的结论性意见：

关于说明书：

申请的内容属于专利法第 5 条规定的不授予专利权的范围。

说明书不符合专利法第 26 条第 3 款的规定。

说明书不符合专利法第 33 条的规定。

说明书的撰写不符合专利法实施细则第 20 条的规定。



国家知识产权局

关于权利要求书：

- 权利要求_____不符合专利法第2条第2款的规定。
 权利要求_____不符合专利法第9条第1款的规定。
 权利要求_____不具备专利法第22条第2款规定的新颖性。
 权利要求1-12不具备专利法第22条第3款规定的创造性。
 权利要求_____不具备专利法第22条第4款规定的实用性。
 权利要求_____属于专利法第25条规定的不授予专利权的范围。
 权利要求_____不符合专利法第26条第4款的规定。
 权利要求_____不符合专利法第31条第1款的规定。
 权利要求_____不符合专利法第33条的规定。
 权利要求_____不符合专利法实施细则第22条的规定。
 权利要求_____不符合专利法实施细则第23条的规定。
 权利要求_____不符合专利法实施细则第24条的规定。
 权利要求_____不符合专利法实施细则第25条的规定。

申请不符合专利法第26条第5款或者实施细则第29条的规定。

申请不符合专利法第19条第1款的规定。

申请不符合专利法实施细则第11条的规定。

分案申请不符合专利法实施细则第49条第1款的规定。

上述结论性意见的具体分析见本通知书的正文部分。

7. 基于上述结论性意见，审查员认为：

申请人应当按照通知书正文部分提出的要求，对申请文件进行修改。

申请人应当在意见陈述书中论述其专利申请可以被授予专利权的理由，并对通知书正文部分中指出的不符合规定之处进行修改，否则将不能授予专利权。

专利申请中没有可以被授予专利权的实质性内容，如果申请人没有陈述理由或者陈述理由不充分，其申请将被驳回。

8. 申请人应注意下列事项：

(1) 根据专利法第37条的规定，申请人应在收到本通知书之日起的4个月内陈述意见，如果申请人无正当理由逾期不答复，其申请被视为撤回。

(2) 申请人对其申请的修改应当符合专利法第33条的规定，不得超出原说明书和权利要求书记载的范围，同时申请人对专利申请文件进行的修改应当符合专利法实施细则第57条第3款的规定，按照本通知书的要求进行修改。

(3) 申请人的意见陈述书和/或修改文本应邮寄或递交国家知识产权局专利局受理处，凡未邮寄或递交给受理处的文件不具备法律效力。

(4) 未经预约，申请人和/或代理师不得前来国家知识产权局专利局与审查员举行会晤。

(5) 对进入实质审查阶段的发明专利申请，在第一次审查意见通知书答复期限届满前（已提交答复意见的除外），主动申请撤回的，可以请求退还50%的专利申请实质审查费。

9. 本通知书正文部分共有4页，并附有下述附件：

引用的对比文件的复印件共1份69页。

审查员：和弦

联系电话：028-62968436

审查部门：专利审查协作四川中心



210401
2023.03

纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路6号 国家知识产权局专利局受理处收
电子申请，应当通过电子专利申请系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。



第一次审查意见通知书

申请号:2022101122848

本申请请求保护金融产品价格影响因素传导分析方法及系统。经审查，现提出如下审查意见：

权利要求1-12不符合专利法第22条第3款有关创造性的规定

1、权利要求 1 请求保护一种金融产品价格影响因素传导分析方法。对比文件 1 (“面向金融领域的事理图谱构建关键技术研究”,廖阔,《中国优秀硕士学位论文全文数据库信息科技辑》,第 2 期,I138–2530 页,2021 年 2 月 15 日) 为最接近的现有技术,具体公开了(参见第 5 章):

我们基于第二至第四章中介绍的事件因果关系抽取、事件表示学习、因果关系强度计算等关键技术,设计了如图 5–1 所示的事理图谱构建流程,并实现了面向金融领域的事理图谱构建系统,下面具体介绍该系统中的每个模块。

(1) 数据获取

我们基于 Scrapy 工具包编写了爬虫,从腾讯、网易、股吧、和讯等十余个网站的金融板块爬取了新闻、研究报告等金融领域的文本,目前共收集了 25G 的语料数据,包含 1200 多万篇文档,作为事理图谱的数据来源。该数据获取模块目前仍在持续运行,不断地从互联网上获取新的金融领域文本,对事理图谱数据进行补充。

(2) 数据预处理

我们首先对互联网上爬取的金融语料进行清洗,以去除其中的标记语言与链接等。之后,我们对语料中的文本进行归一化,其中包括将除小数点外的标点符号统一为中文,将连续的空自字符合并为一个,将全角的数字及英文字符统一为半角等。我们进一步使用 LTP 自然语言处理工具包对语料进行预处理,使用用 LTP 的分句工具将原始语料切分为句子,并对其进行分词、词性标注与依存句法分析。预处理步骤得到的结果与语料一起保存在文件中,共后续处理步骤使用。(相当于对自动获取的原始数据进行预处理)

(3) 事件因果关系抽取

我们使用第二章提出的基于噪声模型的半监督学习方法训练事件因果关系抽取模型,并将训练好的模型应用于事理图谱的构建。爬取好的金融领域文本经预处理及分词后输入事件因果关系抽取模型,模型对输入单词序列生成符合 BIO 标注规范的标签序列,最后由该标签序列解码得到原因事件短语与结果事件短语。我们将抽取出的原因、结果事件与因果对所在的上下文文本一起保存在事理图谱中。(相当于基于预处理后的数据抽取事件因果对,基于时间因果对构建事理图谱)

(4) 事件相似度计算

(5) 事件泛化

(6) 因果关系强度计算



我们为构建好的金融事理图谱搭建了可视化系统进行展示。

权利要求 1 请求保护的技术方案与对比文件 1 的区别在于：基于语言预训练模型的构建分类算法对步骤 2 所构建的事理图谱中的因果事件对自动赋予标签，基于语言预训练模型自动赋予的标签用于价格影响因素的传导分析。由此可以确定，相比于最接近的现有技术，权利要求 1 请求保护的技术方案实际要解决的技术问题是提供一种对金融数据进行分类的方法。

对于上述区别技术特征，对比文件 2 (CN113887230A) 公开了一种面向金融场景的端到端自然语言处理训练框架与方法，具体公开了（参见说明书第 64–68 段）：本发明还提供了一种面向金融场景的端到端自然语言处理训练方法，方法包括以下步骤：步骤 1、FinBERT 预训练；步骤 2、基于类似 self-training 思想从外部相关数据中挖掘新数据，以扩充下游任务的语料总量；步骤 3、下游任务语料上进行预训练，将步骤 2 所获得的外部语料与原始的下游任务语料共同构成扩充后的下游任务语料库，并用来对 FinBERT 进行再一次预训练，得到的模型称之为 TASK FinBERT；步骤 4、用半监督学习的框架来充分利用无标签语料，在 TASK FinBERT 的基础上，利用半监督学习的框架，在下游任务语料库上训练得到的模型称之为 UDA FinBERT；步骤 5、蒸馏学习，将学习到的知识和特征蒸馏到轻量级模型上。由此可见，对比文件 2 给出了基于语言预训练模型构建分类算法对金融领域任务进行自然语言处理任务的技术教导，本领域技术人员基于对比文件 1 为了对金融数据进行标签处理与分类有动机采用上述方法，根据实际使用需求能够将其用于对构建的事理图谱的因果事件对自动赋予标签以用于价格影响因素的传导分析。

综上，在对比文件 1 的基础上结合对比文件 2 和本领域常规技术手段，获得该权利要求请求保护的技术方案对本领域技术人员而言是显而易见的，其不具有突出的实质性特点和显著的进步，不具备专利法第 22 条第 3 款规定的创造性。

2、从属权利要求2–6对其引用的权利要求进一步限定。

对于权利要求2，对比文件1公开了编写爬虫从网站金融板块爬取包括新闻、研究报告等金融领域的文本，对其进行清洗、归一化、分词、词性标注等预处理。在此基础上，原始数据可根据实际需要采用分布式爬虫系统自动从互联网抓取，去重属于常规数据预处理手段。

对于权利要求3，对比文件1公开了（参见第2.1节）可以先根据因果触发词定位出潜在的因果对，再有针对性地抽取出该因果对涉及的原因与结果事件。因此，本课题将探索基于序列标注模型的事件-因果关系联合抽取方法，同时在文本中标注出原因事件与结果事件，以提升事件因果关系抽取的效率（相当于基于预处理后的数据抽取触发词及所在句，抽取原因句与结果句并确定事件表达构建事理图谱）；由对比文件1还可知在因果关系抽取方法上，基于规则的因果抽取方法依然在使用中占据主要地位（参见第1.3节）。因此本领域技术人员能够根据实际抽取质量与成本等需求使用基于规则模板的正则匹配进行事件抽取，相应针对原因、结果子句集合抽取出显式因果语句中的原因子句和结果子句，基于依存句法分析确定事件表达对于本领域技术人



员而言是容易做到的。

对于权利要求4，对比文件2公开了（参见说明书第95段）：在步骤4中，运用半监督学习框架微调二次预训练后的TASK FinBERT。通常一个下游任务中，有很大一部分语料都不具有标签，因此在此阶段，本发明实施例引入半监督学习框架Unsupervised Data Augmentation(UDA)。核心目标是，在不引入额外人工标注情况下，充分利用无标签的数据，用来指导模型的学习。UDA是Google在2019年提出的半监督学习算法。该算法超越了所有之前的半监督学习方法，并实现了仅使用极少量标记样本结合大量无标签样本即可达到甚至超过使用大量标记样本训练集的模型精度。UDA框架的核心思想是一致性假设，即要求模型在输入数据的附近空间应该是平坦的，即使输入数据发生微弱变化或者发生语义不变而仅仅是形式变化时，模型的输出也能够基本保持不变。从数据流角度来看，UDA框架获取学习信号的核心思想，即对于有标签部分的数据仍然采用交叉熵，而对于无标签部分的数据则通过一致性正则来获取学习信号。本发明实施例在TASKFinBERT的基础上，利用半监学习框架，在下游任务相关语料上训练得到的模型称之为UDAFinBERT。由此可见，对比文件2给出了对有标签的文本进行有监督学习，对无标签的文本进行一致性训练，针对无标签数据和有标签数据使用半监督学习的学习框架进行训练的技术教导，得到损失函数作为模型目标优化函数用于训练属于本领域常规技术手段。

对于权利要求5，数据增强是有监督学习中常规技术手段，可根据数据处理实际需求选用。

对于权利要求6，对比文件1公开了为构建好的金融事理图谱搭建了可视化系统进行展示，由图5-2可知其构建了以事件为节点，事件关系为边并存入图数据库的可视化效果。

因此，在引用的权利要求不具备创造性的情况下，上述权利要求也不具备专利法第22条第3款规定的创造性。

3、权利要求 7 请求保护一种金融产品价格影响因素传导分析系统，其是与方法权利要求 1 中各步骤相对应的产品权利要求。对比文件 1 为最接近的现有技术，具体公开内容如前所述。该权利要求请求保护的技术方案与对比文件 1 的区别在于：权利要求 1 与对比文件 1 的不同，设置了实现相应功能的模块。由此可以确定，相比于最接近的现有技术，该权利要求请求保护的技术方案实际要解决的技术问题是提供一种对金融数据进行分类的方法。对于上述区别技术特征，参见权利要求 1 的相关评述，根据实际功能需求不难设置对应模块。综上，在对比文件 1 的基础上结合对比文件 2 和本领域常规技术手段，获得该权利要求请求保护的技术方案对本领域技术人员而言是显而易见的，其不具有突出的实质性特点和显著的进步，不具备专利法第 22 条第 3 款规定的创造性。

4、从属权利要求8-12对其引用的权利要求进一步限定。参见前述权利要求2-6的相关评述，根据实现功能实际需求不难设置对应模块。因此，在引用的权利要求不具备创造性的情况下，上述权利要求也不具备专利法第22条第3款规定的创造性。



国家知识产权局

基于上述理由，本申请的权利要求都不具备创造性，同时说明书中也没有记载其他任何可以授予专利权的实质性内容，因而即使申请人对权利要求进行重新组合和／或根据说明书记载的内容作进一步的限定，本申请也不具备被授予专利权的前景。如果申请人不能在本通知书规定的答复期限内提出表明本申请具有创造性的充分理由，本申请将被驳回。

审查员姓名:和弦
审查员代码:30141376



国家知识产权局

检索报告

申请号：2022101122848	申请日：2022年01月29日	首次检索	
申请人：上海金融期货信息技术有限公司	最早的优先权日：		
权利要求项数：12	说明书段数：124+6		
审查员确定的 IPC 分类号：G06Q 40/04,G06F 16/36,G06N 5/02,G06F 40/289,G06F 40/268,G06F 40/211,G06F 40/186,G06F 16/951,G06N 20/00			
检索记录信息：CN113887230A: 713 CNTXT, (标签 OR label?) and 半监督 AND 金融; 语义排序,语义基准:2022101122848 CN111222648A: CNTXT 语义检索,语义基准:步骤 3.1: 对有标签的文本进行有监督学习; 步骤 3.2: 使用无监督学习框架对无标签的文本进行一致性训... CN113934909A: CNTXT 语义检索,语义基准:2022101122848 CN113590824A: 60 CNTXT, 金融 and 因果 and 事理图谱; 语义排序,语义基准:2022101122848 CN110727803A: 引文关联浏览 面向金融领域的事理图谱构建关键技术研究:CNKI, TKA/ (金融 and 事理图谱) a financial derivatives related multi-label text classification algorithm based on financial knowledge graph:追踪发明人 一种半监督学习的金融新闻文本分类算法:CNKI, TKA/ (半监督 and 标签 and 文本) https://zhuanlan.zhihu.com/p/151021586 :互联网检索, 金融 半监督			

相关专利文献

类型	国别以及代码[11] 给出的文献号	代码[43]或[45] 给出的日期	IPC 分类号	相关的段落 和 / 或图号	涉及的权 利要求
Y	CN113887230A	2022-01-04	G06F40/295	说明书第 64-68、95 段	1-12
A	CN111222648A	2020-06-02	G06N20/00	全文	1-12
A	CN113934909A	2022-01-14	G06F16/951	全文	1-12



国家知识产权局

A	CN113590824A	2021-11-02	G06F16/35	全文	1-12
A	CN110727803A	2020-01-24	G06F16/36	全文	1-12

相关非专利文献					
类型	书名(包括版本号和卷号)	出版日期	作者姓名和出版者名称	相关页数	涉及的权利要求
类型	期刊或文摘名称 (包括卷号和期号)	发行日期	作者姓名和文章标题	相关页数	涉及的权利要求
Y	中国优秀硕士学位论文全文 数据库信息科技辑,第02期	2021-02-15	廖阔,面向金融领域的事 理图谱构建关键技术研 究	I138-2530	1-12
T	proceedings of SPIE	2023-12-31	MA H 等,a financial derivatives related multi-label text classification algorithm based on financial knowledge graph	12636	1-12
T	大数据,第08卷,第02期	2022-03-18	张晓龙等,一种半监督学 习的金融新闻文本分类 算法	134-144	1-12
类型	网址	网络发布日 或公开日	作者姓名和网页标题	相关部分	涉及的权利要求



国家知识产权局

A	https://zhuanlan.zhihu.com/p/ 151021586	2020-07-05	李渔,半监督学习在金融 文本分类上的探索和实 践	全文	1-12
---	--	------------	--------------------------------	----	------

表格填写说明事项：

1. 审查员实际检索领域的 IPC 分类号应当填写到大组和 / 或小组所在的分类位置。
2. 期刊或其它定期出版物的名称可以使用符合一般公认的国际惯例的缩写名称。
3. 相关文件的类型说明：

X：单独影响权利要求的新颖性或创造性的文件；

Y：与本检索报告中其他 Y 类文件组合后影响权利要求的创造性的文件；

A：背景技术文件，即反映权利要求的部分技术特征或者有关的现有技术的文件；

R：任何单位或个人在申请日向专利局提交的、属于同样的发明创造的专利或专利申请文件。

P：中间文件，其公开日在申请的申请日与所要求的优先权日之间的文件，或者会导致需要核实该申请优先权的文件；

E：单独影响权利要求新颖性的抵触申请文件；

T：申请日或优先权日当天或之后公布的，可以对所要求保护发明的理论或原理提供清楚解释的文件，或者可显示出所要求保护发明的推理或事实不成立的文件；

L：除 X、Y、A、R、P、E 和 T 类文件之外的文件。

审 查 员：和弦

2025 年 03 月 18 日

审查部门：专利审查协作四川中心

硕士学位论文

面向金融领域的事理图谱构建关键技术研究

**RESEARCH ON KEY TECHNOLOGIES OF
FINANCIAL DOMAIN-ORIENTED EVENTIC
GRAPH CONSTRUCTION**

廖阔

哈尔滨工业大学

2020 年 6 月

国内图书分类号：TP391.1
国际图书分类号：004.8

学校代码：10213
密级：公开

工学硕士学位论文

面向金融领域的事理图谱构建关键技术研究

硕士研究生：廖阔
导师：黄虎杰教授
申请学位：工学硕士
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

RESEARCH ON KEY TECHNOLOGIES OF FINANCIAL DOMAIN-ORIENTED EVENTIC GRAPH CONSTRUCTION

Candidate:

Liao Kuo

Supervisor:

Prof. Huang Hujie

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

传统的知识图谱大多关注实体的属性与关系知识，而忽视了事件间的演化规律知识，为了弥补这一不足，研究者们提出了事理图谱的概念。事理图谱中的节点是高度泛化的事件，边是事件间的演化关系，例如因果关系与顺承关系。自然语言处理技术的发展使得信息抽取的准确率得到提升，也使得从文本中自动挖掘事理知识、构建事理图谱成为可能。本文从金融领域入手，对事理图谱自动构建的关键技术展开研究，具体内容包括端到端的事件因果关系抽取、常识信息增强的事件表示学习以及数据驱动的因果关系强度计算。

事理知识的获取是构建事理图谱的基础，本文针对事件间因果关系的获取进行了探索。本文将事件因果关系抽取建模为序列标注任务，提出了基于预训练模型的因果抽取方法，以端到端的方式同时进行因果关系的识别与相关事件的抽取。为缓解有标注数据不足的问题，本文进一步提出基于噪声模型的半监督学习方法，利用大量无标注数据提升因果抽取的效果。中英文两个因果抽取数据集上的实验结果证明了该方法的有效性。

事件是事理图谱的核心元素，为了更好地建模事件语义，本文提出了常识信息增强的事件表示学习方法，使学习到的事件表示中融入意图、情感、实体关系等常识信息，以更好地帮助事理图谱的构建以及在其他任务上的应用。事件相似度、脚本事件预测、股市预测等多个任务上的实验结果表明我们的方法可以更准确地建模事件语义，并提升下游任务上的效果。

为了更好地建模事件间因果关系的强度，本文探索了基于统计与基于预训练模型的因果强度计算方法，从大量因果事件对中自动学习因果强度信息。COPA 因果推理数据集上的实验结果表明，预训练模型可以有效地从大量因果事件对中学习因果知识，并准确地建模因果关系强度。

最后，本文基于上述研究成果设计并实现了金融领域事理图谱构建系统，并在大规模金融语料上构建了包含数百万事件与因果关系的事理图谱，验证了本文提出的事理图谱构建方法的可行性。

关键词：事理图谱；因果关系抽取；事件表示学习；因果强度

Abstract

Traditional knowledge graphs mostly center on attribute and relation knowledge of entities, but ignores the knowledge of evolutionary patterns between events. In order to make up for this deficiency, researchers gradually put forward the concept of eventic graph. The nodes in eventic graph are highly generalized events, and the edges are the evolutionary relationships between events, such as causal relationship and sequential relationship. With the development of natural language processing technology, the accuracy of information extraction has been improved, making it possible to automatically mine eventic knowledge and build the eventic graph from the text. Starting from the financial domain, this paper explores the key technologies of financial domain-oriented eventic graph construction, including the extraction of event causality based on sequence labeling, the commonsense-enhanced event representation learning and the data-driven method for causal strength calculation.

The acquisition of event knowledge is the basis of eventic graph construction. This paper explores the acquisition of causal relationship between events. In this paper, event causality extraction is modeled as a sequence labeling task, and a method of causality extraction based on pre-trained model is proposed. To alleviate the problem of insufficient labeled data, this paper further proposes a semi-supervised learning method based on noise model, which uses a large number of unlabeled data to improve the effect of causal extraction. The experimental results on two causal extraction datasets in Chinese and English prove the effectiveness of this method.

Event is the core concept of eventic graph. In order to model event semantics better, this paper proposes a commonsense-enhanced event representation learning method, which integrates commonsense information such as intent, sentiment and entity relationship into the learned event representation, to help the construction of event graph and its application in other tasks. The experimental results on event similarity, script event prediction, stock market prediction and other tasks show that our method can model event semantics more precisely and improve the results on downstream tasks.

Towards better modeling the causal strength between events, this paper explores the calculation of causal strength based on statistical method and pre-trained models, which automatically learn causal strength information from a large number of causal event pairs. The experimental results on COPA causal inference dataset show that pre-trained models can effectively learn causal knowledge from a large number of causal event pairs, and accurately model the strength of causal relationship.

Finally, based on the above researches, this paper designs and implements an

Abstract

eventic graph construction system in financial domain, and constructs a eventic graph that contains millions of events and causal relationships from large-scale financial corpus. The experiment verified the feasibility of eventic graph construction methods proposed in this paper.

Keywords: eventic graph, causality extraction, event representation learning, causality strength

目 录

摘要	I
ABSTRACT	II
第1章 绪 论	1
1.1 课题来源及研究的背景和意义	1
1.1.1 课题来源	1
1.1.2 课题研究的背景和意义	1
1.2 国内外在该方向的研究现状及分析	2
1.2.1 事理知识库的构建及应用	2
1.2.2 因果关系抽取	3
1.2.3 事件表示学习	4
1.3 国内外文献综述的简析	5
1.4 主要研究内容	6
1.5 本文章节安排	7
第2章 端到端的事件因果关系抽取	9
2.1 引言	9
2.2 任务定义	10
2.3 基于预训练模型的有监督事件因果关系抽取	10
2.4 基于自训练的半监督事件因果关系抽取	14
2.5 实验	17
2.5.1 实验数据	17
2.5.2 实验设置	18
2.5.3 实验结果	19
学习，从而为CRF输出层带来提升。	21
2.6 本章小结	21
第3章 常识信息增强的事件表示学习	22
3.1 引言	22
3.2 任务定义	23
3.3 基于张量神经网络的事件表示学习	23
3.4 常识信息增强的事件表示学习	25
3.4.1 融合实体关系信息的事件表示学习	25

3.4.2 融合意图信息的事件表示学习	26
3.4.3 融合情感信息的事件表示学习	27
3.4.4 融合实体关系、意图、情感的事件表示联合学习框架	27
3.5 实验	29
3.5.1 实验数据	29
3.5.2 实验设置	30
3.5.3 实验结果	31
3.6 本章小结	34
第4章 数据驱动的因果关系强度计算	35
4.1 引言	35
4.2 任务定义	35
4.3 基于统计的因果关系强度计算	36
4.4 基于预训练模型的因果关系强度计算	37
4.5 实验	39
4.5.1 实验数据	39
4.5.2 实验设置	40
4.5.3 实验结果	43
4.6 本章小结	46
第5章 金融事理图谱构建系统的设计与实现	47
5.1 引言	47
5.2 系统介绍与结构设计	48
5.3 系统展示	49
5.4 本章小结	51
结 论	52
参考文献	54
攻读硕士学位期间发表的论文及其它成果	59
哈尔滨工业大学学位论文原创性声明和使用权限	60
致 谢	61

第1章 绪 论

1.1 课题来源及研究的背景和意义

1.1.1 课题来源

事件演化的事理逻辑规律是一种非常重要且普遍的知识，这种知识对于事件的预测和推理等任务具有重要的作用。然而，现有的知识图谱、语义网络等知识库往往是以实体为核心的，关注实体的属性与实体间的关系，而忽略了事件演化的事理逻辑知识。为了弥补事理逻辑相关研究的缺失，哈尔滨工业大学社会计算与信息检索研究中心提出了事理图谱的概念，事理图谱是一种以事理逻辑为核心的知识库，形式上为有向图，图中的节点表示事件，有向边表示事件之间的演化关系，用于刻画和记录人类行为活动和事件客观演化规律。随着自然语言处理和信息抽取等技术的发展，从无结构文本自动挖掘知识、构建知识图谱的方法日趋成熟，这些方法同理可应用到事理图谱的自动构建中。本课题从金融领域着手，旨在探索事理图谱自动构建中的关键技术，以提升自动构建的事理图谱质量，并为开放域事理图谱的自动构建提供参考。

1.1.2 课题研究的背景和意义

随着深度学习等技术的兴起，人工智能迎来了新的发展高潮。人工智能发展的一个重要问题在于让机器掌握人类知识，例如，人类能够轻易掌握“吃过饭”后就“不饿”这一常识知识，而让机器理解并掌握大量这样的知识是一件很困难的事情，这也是人工智能从感知走向认知的必由之路。

在众多类型的人类知识中，事理逻辑是另一种非常重要且普遍存在的知识，许多人工智能应用依赖于对事理逻辑知识的深刻理解。例如在对话系统中，只有让机器理解“吃过饭”之后“人不饿了”，“看电影”之前要“先买票”这样的常识事理，对话系统才能根据不同语境作出更有逻辑性的回复。在金融领域，新闻中的事件对股市涨跌的预测具有重要意义，如果可以挖掘出“粮食减产”导致“农产品价格上涨”，再导致“通胀”，进而导致“股市下跌”这样的远距离事件依赖，对于事件驱动的股市预测非常有价值。

事件是人类社会的核心概念之一，人们的社会活动往往是事件驱动的。事件之间在时间上相继发生的演化规律和模式是一种非常有价值的知识，挖掘这种事理逻辑知识对认识人类行为和社会发展变化规律有十分重要的意义。然而，当前的知识图谱、语义网络等知识库均以实体为研究对象，聚焦于实体的属性

与实体间的关系，而缺乏对事理逻辑知识的挖掘。为了揭示事件演化规律和发展模式，哈尔滨工业大学社会计算与信息检索研究中心提出了事理图谱的概念，旨在将事件的演化规律和模式构建成一个有向图形式的事理知识库，其中节点表示事件，有向边表示事件之间的演化关系，用于刻画和记录人类行为活动和事件客观演化规律。

随着自然语言处理和信息抽取等技术的发展，从无结构文本中自动挖掘知识并构建知识图谱成为了新的研究热点。这一技术使知识图谱的构建不再完全依赖人类专家，使知识图谱的构建从繁重的人类劳动中解放出来。在金融领域中，存在着大量质量较高、易于获取的财经新闻，以及由专业人士撰写的财报等，可以期望从这些文本中挖掘出高质量的事理逻辑知识。本课题从金融领域着手，旨在探索事理图谱自动构建的关键技术，包括事件因果关系的抽取、因果关系强度的建模等，以充分挖掘金融文本中的事理逻辑知识，以期以自动化的方式构建较高质量的金融事理图谱，并为通用域事理图谱构建提供参考。

1.2 国内外在该方向的研究现状及分析

1.2.1 事理知识库的构建及应用

事理知识是现实世界中一种重要的知识形式，近年来，已经有一些学者围绕事理知识展开了研究，尝试构建事理知识库，并利用事理知识库取得了下游任务的提升。这些工作主要围绕事件间的顺承关系和因果关系展开，并关注抽象、泛化的事件表示形式，并在因果推理、事件预测、股市预测等任务上取得了引人瞩目的效果。

2016 年，Luo 等人^[1]从文本中抽取因果事件对，并构建了词级别的因果网络 CausalNet。该网络中的节点为单词，若两个单词分别出现同一因果对的原因事件和结果事件中，则网络中两单词间存在一条有向边。该工作提出了数据驱动的因果强度计算方法，由构建的因果网络计算词对的因果强度，并从“必要性”“充分性”两个方面对因果强度进行建模。这一方法在 COPA^[2]因果推理数据集(给定一个事件，选择正确的原因或结果)上取得了当时最佳的效果。

2017 年，Zhao 等人^[3]提出使用抽象事件间的因果关系解决下游任务。该工作首先从新闻标题中抽取满足因果关系的具体事件，并使用 VerbNet^[4]和 WordNet^[5]对事件短语中的动词和名词进行泛化，得到抽象的事件表示，由此构建抽象事件网络，并提出 Dual-CET 模型在网络中学习事件表示，在事件预测、事件聚类、股市预测等任务上取得了较好的效果。

2017 年，Li 等人^[6]首次在论文中正式提出事理图谱的概念，将事理图谱

定义为描述事件之间演化规律和模式的事理逻辑知识库，形式为有向有环图，图中的节点表示抽象事件，定义为泛化、语义完备的谓词短语或片段；图中的有向边表示事件之间的顺承关系与因果关系。该工作探索了事件间顺承关系与顺承关系方向的识别方法，提出了以顺承关系为核心的事理图谱构建框架。该方法首先从文本中抽取事件，再使用分类器判断事件间是否存在顺承关系以及顺承关系的方向。作者使用该框架在中文出行领域语料上构建了事理图谱，包含约3万个事件节点，23万个顺承关系。

2018年，Li等^[7]提出从叙事事件链条中构建事理图谱，并应用于脚本事件预测任务中。脚本事件预测任务定义为给定一系列上下文事件，要求从若干个候选事件中选择后续最有可能发生的事件。该工作从文本中自动抽取事件链条，并从中构建事理图谱，使用可扩展的图神经网络（SGNN）在事理图谱上学习事件表示，在该任务上取得了当时最佳的效果。

1.2.2 因果关系抽取

事件间的因果关系是一种重要的知识类型，相对于一般的顺承关系，因果关系强调事件演化的本质规律，从而避免了事件的偶然共现或表面上的相关性带来的噪声。1.2.1节中的工作验证了因果知识在因果推理、事件预测、股市预测等任务上的有效性。获取因果关系的“黄金标准”是进行随机对照实验，但是这种方法常常耗费过高或根本不可行^[8]。另一种方法是从文本中抽取出因果关系，文本中包含了人们对现实世界因果规律的叙述和总结，尤其是许多专业领域文本中包含很多对事件前因后果的分析，因此可以期望通过信息抽取技术从文本中自动挖掘因果关系。

因果关系自动抽取的研究起源于20世纪90年代^[9]，早期研究围绕基于规则的方法展开。Grishman（1990）等人^[10]提出了PROTEUS工具，使用语法和语义信息自动抽取文本中的时序关系和因果关系。Kaplan等人（1991）^[11]将文本表示为命题的集合，每个命题包含一个谓词（通常为动词）和多个论元，通过定义命题模板的方式抽取命题中的因果关系。Garcia（1997）^[12]提出了COATIS工具，该工具使用包含23个因果性动词的语言学模板自动从法语文本中抽取因果关系。Chan,K等人（2005）^[13]提出了基于语义期望的因果抽取系统SEKE，该系统使用了语义模板、句子模板、原因模板、后果模板四种不同层次的模板信息。Girju（2002）^[14]等人提出使用WordNet信息帮助自动发现表达因果含义的动词，用于抽取“NP1-Verb-NP2”形式的因果关系，其中NP为名词短语，Verb为动词。

2000年后，因果关系抽取研究逐渐转移到基于统计与机器学习的方法。

Girju (2003)^[15]提出了使用 C4.5 决策树^[16]判断“NP1-Verb-NP2”元组是否构成因果关系。Chang 等人 (2004)^[17]使用无监督的方法从文本中学习短语指示因果关系的概率与词对出现在因果关系中的概率,作为额外特征训练贝叶斯分类器判断因果关系是否成立。Blanco 等人 (2008)^[18]使用 C4.5 决策树^[16]结合 Bagging 方法^[19]进行因果关系识别。2007 年, SemEval Task 4^[20]中提出了其中常见语义关系分类任务,其中包含因果关系,Girju 等人^[21]使用 SVM 在该任务上取得了当时最佳的效果。Sil 等人 (2010)^[22]提出了 PREPOST 系统,使用基于 RBF 核的 SVM 结合 PMI 特征判断因果关系是否成立。付剑锋等人 (2011)^[23]提出使用层叠条件随机场的方法,在事件序列上同时标注多个因果关系对。Zhao 等人 (2016)^[24]提出基于因果连接词类别信息的受限隐朴素贝叶斯模型提升因果抽取任务的效果。Tharini 等人 (2017)^[25]将 CNN 应用于因果关系抽取任务中。Kruengkrai 等人 (2017)^[26]进一步使用 Multi-Column CNN 并结合外部知识进行因果关系抽取。这一系列的研究将因果关系抽取建模为分类任务,输入是已经抽取好的事件或实体及其上下文,判断它们是否构成因果关系,而并未关注原因、结果事件或实体本身的抽取工作。

随着自然语言处理技术的发展,一些研究者开始尝试使用序列标注模型解决因果关系抽取问题。Dasgupta 等人 (2018)^[27]提出将因果关系抽取建模为序列标注任务,在文本中直接标注出原因提及与结果提及的短语。该工作使用双向 LSTM^[28]作为序列标注模型,并加入额外的语言学特征提升该任务上的效果。Dunietz 等人 (2018)^[29]提出了 DeepCx 模型用于浅层语义分析,并在因果关系抽取任务上进行了测试。Li 等人 (2019)^[30]提出使用带有自注意力机制的双向 LSTM-CRF 模型,结合经过领域迁移的词向量应用于因果抽取任务。基于序列标注模型的因果抽取方法是一种端到端的方法,可以直接由文本得到因果事件或实体对,为因果关系抽取研究指明了新的方向。

1.2.3 事件表示学习

从文本中可以挖掘出具体事件间的因果关系,但具体事件往往是非常稀疏的,抽取的两个因果对中包含完全相同事件的可能性非常小,因此很难将因果对连接成稠密的网络。另一方面,具体事件间的因果关系也较难用在预测和推理中,因为目标事件很可能没有在因果网络中出现过,这样就无法进行预测和推理。因此,需要对抽取的具体事件进行抽象和泛化,从具体事件间的因果上升到抽象事件间的因果,从而发现更为一般的因果律。对事件进行抽象和泛化的一个有效方法是发现具有相似关系的事件集合,再从这些相似事件中提取出事件的公共成分。为克服事件的稀疏性,需要将事件嵌入到低维的向量空

间中，由此提出了事件表示学习的任务。

最简单的事件表示方式是对事件论元的词向量进行“加性”组合。Li 等人（2018）^[7]使用事件论元的词向量拼接作为事件表示输入后续网络；Weber 等人（2018）^[31]使用词向量均值，以及词向量拼接后输入神经网络的方法作为基线方法。这类方法中，来自不同论元的信息以加法运算进行组合，使相同词较多的事件具有较为相似的事件表示。

2015 年，Ding 等人^[32]提出了基于张量神经网络(NTN)的事件表示模型。该模型的输入为（主语，谓语，宾语）形式的事件三元组，模型使用双线性张量运算分别组合主语与谓语、谓语与宾语的词向量，最后再用双线性张量运算组合前面得到的两个中间向量。作者使用该模型学习事件表示，结合卷积神经网络从事件表示中提取特征，在股市预测任务上取得了很好的效果。

2016 年，Ding 等人^[33]提出将外部知识融入事件表示的学习中。该工作在 NTN 模型的基础上，引入来自知识图谱的实体知识，并通过将知识图谱中的实体与事件主体映射到同一个向量空间中，使事件表示融入来自知识图谱的知识，并在股市预测任务上取得了当时最佳的效果。

2018 年，Weber 等人^[31]提出将场景信息建模到事件表示中，体现为以预测文本中周围事件作为模型的训练目标。该工作依然基于张量组合计算事件表示，但提出了两个更为简单的模型以减少参数的数量，并在脚本事件预测和两个事件相似度任务上取得了较好的结果。

1.3 国内外文献综述的简析

事理逻辑知识作为现实世界中一种重要的知识类型，正受到越来越多研究者的关注。2.1 节中围绕事理逻辑的研究主要围绕事件间的因果关系和顺承关系展开，并且主要关注抽象事件而非具体事件间的因果和顺承关系，这为后续研究指明了方向。然而，这些研究各自有其局限性。Luo 等人（2016）^[1]将事件拆分为词，存储单词在因果对中的共现关系，虽然解决了事件稀疏问题，但破坏了事件原有的结构，引入了过多的噪声，同时也使人难以介入到知识库的构建中。Zhao 等人（2017）^[3]与 Li 等人（2018）^[7]基于特定任务构建事理知识库，该知识库可能无法很好地泛化到其他任务，并且这两个工作均使用简单的方法抽取事件与关系，没有在知识库的构建上做太多探索。Li 等人（2017）^[6]探索了顺承事理图谱的构建方法，并构建了一个大规模的、独立于任务的出行领域事理图谱，但缺少对因果关系抽取的探索。因此，本课题提出以因果关系为核心的事理图谱构建方法，以期能填补相关工作在这一方向上的空白。

在因果关系抽取方法上，基于规则的因果抽取方法依然在使用中占据主要

地位，但这一系列方法依赖人工设计的抽取规则，需要耗费较大的人力，并且容易过拟合于领域数据，无法在抽取质量和泛化能力上取得较好的平衡。基于机器学习的因果抽取方法将因果抽取建模为分类任务，即先由其他步骤如句法分析、事件抽取等识别出潜在的原因和结果短语，再使用分类器判断短语间是否存在因果关系，这一系列的方法在实践中体现为流水线机制，即原因、结果提及的抽取和因果关系的判断是分开进行的，导致的一个缺点是容易受到级联错误限制。另外，这些方法本身没有关注原因、结果提及的抽取质量。近年来，随着自然语言处理和人工智能的发展，基于序列标注的端到端因果抽取方法逐渐被提出（Dasgupta 等人，2018^[27]；Li 等人，2019^[30]），这些方法同时进行因果关系的发现以及原因、结果提及的抽取，在实用性上大大超越基于规则的方法与基于分类的方法，为因果关系抽取研究指明了新的方向。

基于文本中抽取的因果事件对构建事理图谱，需要对抽取出的事件进行抽象和泛化，其中一个简单可行的方法是将事件表示为低维稠密的向量，以此发现彼此相似的事件集合，由此引入事件表示学习任务。简单的“加性”方法如词向量均值、通过简单神经网络组合等仅仅将事件论元的词向量做加法组合，难以避免“相同词多的事件表示相似”这一问题。Ding 等人（2015）^[32]与 Weber 等人（2018）^[31]提出的基于张量神经网络的方法对事件论元做“乘性”组合，使模型能够捕捉到输入特征中的微小区别，但需要合适的训练目标使模型能够正确捕捉到有鉴别的差异。Ding 等人（2016）^[33]提出将实体知识图谱信息融入事件表示学习中，将实体关系的学习作为辅助训练目标，为事件表示学习指出了新方向。人类主观层次的常识信息，如意图、情感等，也可为事件表示学习提供额外训练目标，以期取得更好的效果。

1.4 主要研究内容

本课题主要研究面向金融领域的事理图谱构建关键技术，探索如何从大规模自然语言文本中准确、高效地获取事件演化规律知识，构建事理知识库，具体研究内容包括端到端的事件因果关系抽取、常识信息增强的事件表示学习、数据驱动的因果关系强度建模以及金融领域事理图谱构建系统的设计与实现。

构建事理图谱的第一步是获取事件演化规律知识，为了从自然语言文本中挖掘这种知识，本文使用 BERT 模型以序列标注的形式端到端地抽取事件及事件间的因果关系。为了缓解该任务上有标注数据不足的问题，本文使用自训练的方法，同时利用少量有标注数据与大量自标注数据训练模型，并引入噪声模型建模自标注数据中的噪声，以缓解模型在自标注数据上学习到错误信息的情况。该方法在中、英文两个事件因果关系抽取任务上都取得了提升。

事件表示学习将事件表示为计算机可以理解的形式，是计算机处理事件相关任务的基础。要准确地理解事件的语义信息，计算机不仅需要理解事件各组成元素的字面含义，还要理解事件背后所蕴含的意图、情感等常识信息，以及事件所涉及实体背后的关系信息，而这些信息通常难以从事件的字面表示中获取。本文提出了常识信息增强的事件表示学习方法，采用多任务学习的方式，将外部知识库中的意图、情感、实体知识等信息融入事件表示模型中，并在事件相似度、脚本事件预测、股市预测三个事件相关的任务上进行了测试。实验结果表明，该方法可以有效地在事件表示中融入常识信息，并在下游任务上取得显著的提升。

事理图谱是一个有向有环图，其边上的权重可以为下游任务提供有价值的信息。对于事件间的因果关系，其边上的权重可以视为对因果强度的建模。为了更好地建模事件间的因果强度，本文探索了基于 PMI 的因果强度建模方法以及基于预训练语言模型的因果强度建模方法，并在 COPA 因果推理数据集上进行了测试。实验结果表明，在大量因果数据上预训练的语言模型能够更准确地建模事件间的因果关系强度。

基于上述研究成果，本文设计并实现了一个中文金融领域事理图谱构建系统。该系统首先使用端到端的事件因果关系抽取方法从自然语言文本中抽取因果事件对，并通过计算事件间的相似度对抽取出的具体事件进行泛化，将泛化后的事件以因果关系为边连接成图的形式，最后计算泛化后事件对的因果强度作为因果边上的权重，从而得到事理图谱。本文进一步使用该系统在中文大规模金融领域数据上构建了包含数百万事件及因果关系的事理图谱，并构建了演示系统将图谱可视化，以供用户浏览与访问。

1.5 本文章节安排

本文按照研究内容将各章节安排如下：

第 1 章，首先介绍本课题的来源、研究背景及研究意义，总结国内外在事理知识库的构建及应用、事件因果关系抽取、事件表示学习等相关方向上的研究现状，之后提出本文的研究内容。

第 2 章，首先介绍事件因果关系抽取任务的定义，之后依次介绍基于预训练模型的事件因果关系抽取方法，以及在预训练模型基础上结合自训练与噪声模型的半监督事件因果关系抽取方法。最后介绍上述方法在中、英文两个事件因果关系抽取数据集上的实验设置及实验结果，并对实验结果进行分析。

第 3 章，首先介绍事件表示学习的定义，并介绍基于张量神经网络与低

秩张量分解的事件表示学习方法。之后，介绍本文提出的融合事件意图、情感进行以及实体关系等多种外部知识的事件表示学习方法，以及在事件相似度、脚本事件预测、股市预测三个任务上的实验设置及实验结果，并对实验结果进行分析。

第4章，首先介绍因果关系强度的定义，之后依次介绍基于PMI的因果强度建模方法，以及基于预训练语言模型的因果强度计算方法。最后，我们在COPA因果推理数据集上对上述方法进行了实验，汇报了实验设置及实验结果，并对实验结果进行分析。

第5章，介绍基于上述研究成果设计与实现的中文金融领域事理图谱构建系统，并对构建好的事理图谱进行展示与分析。

本工作全文的结构框架如图1-1所示。

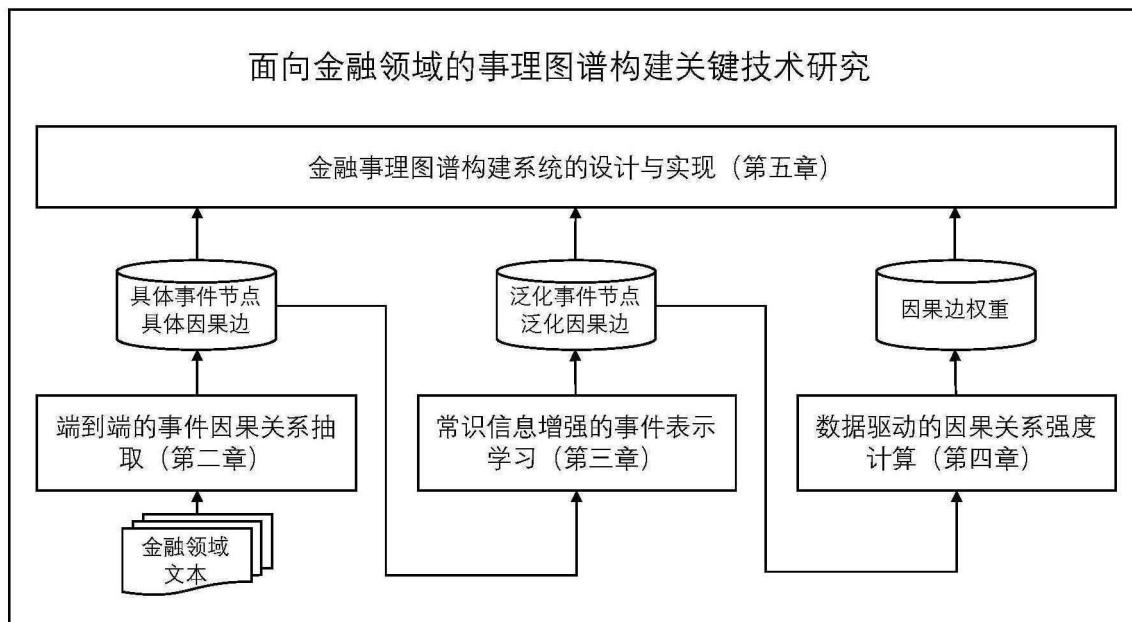


图1-1 论文结构框架

第 2 章 端到端的事件因果关系抽取

2.1 引言

构建事理图谱的首要步骤是获取大量的事件演化规律知识，即事件与事件间的关系。获取这些知识的一种方法是由人类专家手工整理，这样获取的事理知识准确度很高，但会消耗较大的人力，难以用此方法获取大规模的事理知识。而新闻、研究报告等形式的自然语言文本中往往涉及大量对事件与事件间关系的表述，蕴含着大量事件演化规律信息，因此可以运用信息抽取技术自动化地从海量文本中获取事理逻辑知识。随着自然语言处理技术的发展，预训练语言模型^{[34][35][36]}的出现大幅刷新了包括信息抽取在内的多项自然语言处理任务的最高水平，因此通过信息抽取技术从海量文本中自动化、低成本地获取大量事理逻辑知识是值得期待的。事件间的关系包含多种类型，其中因果关系是一种非常重要的关系，对未来事件的预测与推理起到重要作用，本章从因果关系这一种事件间关系出发，探索自动化抽取事件间因果关系的方法。

在传统的事件关系抽取系统中，往往先抽取单个事件，再使用分类模型判断任意两个事件之间是否存在关系。这种方法并不适用于事件因果关系的抽取，因为文本中事件间因果关系是十分稀疏的，根据粗略统计，即使在因果分析较多的研报文本中，平均一篇文档也只能抽取出少于 3 个因果对。因此，先抽取所有事件，再判断事件间因果关系的方法是不经济的。另一方面，由于因果触发词的存在，因果关系的识别往往是较为简单的，因此可以先根据因果触发词定位出潜在的因果对，再有针对性地抽取出该因果对涉及的原因与结果事件。因此，本课题将探索基于序列标注模型的事件-因果关系联合抽取方法，同时在文本中标注出原因事件与结果事件，以提升事件因果关系抽取的效率。

预训练语言模型如 BERT^[34]在命名实体识别等多个序列标注任务上取得了最佳的结果，本章首先探索基于 BERT 模型的有监督事件因果关系抽取方法。事件因果关系抽取的一个难点在于有标注数据的匮乏，限制了有监督方法的应用，因此，本章进一步提出了基于自训练的半监督事件因果关系抽取方法，充分利用大量无标注数据来弥补有标注数据的不足，并在中文、英文两份事件因果关系抽取数据上验证了上述方法的效果。

2.2 任务定义

本课题考虑从单个句子中抽取因果事件对，并假设一个句子中最多只包含一个因果事件对。本课题将事件因果关系抽取建模为序列标注任务，输入为一个具有 n 个词的句子 $S = \{w_1, \dots, w_i, \dots, w_n\}$ ，其中 w_i 是句子中的第 i 个单词，任务的输出为长度为 n 的标签序列 $T = \{t_1, \dots, t_i, \dots, t_n\}$ ，其中 t_i 与 w_i 一一对应。标签采用 BIO 标注规范， $t_i \in \{O, B - cause, I - cause, B - effect, I - effect, B - trigger, I - trigger\}$ ，具体地， $B - cause, B - effect, B - trigger$ 分别表示单词位于原因事件、结果事件、因果触发词短语的开始， $I - cause, I - effect, I - trigger$ 分别表示单词位于原因事件、结果事件、因果触发词短语的中间， O 标签表示单词不属于原因事件、结果事件或因果触发词短语。可以进一步根据标注规范，从标签序列中解码出原因事件、结果事件及因果触发词短语。

2.3 基于预训练模型的有监督事件因果关系抽取

BERT 模型由 Devlin 等人^[34]于 2018 年提出，是一种基于双向 Transformer^[37]编码器的预训练语言模型，在阅读理解、信息抽取等多项自然语言处理任务上取得了最佳结果。BERT 的输入为一个单词序列，并通过堆叠多个双向 Transformer 层，为序列中的每个单词计算一个融合上下文信息的向量表示。基于 BERT 的事件因果关系抽取方法首先使用 BERT 为句子中每个单词计算上下文相关的向量表示，之后将这些向量输入输出层模型，从中解码出句子对应的标签序列。本节首先介绍 BERT 模型的原理，之后介绍两种输出层模型结构。

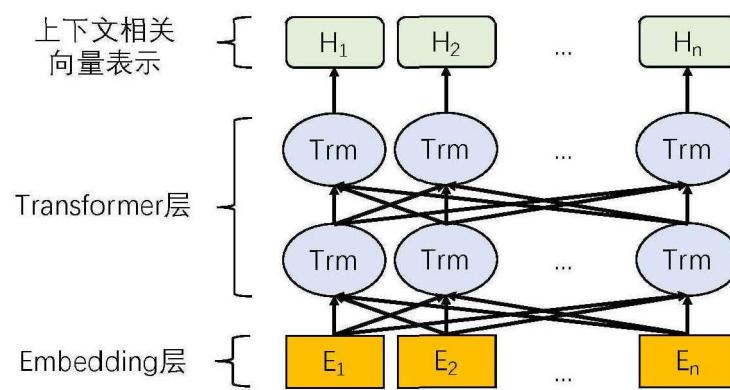


图 2-1 BERT 模型结构

BERT 由一个 Embedding 层与多个双向 Transformer 层堆叠而成，其模型结构如图 2-1 所示。待编码的文本经过 WordPiece 算法^[38]进行分词，在每段待编码的文本后添加特殊的分隔符[SEP]并拼接在一起，并在第一段文本前添加特殊

的符号[CLS]。经过上述处理的文本输入 Embedding 层得到初始的向量表示。Embedding 层由 Token Embedding, Segment Embedding 与 Position Embedding 三部分组成, 如图 2-2 所示, 其中 Token Embedding 是每个单词上下文无关的向量表示, Segment Embedding 用于在输入包含多段文本时区分每一段文本, Position Embedding 用于编码单词在文本中的位置信息。Embedding 层的输出 E_i 为每个单词 w_i 的 Token Embedding, Segment Embedding 与 Position Embedding 之和。

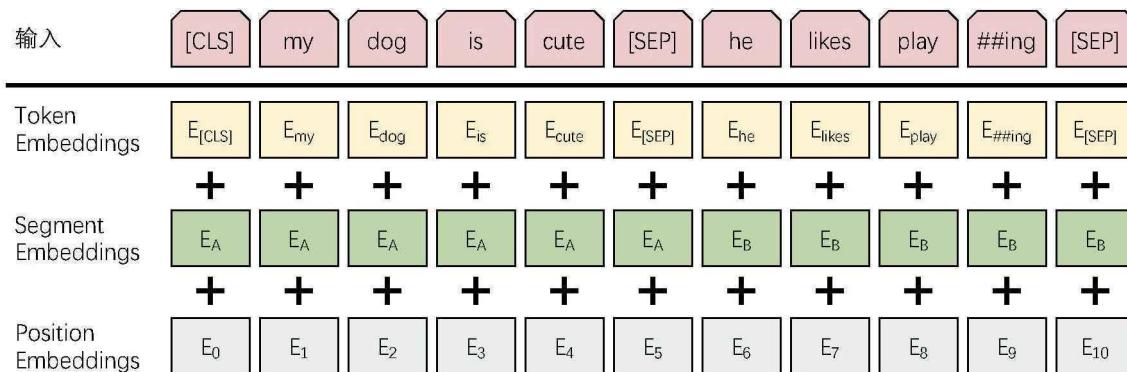


图 2-2 BERT Embedding 层结构

BERT 通过堆叠多个双向 Transformer 层为每个单词计算上下文相关的向量表示。每个 Transformer 层的输入为前一个 Transformer 层的输出, 第一个 Transformer 层的输入为 Embedding 层的输出。每个 Transformer 层由一个多头自注意力模块与一个前馈神经网络模块组成, 这两个模块分别由一个残差连接包裹, 并对其输出进行层归一化 (Layer Normalization)。Transformer 层的结构如图 2-3 所示。下面首先介绍多头自注意力模块。

注意力机制由查询 Q 、键 K 与值 V 三部分要素组成, 其中查询 Q 为一个向量, 键 K 与值 V 为 n 个向量, Q, K 与 V 的取值来自同一个集合的情况称为自注意力机制。注意力机制首先由查询 Q 与键 K 为每个值 V 计算一个权重, 再用该权重对值 V 求加权和, 得到注意力机制的输出。BERT 中使用缩放点积注意力, 其计算方式如下, 其中 d_K 为键 K 的维度:

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

多头自注意力机制通过多个不同的线性变换将 Q, K, V 映射到多个不同的向量空间中, 并在每个向量空间中分别使用自注意力机制进行计算, 每个向量空间中的计算结果称为多头自注意力的一个“头”。每个“头”进行拼接后送入一个全连接层得到多头自注意力机制的输出:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2-2)$$

$$\text{MultiHead}(Q, K, V) = \text{concat}(\text{head}_1, \dots, \text{head}_i, \dots, \text{head}_h)W^O \quad (2-3)$$

其中, h 为多头自注意力中“头”的个数, $i \in \{1, \dots, h\}$, W_i^Q, W_i^K, W_i^V, W^O 为多头自注意力的参数。在 BERT 的 Transformer 层中, Q 为前一层输出的当前单词的向量表示, K 和 V 为前一层输出的所有单词的向量表示组成的矩阵。多头自注意力机制的输出经过残差连接与层归一化后得到多头自注意力模块的输出。该输出进一步送入一个全连接层与 GeLU 激活函数, 同样经过残差连接与层归一化后得到 Transformer 层的输出。

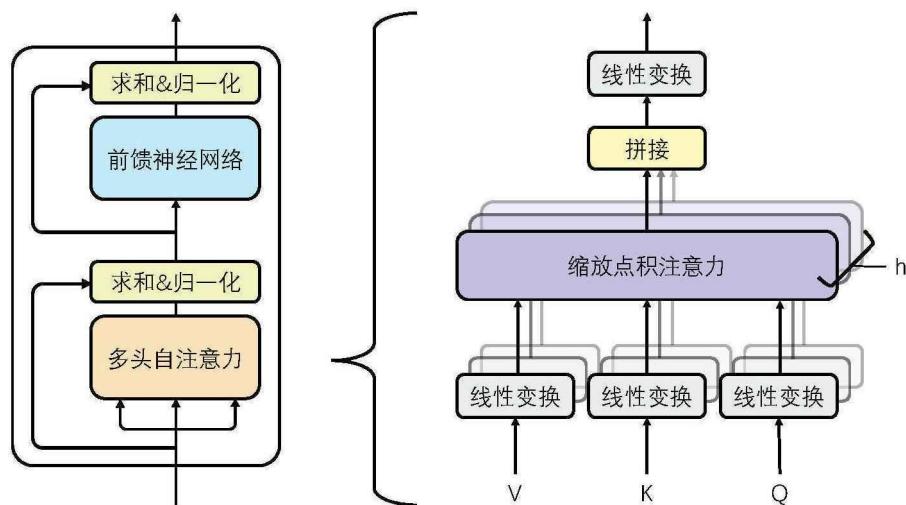


图 2-3 Transformer 层结构

我们使用 BERT 最后一个 Transformer 层的输出 H_i 作为每个单词上下文相关的向量表示, 并将其输入一个输出层模型得到单词序列所对应的标签序列。我们考虑了两种输出层模型结构: Softmax 输出层与条件随机场 (CRF) 输出层, 如图 2-4 所示。

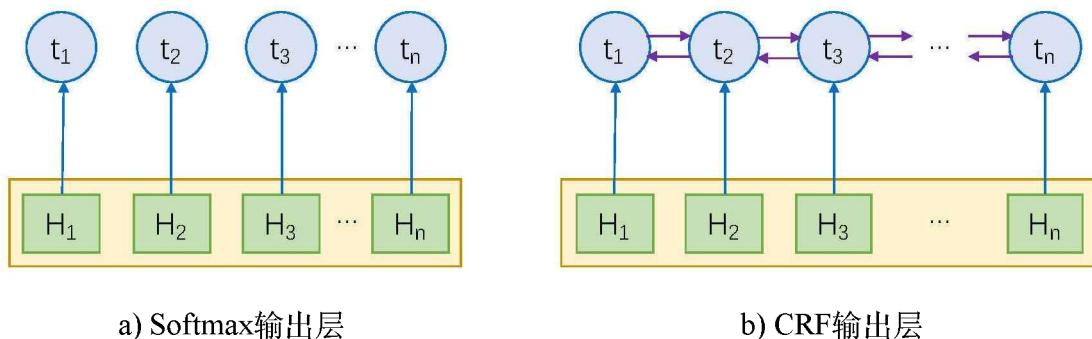


图 2-4 两种输出层模型结构

Softmax 输出层直接将每个单词的上下文相关向量表示输入一个全连接神经网络, 得到该单词属于每个标签类别的分数, 并由 Softmax 函数将这些分数

归一化为单词属于每个标签类别的概率。模型的训练目标为最小化所有单词正确标签的负对数似然，即计算所有单词标签分类的交叉熵损失。该方法将序列标注问题简化为每个单词上的标签分类问题，虽然思想十分简单，但基于 BERT 模型强大的学习能力，依然在多项自然语言处理任务上取得了较高的结果。具体计算过程如下：

$$Y_i = H_i W + b \quad (2-4)$$

$$P_i = \text{softmax}(Y_i) \quad (2-5)$$

$$L_{\text{Softmax}} = - \sum_i \log P_i^j = \sum_i \left(\log \sum_j \exp(P_i^j) - P_i^j \right) \quad (2-6)$$

其中， W 与 b 为全连接神经网络的参数， $Y_i \in \mathbb{R}^{|L|}$ 为全连接神经网络计算的单词 w_i 属于每个标签类别的分数， L 为所有标签的集合， P_i 为 Y_i 经过 Softmax 函数归一化后得到的概率， P_i^j 为单词 w_i 属于标签类别 j 的概率， j 为训练集中标注的单词 w_i 的正确标签。

Softmax 输出层只考虑了单词标签对单词向量表示的依赖关系，而未考虑不同位置上单词标签间的依赖关系，而条件随机场输出层同时考虑了两种类型的依赖关系。条件随机场^[39]是一种无向图模型，定义为给定随机变量 X 的条件下，随机变量 Y 的马尔可夫随机场。本文考虑线性链条件随机场，我们以序列标注问题中的单词序列 S 作为条件随机场中的随机变量 X ，以标签序列 T 作为条件随机场中的随机变量 Y ，则给定单词序列 S 的条件下，标签序列 T 的概率为定义为：

$$P(T|S) = \frac{1}{Z} \prod_i \psi_i(t_i|S) \quad (2-7)$$

$\psi_i(t_i|S)$ 是定义在条件随机场无向图模型中每个最大团上的势函数，定义为：

$$\psi_i(t_i|S) = \exp(\text{Emit}(t_i, w_i) + \text{Trans}(t_{i-1}, t_i)) \quad (2-8)$$

其中， $\text{Emit}(t_i, w_i)$ 是条件随机场的发射函数，建模标签对单词的依赖关系，输出给定单词 w_i 的条件下标签 t_i 的概率； $\text{Trans}(t_{i-1}, t_i)$ 是条件随机场的转移函数，建模不同位置上单词标签间的依赖关系，输出给定 $i-1$ 位置上标签为 t_{i-1} 的条件下、 i 位置上标签为 t_i 的概率。具体地，在基于 BERT 的序列标注模型中，两函数定义为：

$$\text{Emit}(t_i, w_i) = H_i W + b \quad (2-9)$$

$$\text{Trans}(t_{i-1}, t_i) = M_{t_{i-1}, t_i} \quad (2-10)$$

发射函数与 Softmax 输出层相同，仍由一个全连接神经网络建模，将 BERT 输出的上下文相关向量表示映射为单词属于每个标签类别的分数，其参数为

W, b 。转移函数由一组另外的模型参数 $M \in \mathbb{R}^{|L| \times |L|}$ 建模，其中 $M_{i,j}$ 为标签 i, j 出现在两个相邻位置上的概率。

Z 为归一化因子，确保 $P(T|S)$ 满足概率的性质。设任意可能的标签序列为 $T' = \{t'_1, \dots, t'_i, \dots, t'_n\}$ ，则归一化因子为 $\prod_i \psi_i(t'_i|S)$ 对于任意可能的 T' 求和：

$$Z = \sum_{T'} \prod_i \psi_i(t'_i|S) \quad (2-11)$$

条件随机场模型的训练目标为最小化正确标签序列的负对数似然：

$$\begin{aligned} L_{CRF} &= -\log P(T|S) \\ &= -\log \frac{\exp(\sum_i (\text{Emit}(t_i, w_i) + \text{Trans}(t_{i-1}, t_i)))}{\sum_{T'} \exp(\sum_i (\text{Emit}(t'_i, w_i) + \text{Trans}(t'_{i-1}, t'_i)))} \end{aligned} \quad (2-12)$$

在训练时，该训练目标可通过前向-后向算法高效地计算；在预测时，需要对所有可能的标签序列 T' 计算其概率，并选择概率最大的标签序列作为输出，此时可通过维特比算法高效地寻找概率最大的标签序列。

2.4 基于自训练的半监督事件因果关系抽取

事件因果关系抽取任务的一个难点在于有标注数据的匮乏，限制了有监督方法的应用。仅在少量有标注数据上训练，一方面模型难以从少量数据中学习到有用的知识，另一方面容易导致模型过拟合，使其仅在训练数据上表现很好，但在其他数据上的泛化能力很差。

半监督学习是缓解有标注数据匮乏的方法之一。半监督学习同时在有标注数据与无标注数据上训练模型，相对于有标注数据，无标注数据的获取成本往往是非常低廉的，因此可以很容易地收集到大量无标注数据，而半监督学习可以充分地利用无标注数据中的信息帮助模型取得更好的性能。

在各种半监督学习方法中，自训练^[40]是一种应用非常广泛的方法。这一方法首先在少量有标注数据上训练模型，并使用模型对无标注数据进行标注，本文中使用“自标注”代指这一步骤，用“自标注数据”代指这一步骤得到的数据。之后，自标注得到的一部分数据被加入训练集中，用于进一步对模型进行训练。通常，这一步骤选取模型在标注时置信度最高的数据加入训练集。自训练方法不断地迭代上述的训练模型、自标注、更新训练集三个步骤。自训练在自然语言处理的各项任务中均有应用，例如词义消歧^[40]、情感分析^[41]以及句法分析^[42]等。

自训练方法最明显的特点是使用模型自标注的数据反过来训练模型自身。然而，自标注数据往往存在错误，尤其是当模型只使用少量有标注数据训练时。

这些被错误标注的数据反而会使模型学习到有害的信息，甚至降低模型的性能。因此，更为合理的做法是将模型自标注的数据视为不可靠数据而非真实答案，并针对自标注数据的这一特点对自训练方法进行改进。

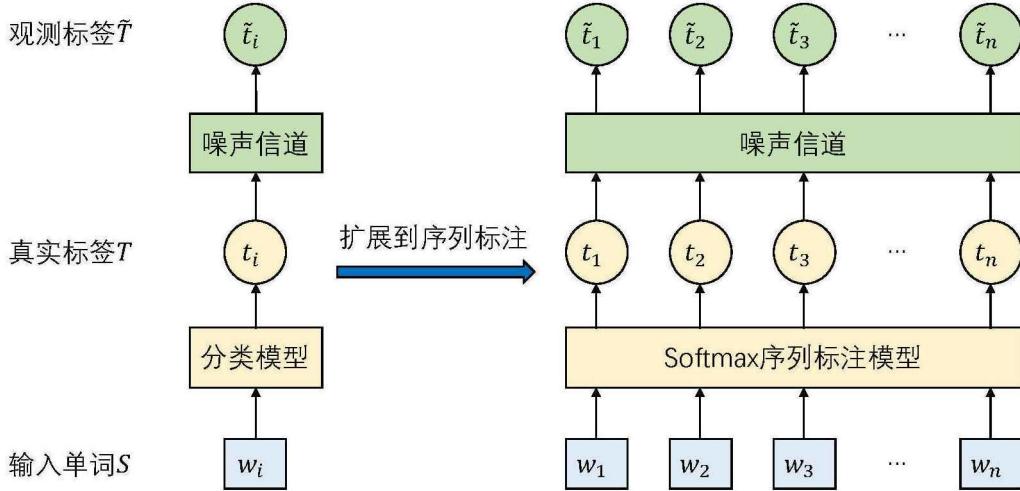


图 2-5 噪声模型示意图

Bekker 等人^[43]首先提出了使用噪声模型改进在不可靠数据上的训练过程。该工作考虑分类任务，将不可靠数据上的标签视为由真实标签经过一个噪声信道得到，并使用噪声模型对噪声信道进行建模，如图 2-5 所示。由原模型输出的单词属于每个类别的概率 $P(t_i|w_i)$ 并不直接与数据集中的标签计算损失，而是再输入噪声模型得到加入噪声的概率分布 $P(\tilde{t}_i|w_i)$ 后进行损失计算。加入噪声的概率分布可以由边缘概率求和得到： $P(\tilde{t}_i|w_i) = \sum_{t_i} P(\tilde{t}_i|t_i)P(t_i|w_i)$ 。这一方法不使用原模型输出的概率拟合数据集中的标签分布，而是使用加入噪声的概率拟合数据集中的标签分布，避免了原模型直接学习数据集中有噪声的标签，缓解了不可靠数据带来的影响。在 Bekker 等人的方法中， i 位置上标签 j 通过噪声信道变为标签 k 的概率 $\mathcal{N}_i^{NLNN}(j, k)$ 使用一个所有单词共享的噪声矩阵 N 建模，我们将这种建模噪声的方法记为 NLNN：

$$\mathcal{N}_i^{NLNN}(j, k) = N_{j,k} \quad (2-13)$$

Goldberger 等人^[44]进一步假设每个单词位置上的噪声分布是不同的，并提出使用全连接神经网络为每个单词独立地计算噪声，我们将这种建模噪声的方法记为 NLNN-Adapt，标签 j 经过噪声信道后变为标签 k 的概率计算方式如下，其中 $|L|$ 个全连接神经网络的权重组成一个三维张量 $W \in \mathbb{R}^{|L| \times |L| \times d}$ ，偏置组成矩阵 $b \in \mathbb{R}^{|L| \times |L|}$ ：

$$\mathcal{N}_i^{NLNN-Adapt}(j, k) = \text{softmax}(H_l W_{j,k} + b_{j,k}) \quad (2-14)$$

考虑到序列标注任务的特点，我们还提出了新的假设：每个单词位置上的

噪声不仅依赖于该单词，还依赖于序列中的其他单词，并提出使用循环神经网络（RNN）对噪声信道进行建模，将该方法记为 NLNN-Adapt-RNN。具体地，我们使用双向长短时记忆网络（BiLSTM）^[28]计算真实标签经过噪声信道变为错误标签的概率。双向长短时记忆网络包含两个长短时记忆网络组件，分别从前向后与从后向前地处理输入的文本序列。在每个方向上，长短时记忆网络依次读入文本序列中的每个单词的向量表示，并更新自己的隐层状态及细胞状态。以前向的长短时记忆网络为例，记其初始的隐层状态为 \vec{h}_0 ，初始的细胞状态为 \vec{c}_0 ，该网络依次读入输入序列中的每个单词 $x_0, \dots, x_{t-1}, \dots, x_n$ ，并按下列公式循环地计算其隐层状态 $\vec{h}_1, \dots, \vec{h}_t, \dots, \vec{h}_{n+1}$ 及细胞状态 $\vec{c}_1, \dots, \vec{c}_t, \dots, \vec{c}_{n+1}$ ：

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}\vec{h}_{t-1} + b_{hi}) \quad (2-15)$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}\vec{h}_{t-1} + b_{hf}) \quad (2-16)$$

$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}\vec{h}_{t-1} + b_{hg}) \quad (2-17)$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}\vec{h}_{t-1} + b_{ho}) \quad (2-18)$$

$$c_t = f_t * \vec{c}_{t-1} + i_t * g_t \quad (2-19)$$

$$h_t = o_t * \tanh(\vec{c}_t) \quad (2-20)$$

类似地，后向的长短时记忆网络由初始隐层状态 \tilde{h}_{n+1} 与初始细胞状态 \tilde{c}_{n+1} 开始，从后向前地读入单词序列 $x_n, \dots, x_t, \dots, x_0$ ，并循环地计算其隐层状态 $\tilde{h}_n, \dots, \tilde{h}_t, \dots, \tilde{h}_0$ 。这里，我们使用 BERT 输出的上下文相关向量表示 H_i 作为双向长短时记忆网络的输入 x_i ，并使用 $|L|$ 个双向长短时记忆网络分别建模每个标签转移到其他标签的概率，则 $\mathcal{N}_i^{NLNN-Adapt-RNN}(j, k)$ 的计算方式如下，其中， $\vec{h}_i^j \in \mathbb{R}^{|L|}$ 与 $\tilde{h}_i^j \in \mathbb{R}^{|L|}$ 分别是第 j 个双向长短时记忆网络在 i 时刻前向与后向的隐层状态：

$$\mathcal{N}_i^{NLNN-Adapt-RNN}(j, k) = \text{softmax}(\text{concat}(\vec{h}_i^j, \tilde{h}_i^j))_k \quad (2-21)$$

上述方法独立地建模每个单词位置上的噪声，认为每个观测标签是由该位置上的真实标签独立地经过噪声信道得到的。受条件随机场模型启发，我们考虑建模由整个真实标签序列转移到整个观测标签序列的噪声，并考虑真实标签序列与观测标签序列中相邻标签的依赖关系。具体地，设真实标签序列为 $T = \{t_1, \dots, t_i, \dots, t_n\}$ ，观测标签序列为 $\tilde{T} = \{\tilde{t}_1, \dots, \tilde{t}_i, \dots, \tilde{t}_n\}$ ，我们使用另一个条件随机场模型建模真实标签序列转移到观测标签序列的概率：

$$\begin{aligned} P(\tilde{T}|T) &= \frac{1}{Z} \prod_i \psi_i(\tilde{t}_i|T) \\ &= \frac{1}{Z} \prod_i \exp(\widetilde{\text{Emit}}(\tilde{t}_i, t_i) + \widetilde{\text{Trans}}(\tilde{t}_{i-1}, \tilde{t}_i)) \end{aligned} \quad (2-22)$$

该条件随机场的转移函数 $\widehat{\text{Trans}}(\tilde{t}_{i-1}, \tilde{t}_i)$ 由另一个转移矩阵 \tilde{M} 建模, $\widehat{\text{Trans}}(\tilde{t}_{i-1}, \tilde{t}_i) = \tilde{M}_{\tilde{t}_{i-1}, \tilde{t}_i}$ 。发射函数 $\widehat{\text{Emit}}(\tilde{t}_i, t_i)$ 可由噪声矩阵 \mathcal{N} 建模, 其中 \mathcal{N} 可采用与 NLNN、NLNN-Adapt 或 NLNN-Adapt-RNN 相同的方法计算。我们将使用 NLNN、NLNN-Adapt、NLNN-Adapt-RNN 中方法计算 $\widehat{\text{Emit}}$ 的变种分别记为 NLCRF、NLCRF-Adapt、NLCRF-Adapt-RNN。模型的训练目标变为最小化观测序列 \tilde{T} 的负对数似然:

$$L_{\text{NLCRF}} = -\log \sum_T P(\tilde{T}|T)P(T|S) \quad (2-23)$$

该训练目标依然可通过前向-后向算法高效地计算。NLCRF 系列模型的示意图如图 2-6 所示。

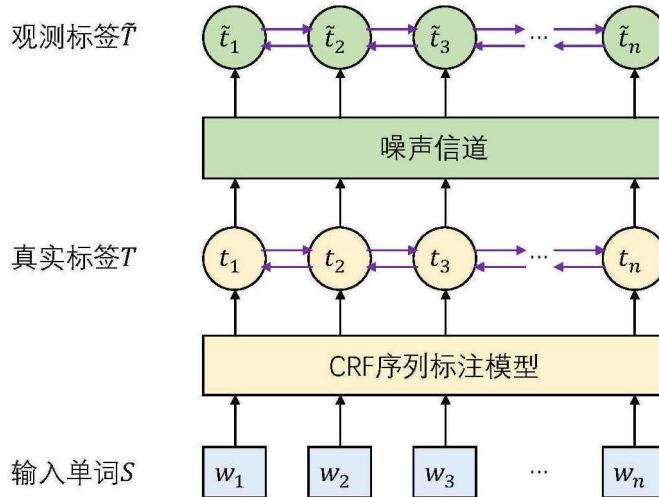


图 2-6 NLCRF 系列模型示意图

自标注数据中的标签带有噪声, 而有标注数据中的标签可以认为是不带有噪声的真实标签, 因此应对两部分数据采用不同的训练设置。我们采用与 Paul 等人^[45]相同的设置, 在有标注数据上使用原始训练目标训练原序列标注, 在自标注数据上使用考虑噪声的训练目标同时训练原序列标注模型与噪声模型。

2.5 实验

2.5.1 实验数据

我们在中文与英文两个事件因果关系抽取数据集上进行了实验。在中文上, 由于并未调研到现有的事件因果关系抽取数据集, 我们人工标注了一份金融领域的事件因果关系抽取数据, 数据的来源为腾讯、网易、和讯、股吧等十余个网站上金融领域的新闻及研究报告等, 这份数据共包含 13839 条数据, 将其划分为训练集 9687 条, 开发集 1384 条, 测试集 2768 条。在英文上, 我们从 Penn

Discourse Treebank^[46]与 BECauSE^[47]两个数据集中筛选了包含因果关系、因果关系为显式且原因与结果在同一个句子中的数据，共得到数据 2800 条，将其划分为训练集 1960 条，开发集 280 条，测试集 560 条。

2.5.2 实验设置

在中文因果关系抽取数据集上，我们采用谷歌发布的 BERT-base-chinese 模型。对于有监督方法，采用 Softmax 输出层与 CRF 输出层均训练 25 轮得到最好结果。对于半监督方法，各模型均在有标注与自标注数据上继续训练 5 轮。其他超参数采用如下设置：学习率为 1×10^{-5} ，Batch 大小为 32。

在英文因果关系抽取数据集上，我们采用谷歌发布的 BERT-base-cased 模型。两组模型的 Transformer 层数均为 12，隐层大小为 768。对于有监督方法，采用 Softmax 输出层与 CRF 输出层均训练 230 轮得到最好结果。对于半监督方法，各模型均在有标注与自标注数据上继续训练 5 轮。其他超参数采用如下设置：学习率为 1×10^{-5} ，Batch 大小为 32。

两个数据集均使用短语级别 F1 值作为评价指标。对于每种短语类型（原因、结果或因果触发词），其抽取结果的 F1 值计算方法为：

$$Precision = \frac{\text{该类别抽取正确的短语数}}{\text{该类别抽取出的短语数}} \quad (2-24)$$

$$Recall = \frac{\text{该类别抽取正确的短语数}}{\text{该类别正确答案中的短语数}} \quad (2-25)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2-26)$$

对于各种类型的短语，我们还计算其抽取结果的 micro-F1 值：

$$Precision' = \frac{\text{该类别抽取正确的短语数}}{\text{该类别抽取出的短语数}} \quad (2-27)$$

$$Recall' = \frac{\text{该类别抽取正确的短语数}}{\text{该类别正确答案中的短语数}} \quad (2-28)$$

$$micro - F1 = \frac{2 \times Precision' \times Recall'}{Precision' + Recall'} \quad (2-29)$$

对于中文数据，我们评价原因短语、结果短语、因果触发词短语各自的 F1 值以及它们的 micro-F1 值；对于英文数据，因为数据集中没有标注因果触发词，我们评价原因短语、结果短语各自的 F1 值以及它们的 micro-F1 值。每种短语类型的 F1 值以及 micro-F1 值均使用 conlleval 脚本统计。因为测试集数据较少，

对于每组实验设置，我们采用 5 个不同的随机种子进行实验，并汇报其平均结果。

2.5.3 实验结果

我们在中文、英文两个因果抽取数据集上，对本章提出的各种方法进行了实验，以验证加入自训练、加入噪声模型的影响，以及对各种建模噪声的方法进行对比。中、英文的实验结果分别如表 2-1 与表 2-2 所示。其中，BERT 为采用 Softmax 输出层，只使用有标注数据的方法；BERT+CRF 为采用 CRF 输出层，只使用有标注数据的方法；BERT+Naive-ST 为采用 Softmax 输出层，使用自训练，但不使用噪声模型的方法；BERT+CRF+Naive-ST 为采用 CRF 输出层，使用自训练，但不使用噪声模型的方法。其他为 2.4 节中提出的使用自训练并采用各种噪声模型的方法。

对比中、英文的实验结果，可以看出中文上各方法的指标都明显高于英文，考虑到中文实验的有标注数据约为英文的 5 倍，这一实验结果是符合预期的。两个数据集上各模型的结果趋势基本相同，加入自训练后可以带来一定的提升，进一步加入噪声模型后可以取得进一步的提升。

对比 BERT 与 BERT-CRF 两种方法，可以看出无论是在中文（93.07 对比 92.98）还是英文数据（77.95 对比 77.39）上，只使用有监督学习时，CRF 输出层相比 Softmax 输出层并不能带来提升，这一结果与 Chen 等人^[48]在任务型对话上的实验结果是一致的。

对比 BERT 与 BERT+Naive-ST 方法，以及 BERT+CRF 与 BERT+CRF+Naive-ST 方法，可以加入自训练后，无论是 Softmax 输出层还是 CRF 输出层都取得了提升，表明自训练方法能够利用无标注数据中的信息为模型带来提升。此外，加入自训练后，CRF 输出层的提升（中文 0.50，英文 1.71）超过了 Softmax 输出层（中文 0.08，英文 0.55），表明 CRF 输出层能够从无标注数据中学习到更多有效的信息。

对比 BERT+Naive-ST 与 BERT+NLNN-Adapt 方法，以及 BERT+CRF+Naive-ST 与 BERT+NLCRF-Adapt 方法，可以得出使用合适的噪声模型建模自标注数据中的噪声可以为模型带来进一步提升。

对比各种噪声模型的实验结果，可以看出使用全局统一的噪声矩阵（NLNN, NLCRF）建模噪声并不能带来提升，反而会使结果有所下降；使用全连接神经网络（NLNN-Adapt, NLCRF-Adapt）与循环神经网络（NLNN-Adapt-RNN, NLCRF-Adapt-RNN）建模噪声都可以带来提升，但循环神经网络的结果低于全

表 2-1 中文因果抽取实验结果

模型	Micro-F1	原因 F1	结果 F1	因果触发词 F1
BERT	93.07	92.09	92.96	94.10
BERT+Naïve-ST	93.15	92.24	92.99	94.39
BERT+NLPN	93.08	91.90	92.96	94.29
BERT+NLPN-Adapt	93.28	92.36	93.26	94.31
BERT+NLPN-Adapt-RNN	93.29	92.26	93.33	94.30
BERT+CRF	92.98	92.07	92.69	94.12
BERT+CRF+Naïve-ST	93.48	92.30	93.88	94.33
BERT+NLCRF	93.07	91.72	93.38	94.23
BERT+NLCRF-Adapt	93.62	92.58	93.96	94.40
BERT+NLCRF-Adapt-RNN	93.59	92.50	93.94	94.35

表 2-2 英文因果抽取实验结果

模型	Micro-F1	原因 F1	结果 F1
BERT	77.95	81.79	74.14
BERT+Naïve-ST	78.50	81.90	75.25
BERT+NLPN	76.64	80.52	72.92
BERT+NLPN-Adapt	78.91	82.27	75.56
BERT+NLPN-Adapt-RNN	78.72	81.98	75.47
BERT+CRF	77.39	81.55	73.35
BERT+CRF+Naïve-ST	79.10	82.55	75.65
BERT+NLCRF	77.16	80.89	73.74
BERT+NLCRF-Adapt	79.58	83.04	76.11
BERT+NLCRF-Adapt-RNN	79.10	82.60	75.61

连接神经网络，可能是因为循环神经网络结构更为复杂，不利于模型学习正确的噪声分布。

为了进一步分析自训练对 CRF 输出层带来的提升，我们对比了加入自训练与噪声模型前后 CRF 输出层中的转移概率，如表 2-3 与表 2-4 所示。从上述数据可以得出，加入自训练与噪声模型后，CRF 学习到了更为合理的转移概率，不符合 BIO 标注规范的标签转移概率（例如 O 标签转移到 I-cause 标签、I-effect 标签、I-trigger 标签）有所降低，符合标注规范的标签转移概率（例如 B-cause 标签转移到 I-cause 标签、B-effect 标签转移到 I-effect 标签、B-trigger 标签转移到 I-trigger 标签）有所提高，表明大规模自标注数据能够帮助 CRF 转移概率的

表 2-3 加入自训练前 CRF 转移概率

前/后标签	O	B-cause	I-cause	B-effect	I-effect	B-trigger	I-trigger
O	0.1974	0.1082	0.1368	0.1542	0.0656	0.1818	0.1561
B-cause	0.1364	0.1445	0.1962	0.1142	0.1249	0.0939	0.1899
I-cause	0.1088	0.1590	0.0967	0.1444	0.1493	0.1486	0.1932
B-effect	0.1192	0.0755	0.0972	0.1900	0.2440	0.1607	0.1134
I-effect	0.1531	0.0796	0.1052	0.2468	0.1871	0.0920	0.1362
B-trigger	0.1376	0.1667	0.2403	0.0985	0.1235	0.0759	0.1576
I-trigger	0.1333	0.0922	0.1421	0.1021	0.1503	0.2707	0.1093

表 2-4 加入自训练后 CRF 转移概率

前/后标签	O	B-cause	I-cause	B-effect	I-effect	B-trigger	I-trigger
O	0.2115	0.1102	0.1275	0.1568	0.0605	0.1851	0.1483
B-cause	0.1318	0.1445	0.2076	0.1089	0.1249	0.0906	0.1917
I-cause	0.1024	0.1572	0.1019	0.1470	0.1457	0.1520	0.1939
B-effect	0.1126	0.0725	0.0967	0.1917	0.2577	0.1547	0.1140
I-effect	0.1418	0.0807	0.1024	0.2461	0.1998	0.0945	0.1337
B-trigger	0.1281	0.1711	0.2363	0.1000	0.1212	0.0753	0.1679
I-trigger	0.1323	0.0934	0.1406	0.1046	0.1498	0.2698	0.1096

学习，从而为 CRF 输出层带来提升。

2.6 本章小结

本章主要介绍了基于 BERT 的有监督因果抽取方法与基于噪声模型的半督因果抽取方法的模型结构，在中、英文两个数据集上的实验设置、实验结果以及结果分析。实验结果表明，加入自训练与噪声模型在两个数据集上都取得了提升，表明该方法可以有效地在有标注数据较少时利用无标注数据提升模型性能。对比各种噪声模型的实验结果，可以得出使用全连接神经网络单独建模每个单词上的噪声是最为合理的。对比 CRF 与 Softmax 两种输出层的实验结果，可以得出自训练能够为 CRF 输出层带来更大的提升，对模型权重的进一步分析表明这种提升一定程度上来自于自训练过程使 CRF 学习到了更为合理的标签转移概率。

第3章 常识信息增强的事件表示学习

3.1 引言

事件是事理图谱的核心元素，准确地建模事件语义对于事理图谱的构建与应用都有着至关重要的作用。事件表示学习将事件表示为计算机可以理解的形式，为使用计算机对事件进行分析预测打下基础。早期研究多使用离散的事件表示，将事件表示为一个事件动作与多个事件元素构成的元组。随着分布式词表示即词嵌入^{[55][56]}的兴起，稠密的事件表示应运而生，稠密的事件表示将事件嵌入到低维向量空间中，有效地缓解了离散的事件表示带来的稀疏性问题，这一系列方法大多建立在词嵌入的基础上，使用神经网络对事件动作与事件元素的词嵌入进行组合，以使事件的向量表示充分包含事件动作与事件元素的语义信息。

然而，这些稠密的事件表示方法仅仅建模了事件动作与事件论元字面上的语义信息，而忽略了事件动作与事件元素背后的领域知识与常识知识。例如，“乔布斯离开苹果公司”与“小明离开星巴克”可能会具有相似的向量表示，但考虑事件实施者与受事者间的实体关系，“乔布斯”是“苹果公司”的CEO，而“小明”与“星巴克”并没有什么联系，因此两个事件会对其受事者造成完全不同的影响。再如，“某人甲扔篮球”与“某人乙扔炸弹”在施事者与事件动作上具有较高的相似性，仅考虑字面信息相似很高，但两个事件的意图完全不同，“扔篮球”的意图是锻炼身体，“扔炸弹”的意图是杀伤敌人，因此从意图角度考虑两个事件的语义相似度很低。又如，“某人甲打破花瓶”与“某人甲打破记录”在字面上依然有很高的相似度，但考虑两个事件的情感极性，“打破花瓶”带有消极的情感，“打破记录”带有积极的情感，因此从情感极性角度考虑，两个事件有较低的语义相似度。

基于上述思考，本章提出了常识信息增强的事件表示学习方法，不仅考虑事件动作与事件元素字面上的语义，同时将事件背后隐含的实体关系、意图、情感极性等常识信息也建模到事件表示中。具体地，本章基于张量神经网络这一稠密的事件表示方法，提出了融合实体关系、意图、情感信息的事件表示联合学习框架，并在事件相似度、脚本事件预测、股市预测三个事件相关的任务上进行了实验，实验结果表明该方法可以有效在事件表示中融入上述常识信息，并显著提升下游任务的结果。

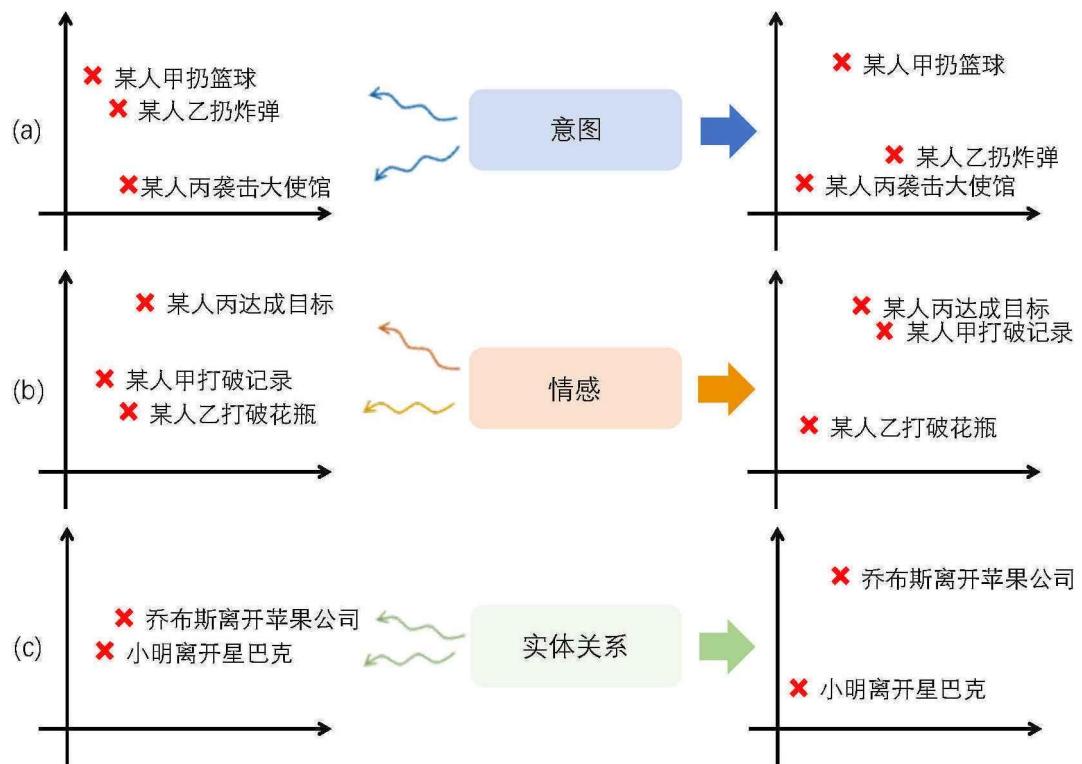


图 3-1 意图、情感以及实体关系对建模事件语义的影响

3.2 任务定义

本章考虑为 (A, P, O) 三元组形式的事件学习向量表示，其中 A 为施事者， O 为受事者， P 为事件动作。 A, P, O 均由多个单词组成， $A = \{w_1^A, \dots, w_i^A, \dots, w_{n_A}^A\}$, $P = \{w_1^P, \dots, w_i^P, \dots, w_{n_P}^P\}$, $O = \{w_1^O, \dots, w_i^O, \dots, w_{n_O}^O\}$ ，其中 n_A 为施事者包含的词数， n_P 为事件动作包含的词数， n_O 为受事者包含的词数。事件表示学习的目标是寻找一个函数 F ，将 (A, P, O) 事件三元组映射为 k 维向量 C ，即 $F(A, P, O) = C$ ，并使得 C 作为特征应用在下游任务上时取得尽可能好的表现。

3.3 基于张量神经网络的事件表示学习

基于张量神经网络的事件表示学习方法首先由 Ding 等人^[32]于 2015 年提出。基于张量神经网络的事件表示学习方法使用双线性张量运算对事件的动作、施事者与受事者进行语义组合，从而使事件表示充分包含事件动作与事件元素相关的信息。

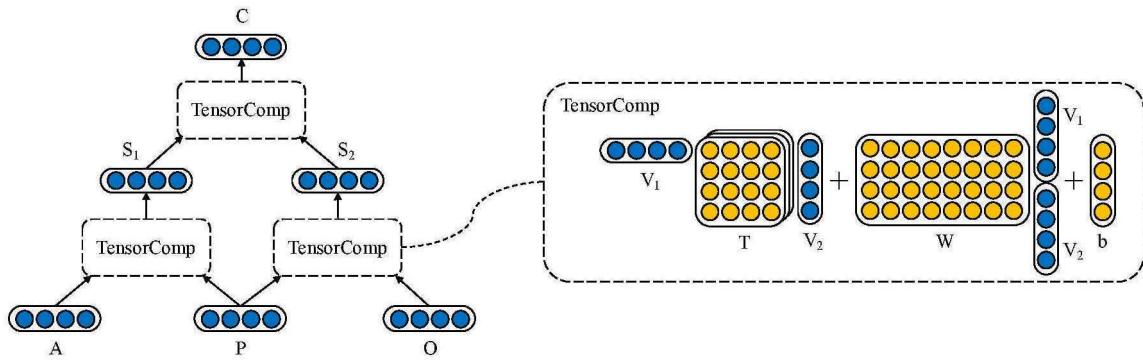


图 3-2 张量神经网络结构

张量神经网络的模型结构如图 3-2 所示。事件的施事者、受事者、事件动作首先分别被表示为其词向量的均值 $\bar{A}, \bar{O}, \bar{P}$ 。之后，使用张量组合运算 $TensorComp$ 分别对事件动作与实施者、事件动作与受事者的向量表示进行语义组合得到向量 S_1, S_2 ，并对 S_1, S_2 进行进一步组合得到事件的向量表示 C ：

$$S_1 = TensorComp(\bar{P}, \bar{A}; \theta_1) \quad (3-1)$$

$$S_2 = TensorComp(\bar{P}, \bar{O}; \theta_2) \quad (3-2)$$

$$C = TensorComp(S_1, S_2; \theta_3) \quad (3-3)$$

其中， $\theta_i = \{T_i, W_i, b_i\}$ 是张量组合运算 $TensorComp$ 的参数，张量组合运算 $TensorComp$ 的具体运算步骤为：

$$TensorComp(V_1, V_2; T, W, b) = f \left(V_1^T T^{[1:k]} V_2 + W \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} + b \right) \quad (3-4)$$

其中 $T^{[1:k]} \in \mathbb{R}^{d \times d \times k}$ 是一个由 k 个 $d \times d$ 维矩阵组成的三位张量， d 为输入向量 V_1, V_2 的维度， k 为输出向量的维度。双线性张量运算的结果 $V_1^T T^{[1:k]} V_2$ 是一个 k 维向量 $r \in \mathbb{R}^k$ ，其每一维上的元素是由输入向量 V_1, V_2 与张量 T 的一个切片运算得到的： $r_i = V_1^T T^{[i]} V_2, i = 1, \dots, k$ 。张量组合运算的其他参数是一个标准的前馈神经网络，其中 $W \in \mathbb{R}^{k \times 2d}$ 是前馈神经网络的权重 $b \in \mathbb{R}^k$ 是前馈神经网络的偏置。 $f = \tanh$ 是一个标准的非线性激活函数。

张量神经网络的权重 $T_i \in \mathbb{R}^{d \times d \times k}$ 包含大量参数，面临“维度灾难”的问题，限制了其在很多领域的应用。因此，对张量神经网络中的三维张量进行压缩以减小其参数数量是十分必要的。我们提出使用低秩张量分解来减小标准张量神经网络中的参数数量，将原张量中的每个矩阵使用两个低维矩阵的乘积加上对角阵来近似，如图 3-3 所示。近似后的张量 T_{appr} 的每个切片为 $T_{appr}^{[i]} = T_1^{[i]} \times T_2^{[i]} + diag(t^{[i]})$ ，其中 $T_1 \in \mathbb{R}^{d \times n \times k}, T_2 \in \mathbb{R}^{n \times d \times k}, t \in \mathbb{R}^{d \times k}$ ， n 是一个超参数，用来调整张量分解的程度。采用低秩张量分解后的张量组合运算 $TensorComp$ 运算过程为：

$$TensorComp(V_1, V_2; T_1, T_2, t, W, b) = f \left(V_1^T T_{appr}^{[1:k]} V_2 + W \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} + b \right) \quad (3-5)$$

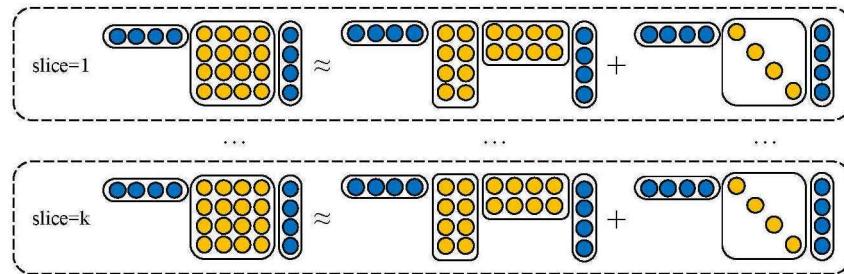


图 3-3 低秩张量分解示意图

我们通过替换训练数据中事件三元组的施事者或受事者构造被破坏的事件三元组。具体地，对于训练集中的事件三元组 $E = (A, P, O)$ ，为其构造被破坏的事件三元组 $E^r = (A^r, P, O)$ 或 $E^r = (A, P, O^r)$ ，其中 E^r 是通过将 E 中的 A 随机替换为词表中的任意词 w^r ，或将 O 替换为词表中的任意词 w^r 得到的。我们认为正确的事件三元组的分数应该高于被破坏的事件三元组的分数，并计算两个事件三元组间的最大边际损失作为训练目标：

$$L_e = loss(E, E^r) = \max(0, 1 - g(E) + g(E^r)) + \lambda \|\theta\|_2^2 \quad (3-6)$$

其中， θ 为张量神经网络的所有参数， $\|\theta\|_2^2$ 为标准的 L2 正则项， λ 为 L2 正则项系数， $g(E)$ 为事件元组 E 的分数，是通过以下方式计算的：

$$g(E) = U^T C_E \quad (3-7)$$

其中 $U \in \mathbb{R}^k$ 是模型的参数， C_E 是张量神经网络输出的事件元组 E 的向量表示。

3.4 常识信息增强的事件表示学习

我们考虑在事件表示中融入实体关系、意图及情感三种类型的常识信息，以使事件表示模型捕获事件文本中未显式提及的语义信息。

3.4.1 融合实体关系信息的事件表示学习

知识图谱中编码了实体关系知识，即两个实体 (e_1, e_2) 是否满足关系 R 。例如，“史蒂夫 乔布斯”与“苹果公司”存在“创始人”的关系。我们采用 Ding 等人^[33]的方法，同时学习实体-关系三元组与事件的向量表示，使事件表示中融入实体关系信息。

在对实体-关系三元组学习向量表示时，我们使用与事件表示模型中相同的张量组合运算。对于知识图谱中的每种关系类型 R ，我们使用一个张量组合运算为满足该类型关系的实体-关系三元组 (e_1, R, e_2) 计算向量表示：

$$EventComp_R(e_1, e_2) = f \left(e_1^T T_R^{[1:k]} e_2 + W_R \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + b_R \right) \quad (3-8)$$

该向量表示被进一步映射为一个分数 $g(e_1, R, e_2)$:

$$g(e_1, R, e_2) = U_R^T EventComp_R(e_1, e_2) \quad (3-9)$$

实体-关系三元组向量表示的训练目标与事件表示相同。对于训练集中每个正确的实体-关系三元组 $T = (e_1, R, e_2)$, 我们将其头实体或尾实体替换为一个随机的实体, 构造被破坏的实体-关系三元组 $T^c = (e_1^c, R, e_2)$ 或 $T^c = (e_1, R, e_2^c)$ 。我们认为正确三元组的分数应该高于被破坏的三元组, 并计算两个三元组间的最大边际损失作为训练目标:

$$L_k = loss(T, T^c) = \max(0, 1 - g(T) + g(T^c)) + \lambda \|\theta\|_2^2 \quad (3-10)$$

在为实体-关系三元组学习向量表示时, 我们为实体-关系三元组中的实体与事件中的施事者、受事者使用同一份词向量。因此, 在学习实体-关系三元组的向量表示时, 来自知识图谱中的实体-关系信息会指导头尾实体词向量的更新, 从而指导事件中施事者与受事者词向量的更新, 使事件表示中融入实体关系知识。

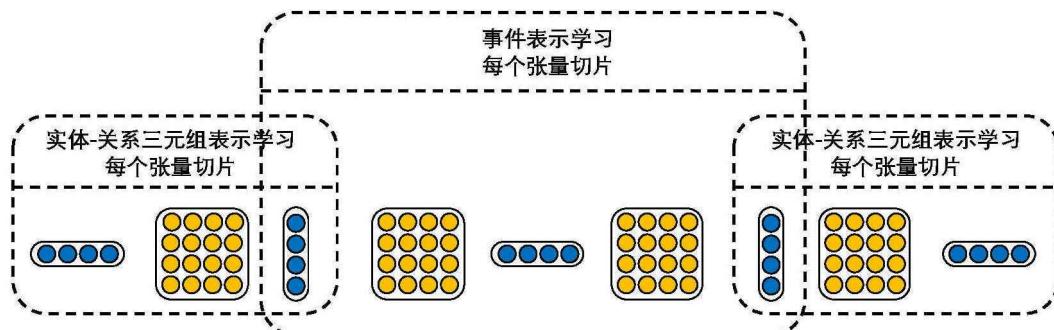


图 3-4 融合实体-关系信息的事件表示学习示意图

3.4.2 融合意图信息的事件表示学习

事件的意图信息解释了为什么事件的施事者要执行事件动作, 例如“某人扔篮球”的意图可能为“进行娱乐”或“锻炼身体”。将意图信息融入事件表示的一个困难是需要大规模标注了事件意图的数据集。Sap 等人^[57]于 2019 年发布了标注了大量事件意图信息的常识知识库 ATOMIC, 为本课题的研究提供了宝贵的资源。在 ATOMIC 数据集中, 事件的意图被标注为一段自然语言文本, 例如, 事件“某人早上喝咖啡”的意图被标注为“某人想保持清醒”。

为了在事件表示中融入意图信息, 我们首先为事件的意图学习向量表示。我们注意到 ATOMIC 数据集中事件的意图被描述为一个句子, 因此, 意图的表示学习实际上是一个句子表示学习的任务。我们采用该任务上广泛使用的双向长短时记忆网络 (BiLSTM)^[28]为意图学习向量表示。双向长短时记忆网络已在 2.4 节中介绍过, 此处不再赘述。我们使用两个长短时记忆网络最后一个时刻的

隐层表示拼接后作为意图的向量表示：

$$\nu_i = \text{concat}(\vec{h}_{n+1}, \vec{h}_0) \quad (3-11)$$

在训练中，我们计算事件向量 ν_e 与意图向量 ν_i 的余弦相似度。对每个事件 e ，其正确的意图标注为 i ，我们为其采样一个错误的意图 i' ，训练事件向量与正确意图的相似度大于其与错误意图的相似度：

$$L_i = \max(0, 1 - \text{cosine}(\nu_e, \nu_i) + \text{cosine}(\nu_e, \nu_{i'})) \quad (3-12)$$

3.4.3 融合情感信息的事件表示学习

事件的情感极性描述了事件施事者在事件发生后的感受。例如，某人“打破记录”后将会很高兴，即处于积极的情感极性；而某人“打破花瓶”后可能会很难过，即处于消极的情感极性。我们使用 ATOMIC 数据集与 SenticNet 数据集^[58]获取事件的情感极性标注。ATOMIC 数据集为每个事件标注了其发生后事件施事者的反应 (X-React)，记录为一段短文本（如“伤心”“后悔”“感到遗憾”“害怕”），我们使用 SenticNet 将这些标注转换为情感极性得分，并进一步二值化为“积极”与“消极”两种情感极性标签。

为了使事件表示中融入情感极性信息，我们引入事件情感极性分类的训练任务。在使用张量神经网络得到事件表示后，将其输入一个分类器预测事件的情感极性标签，并计算分类的交叉熵损失：

$$L_s = - \sum_{x_e \in C} \sum_{l \in L} p_l^g(x_e) \log(p_l(x_e)) \quad (3-13)$$

其中 C 为所有训练样本的集合， L 为情感极性类别的集合， x_e 为事件 e 的向量表示， $p_l(x_e)$ 为分类器输出的 e 属于类别 l 的概率。 $p_l^g(x_e)$ 为 e 的真实类别标注。我们使用标准的前馈神经网络作为分类器的模型结构。

3.4.4 融合实体关系、意图、情感的事件表示联合学习框架

通过结合基于张量神经网络的事件表示学习方法与融合实体关系、意图、情感信息的事件表示学习方法，我们提出了融合多种外部信息的事件表示联合学习框架。给定标注了意图与情感极性的事件语料，我们的框架以多任务学习的形式同时优化上述任务损失函数的线性组合：

$$L = \alpha L_e + \beta L_k + \gamma L_i + \theta L_s \quad (3-14)$$

其中， $\alpha, \beta, \gamma, \theta \in [0, 1]$ 为超参数，用于调整各任务损失的权重。

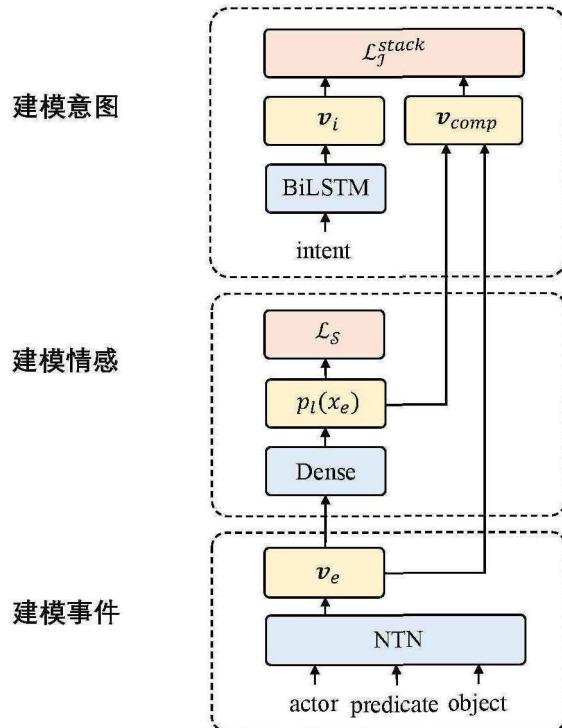


图 3-5 使用栈式传播进行意图与情感建模

我们注意到事件的意图与情感极性并不是相互独立的。例如，“某人全身心地投入学习”表现了积极的情感极性，其意图“想在考试中取得好成绩”也表现了同样的情感极性。因此，事件的情感极性信息可以帮助更好地建模事件意图。另一方面，对事件意图的建模也可以反过来指导事件的情感极性分类。然而，在上述的多任务学习框架中，这两个任务无法显式地获得来自另一个任务的信息。受到 Zhang 等人^[59]提出的“栈式传播”框架的启发，我们采用同样的框架来充分利用情感极性信息指导对意图信息的建模。如图 3-5 所示，在栈式传播框架中，意图建模的训练任务可以直接获得来自情感极性分类任务的特征，并保持了两个任务间的可微性。具体地，我们将事件表示向量与情感极性分类器输出的概率拼接后应用于意图建模，并相应地修改意图建模的损失函数：

$$v_{comp} = \text{concat}(v_e, p_l(x_e)) \quad (3-15)$$

$$L_i^{stack} = \max(0, 1 - \text{cosine}(v_{comp}, v_i) + \text{cosine}(v_{comp}, v_i')) \quad (3-16)$$

并相应地修改整体的损失函数为：

$$L^{stack} = \alpha L_e + \beta L_k + \gamma L_i^{stack} + \theta L_s \quad (3-17)$$

最终，我们提出的事件表示联合学习框架如图 3-6 所示。

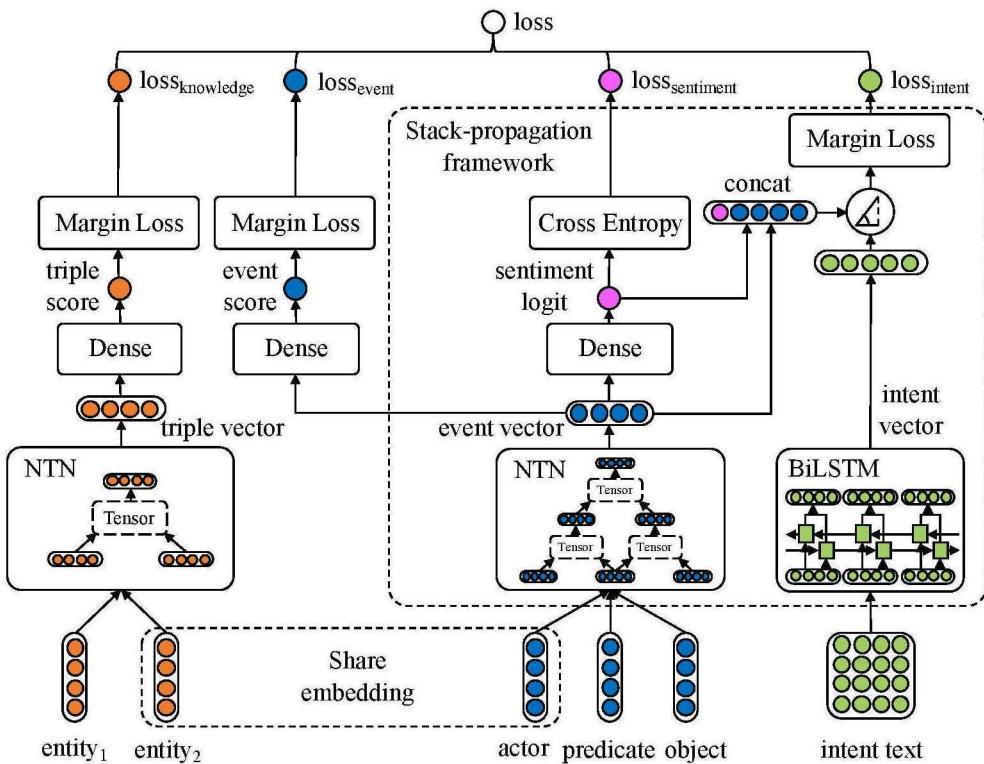


图 3-6 融合实体关系、意图、情感的事件表示联合学习框架

3.5 实验

3.5.1 实验数据

我们在事件相似度、脚本事件预测、股市预测三个事件相关的任务上进行了实验，以验证本文提出的常识信息增强的事件表示学习方法在下游任务上的效果。在事件相似度任务上，我们选用了三个数据集：第一个数据集是 Weber 等人^[31]提出的 Hard Similarity 数据集，该数据集包含 230 个事件对，构成 115 条数据，每条数据包含两个事件对，其中一个事件对字面上的相似度很高但语义上的相似度很低，另一个事件对字面上的相似度很低但语义上的相似度很高，如果模型对后一个事件对的相似度得分高于前一个事件对，则认为这条数据做对，评价指标为准确率（Accuracy）；考虑到 Hard Similarity 数据集规模较小，我们自行标注了一份扩展的 Hard Similarity 数据集，将其扩展到 1000 条测试数据，作为第二个事件相似度数据集；第三个数据集是 Kartsaklis 等人^[60]提出的 Transitive Sentence Similarity 数据集，该数据集包含 108 对及物动词短句，每个短句包含主语、谓语、宾语三种成分，可以很自然地视为（施事者，事件动作，受事者）形式的事件三元组，同时每个事件对标注了一个 1 到 7 之间的相似度。

得分，我们将模型输出的相似度得分与人工标注的相似度得分进行对比，计算斯皮尔曼相关系数作为评价指标。

在脚本事件预测任务上，我们采用与 Li 等人^[7]相同的设置，在标准的 MCNC 数据集上进行实验。该数据集包含从 Gigaword 语料中 New York Times 部分中自动抽取的 160331 个事件链条，每个事件链条包括多个上下文事件，并为链条中最后一个事件设置 5 个选项，任务的形式是要求模型从中选择唯一正确的选项。这些事件链条被划分为 140331 条训练集数据，10000 条开发集数据与 10000 条测试集数据。

在股市预测任务上，我们采用与 Ding 等人^{[32][33]}相同的设置，从路透社与彭博社 2006 年 10 月至 2013 年 11 月的金融新闻标题中抽取事件，并与这段时间内 S&P500 指数的涨跌数据进行对齐。任务的形式为给定前一段时间内发生的事件，预测下一天 S&P500 指数的涨跌情况，即将其建模为一个二分类任务。

3.5.2 实验设置

在事件相似度任务上，我们使用预训练的 100 维 GloVe^[56]词向量作为模型中词向量的初始值。我们首先在 New York Times Gigaword 语料上(LDC2007T07) 使用与 Weber 等人^[31]相同的方法进行预训练，训练目标为给定语料中出现的一个事件，预测其上下文窗口中的其他事件。之后，在 ATOMIC 数据集上，使用 4.4.4 小节的方法，进行融合实体关系、意图、情感信息的事件表示联合学习。在融入实体关系的实验设置中，我们使用来自 YAGO 知识图谱的实体-关系三元组作为训练数据。我们对模型输出的事件向量计算余弦相似度，作为两个事件的相似度分数。

在脚本事件预测任务上，我们使用 Li 等人^[7]在事件链条上预训练的 128 维词向量作为模型中词向量的初始值。我们首先在事件链条数据上采用 Weber 等人^[31]的方法对事件表示进行预训练，之后在 ATOMIC 数据上进行融合常识信息的联合学习。我们使用本文的事件表示方法替换 Li 等人^[7]SGNN 方法中的事件表示部分，并使用修改后的 SGNN 进行脚本事件预测的实验。

在进行股市预测任务的实验时，我们首先使用与事件相似度任务相同的设置训练事件表示模型。之后，我们采用与 Ding 等人^[32]相同的方法，分别选择前 30 天、前 7 天与前 1 天的事件并计算其向量表示作为长期、中期、短期特征，并使用两个一维卷积神经网络对长期、中期特征进行进一步特征提取，最终得到长期、中期、短期三个特征向量，将其拼接后送入前馈神经网络分类器预测下一天股价的涨跌类别。

对于其他超参数，我们在所有实验中采用同样的设置：学习率为 1×10^{-3} ，Batch 大小为 128，L2 正则项系数为 1×10^{-5} 。

3.5.3 实验结果

我们在事件相似度、脚本事件预测与股市预测任务上的实验结果分别如表 3-1、表 3-2 与表 3-3 所示。在所有实验结果中，我们用 NTN 表示不加入常识知识的张量神经网络模型，用 NTN+KG 表示加入实体关系信息的张量神经网络模型，NTN+Int 表示加入意图信息的模型，NTN+Senti 表示加入情感极性信息的模型。

表 3-1 事件相似度实验结果

方法	Hard (小)	Hard (大)	Transitive
Avg	5.2	13.7	0.67
Comp. NN	33.0	18.9	0.63
EM Comp.	33.9	18.7	0.57
Role Factor Tensor	43.5	20.7	0.64
Predicate Tensor	41.0	25.6	0.63
NTN	40.0	37.0	0.60
NTN+KG	52.6	49.8	0.61
NTN+Int	65.2	58.1	0.67
NTN+Senti	54.8	52.2	0.61
NTN+Int+Senti	77.4	62.8	0.74
NTN+KG+Int+Senti (Multi-Task)	77.4	64.6	0.76
NTN+KG+Int+Senti (Stack-Prop)	79.1	69.4	0.74

在事件相似度任务上，我们与基于词向量加性组合的方法(Avg, Comp. NN, EM Comp.) 以及 Weber 等人^[31]提出的另两种基于双线性张量运算的方法(Role Factor Tensor 与 Predicate Tensor) 进行了对比。我们发现，简单的词向量均值方法(Avg) 在 Transitive Sentence Similarity 任务上取得了很高的结果，但在两个 Hard Similarity 数据集上表现很差，主要是因为该方法无法有效区分事件字面上的相似性与语义上的相似性。基于双线性张量运算的方法(NTN, Role Factor Tensor, Predicate Tensor) 在 Hard Similarity 数据集上显著超越了基于词向量加性组合的方法，表明这些方法能够更好地对事件元素的语义进行组合，在一定程度上区分子面相似性与语义相似性。加入常识知识的事件表示方法(NTN+KG, NTN+Int, NTN+Senti) 带来了进一步提升，加入意图(NTN+Int) 与情感

(NTN+Senti) 信息的提升比加入实体关系 (NTN+KG) 信息更为明显, 表明实体无关的常识知识对于这些任务的帮助更大。最终, 融合多种常识知识的方法 (NTN+KG+Int+Senti) 取得了最好的实验结果, 表明本文提出的事件表示联合学习框架能够充分学习多种类型的常识知识, 且多种常识知识能够相互补充, 以更准确地建模事件的语义相似度。

为了更好地展示常识知识对事件向量空间结构的影响, 我们从 ATOMIC 数据集中筛选了饮食、暴力、体育、教育四种主题下的事件, 为其计算向量表示, 并使用主成分分析 (PCA) 将其投影到二维空间中, 如图 3-7 所示。从投影结果可以看出, 加入常识知识后, 相同主题下的事件紧密地聚合在一起, 表明常识知识可以指导模型学习更准确的事件向量表示。

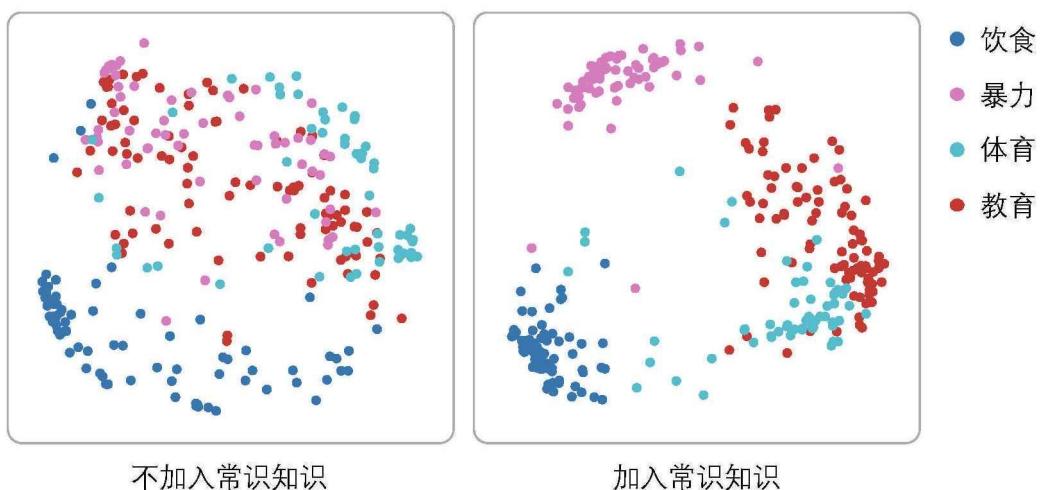


图 3-7 事件表示使用主成分分析在二维空间的投影

在脚本事件预测任务上, 我们首先比较了单模型的实验结果。加入实体知识 (SGNN+KG)、意图 (SGNN+Int)、情感 (SGNN+Senti) 的方法都带来了 1% 以上的提升, 表明三种常识知识都可以为预测未来事件提供帮助。其中, 意图信息 (SGNN+Int) 带来了最为显著的提升, 这可能是因为事件链条中各事件的意图信息是紧密关联的, 因此正确建模事件意图可以为预测后续事件提供很大帮助。加入所有三种知识 (SGNN+KG+Int+Senti) 的方法取得了最好的单模型实验结果, 表明三种信息可以互补地提升该任务上的效果。我们进一步尝试了将基于 SGNN 的方法与另两种脚本预测模型 (PairLSTM 与 EventComp) 进行模型集成, 最终集成所有三种模型并加入所有三种知识的方法 (SGNN+EventComp+PairLSTM+KG+Int+Senti) 取得了最好的实验结果。

在股市预测任务上, 加入实体关系信息 (NTN+KG) 带来了最为显著的提升, 表明实体间的关系 (例如员工与公司的供职关系等) 对于预测金融事件的

表3-2 脚本事件预测实验结果

方法	准确率
SGNN	52.45
SGNN+KG	53.72
SGNN+Int	53.93
SGNN+Senti	53.57
SGNN+Int+Senti	53.88
SGNN+KG+Int+Senti	53.95
SGNN+KG+Int+Senti (Stack-Prop)	54.26
SGNN+PairLSTM	52.71
SGNN+EventComp	54.15
SGNN+EventComp+PairLSTM	54.93
SGNN+PairLSTM+KG	54.10
SGNN+PairLSTM+Int+Senti	54.14
SGNN+EventComp+KG	55.30
SGNN+EventComp+Int+Senti	55.08
SGNN+EventComp+PairLSTM+Int+Senti	56.03
SGNN+EventComp+PairLSTM+KG+Int+Senti	56.06

影响十分重要。加入情感极性信息(NTN+Senti)也带来了一定的提升，这与前人的研究是相符的，即带有积极或消极情感的新闻会影响人们的交易决策，进而影响股市的走势。单独加入意图信息(NTN+Int)虽然没有带来提升，但同时加入意图与情感信息(NTN+Int+Senti)取得了比单独加入情感信息(NTN+Senti)更好的实验结果，表明意图信息可以对情感信息起到很好地补充作用。最终，加入所有三种信息的方法在股市预测任务上取得了最高的准确率。

表3-3 股市预测实验结果

方法	准确率
NTN	65.08
NTN+KG	66.93
NTN+Int	64.17
NTN+Senti	65.24
NTN+Int+Senti	65.78
NTN+KG+Int+Senti	67.3

3.6 本章小结

本章首先介绍了基于张量神经网络的事件表示学习方法。张量神经网络能够很好地组合事件动作与事件元素字面上的语义，但无法捕获事件背后蕴含的常识信息。为了弥补这一缺陷，本章进一步提出了融合实体关系、意图、情感三种常识信息的事件表示联合学习框架，通过在训练事件表示模型时引入外部常识知识库并构造额外的训练目标，使事件表示模型能够捕获事件文本中未显式提及的常识知识。常识信息增强的事件表示在事件相似度、脚本事件预测及股市预测任务上都取得了明显的提升，尤其在事件相似度任务上指标大幅超过基线方法。进一步的分析表明，加入意图信息对于脚本事件预测任务有较大的帮助，而加入实体关系与情感极性信息能够为股市预测任务带来明显的提升。

第4章 数据驱动的因果关系强度计算

4.1 引言

事件间的因果关系往往不是确定性的，而是以一定的概率成立。例如，“下雨”有可能会导致“洪水”，但并非每次下雨都会导致或洪水。由此引出了因果关系强度的概念。本文提出使用因果关系强度度量两个事件间因果关系成立的概率，如果两个事件间的因果强度很低，表明两个事件间不存在因果关系，是两个相互独立的事件；如果两个事件间的因果强度很高，表明两个事件间存在确定性的因果关系，即原因事件发生后结果事件也一定发生，若结果事件发生则原因事件也一定发生过。在事理图谱中，因果关系强度可以作为因果关系边上的权重，为下游任务提供有价值的信息。

本文探索了基于统计的因果关系强度计算方法与基于预训练模型的因果关系强度计算方法。事件间的因果关系强度可以由其在因果关系中的共现信息体现，由此引出了基于统计的因果强度计算方法。Luo 等人^[1]在 2016 年提出利用统计信息建模词级别的因果强度，以缓解事件文本的稀疏性，并从必要性与充分性两方面进行建模，取得了一定的成功。

基于统计的因果强度计算方法忽略了文本的语义信息，例如“下雨”与“降水”具有相似的语义，因此与其他事件间的因果强度也应有相似的数值，但基于统计的方法无法捕获到这种语义上的相似性。预训练的语言模型^{[34][35][36]}如 BERT 等在大规模语料上学习了丰富的语义知识，并在众多自然语言处理任务上取得了优异的结果，因此本文提出基于预训练语言模型的因果强度计算方法。该方法将因果强度计算建模为文本匹配任务，输入原因与结果事件文本，由模型输出两事件的因果关系强度分数。本文提出在因果语料上对语言模型进行进一步预训练，使其学习到建模因果关系强度的能力。

本文在 COPA 因果推理数据集上验证了两种方法的效果，实验显示基于预训练模型的因果强度计算方法可以显著提升 COPA 任务上的实验结果，表明该方法能够更准确地建模事件间的因果关系强度。

4.2 任务定义

本文将事件间因果关系强度的计算建模为文本匹配任务。文本匹配任务的输入为两段自然语言文本： $S_1 = \{w_1^1, \dots, w_i^1, \dots, w_n^1\}$, $S_2 = \{w_1^2, \dots, w_j^2, \dots, w_m^2\}$ ，输出

为两段文本的相关性分数 C 。其中， n 为第一段文本 S_1 所包含的单词数， m 为第二段文本 S_2 所包含的单词数， $w_i^1, \forall i \in \{1, \dots, n\}$ 为第一段文本 S_1 中的单词， $w_j^2, \forall j \in \{1, \dots, m\}$ 为第二段文本 S_2 中的单词。在因果关系强度计算中， S_1 与 S_2 为两段事件文本，相关性分数 C 为两事件的因果关系强度分数，本文进一步规定因果关系强度 C 的取值范围为 0 与 1 之前的实数， $C = 0$ 表示事件 S_1 与 S_2 之间不存在因果关系，即 S_1 与 S_2 的发生是相互独立的； $C = 1$ 表示事件 S_1 与事件 S_2 之间存在确定性的因果关系，即事件 S_1 发生是事件 S_2 发生的充分必要条件。

4.3 基于统计的因果关系强度计算

事件间的因果关系强度可以通过大规模语料中的共现信息计算，例如，如果事件 A 与事件 B 总是出现在同一个因果关系中，那么事件 A 与事件 B 往往存在较强的因果关系，由此引出了基于统计的因果强度计算方法。这种方法的一个缺点是受到事件稀疏性的制约，由于自然语言的随意性，每一个单独的事件文本在语料中的出现次数都非常低，使得有效地统计事件共现信息变得十分困难。Luo 等人^[1] 提出基于词对在因果关系中共现信息计算因果关系强度，并使用词对因果强度的组合作为事件的因果关系强度，缓解了事件文本的稀疏性问题。

另一方面，事件间的因果关系强度体现在必要性与充分性两个方面。考虑一个因果事件对 (S_1, S_2) ，必要性强调若结果事件 S_2 发生，则之前必须有原因事件 S_1 发生；充分性强调若原因事件 S_1 发生，则结果事件 S_2 一定在其后发生。例如，（降雨，洪水）这一事件对具有较强的必要性与较弱的充分性，因为若没有降雨，洪水几乎不可能发生，但即使发生了降雨，也不一定会导致洪水发生。相反，（风暴，损失）这一事件对具有较强的充分性与较弱的必要性，因为风暴几乎必然会带来损失，但造成损失的原因不一定是风暴。从必要性与充分性两个角度考虑，对于原因事件中的单词 i_c 与结果事件中的单词 j_e ，使用如下的方法计算两个单词间的因果强度：

$$CS_{nec}(i_c, j_e) = \frac{P(i_c|j_e)}{P^\alpha(i_c)} = \frac{P(i_c, j_e)}{P^\alpha(i_c)P(j_e)} \quad (4-1)$$

$$CS_{suf}(i_c, j_e) = \frac{P(j_e|i_c)}{P^\alpha(j_e)} = \frac{P(i_c, j_e)}{P^\alpha(j_e)P(i_c)} \quad (4-2)$$

$CS_{nec}(i_c, j_e)$ 从必要性的角度建模了 i_c, j_e 间的因果强度， $CS_{suf}(i_c, j_e)$ 从充分性的角度建模了 i_c, j_e 间的因果强度。直观上看，后验概率 $P(i_c|j_e)$ 越大，反映了因果关系的必要性越强； $(j_e|i_c)$ 越大，反映了因果关系的充分性越强。然而，一些

高频词更可能同时在原因与结果文本中出现，使得后验概率更偏向这些高频词。因此，我们使用单词的先验概率作为惩罚项，对高频词的后验概率进行惩罚。其中 α 为惩罚项系数，本文中将其设置为0.66。

上式中的各项概率值可以通过大规模语料上的统计信息进行估计，具体地：

$$P(i_c) = \frac{\sum_{w \in W} f(i_c, w_e)}{M} \quad (4-3)$$

$$P(j_e) = \frac{\sum_{w \in W} f(w_c, j_e)}{M} \quad (4-4)$$

$$P(i_c, j_e) = \frac{f(i_c, j_e)}{N} \quad (4-5)$$

其中， $f(i_c, j_e)$ 是语料中统计得到的单词*i*出现在原因事件中且单词*j*出现在结果事件中的概率， W 是语料中出现的所有单词集合， M 与 N 为归一化系数，确保计算结果满足概率的性质。

两个单词*i, j*的因果强度 $CS(i_c, j_e)$ 是其考虑必要性与充分性的因果强度的组合：

$$CS(i_c, j_e) = CS_{nec}(i_c, j_e)^\lambda CS_{suf}(i_c, j_e)^{1-\lambda} \quad (4-6)$$

最终，两个事件 S_1, S_2 的因果强度是事件中所有词对因果强度的组合：

$$CS_T(S_1, S_2) = \frac{1}{|S_1| + |S_2|} \sum_{i \in S_1} \sum_{j \in S_2} CS(i_c, j_e) \quad (4-7)$$

4.4 基于预训练模型的因果关系强度计算

随着自然语言处理技术的发展，预训练语言模型^{[34][35][36]}如BERT在多项文本匹配任务上取得了最佳的效果。这些模型在大规模语料上以遮罩语言模型等训练目标进行预训练，使其充分学习到文本分布的先验知识，显著提升了各项下游任务上的效果。

本文提出基于预训练语言模型的因果关系强度计算方法。具体地，我们将两段事件文本拼接后输入预训练语言模型，模型为文本中的每个单词计算一个上下文相关的向量表示，这些向量表示经过池化后得到一个固定长度的向量，该向量进一步由一个全连接网络映射为因果强度分数。

$$H_1, \dots, H_k, \dots, H_{m+n} = PLM(S_1, S_2) \quad (4-8)$$

$$H = Pooler(H_1, \dots, H_k, \dots, H_{m+n}) \quad (4-9)$$

$$C = \sigma(WH + b) \quad (4-10)$$

其中， PLM 函数表示预训练语言模型， $H_1, \dots, H_k, \dots, H_{m+n}$ 为其输出的上下文相关词向量， $Pooler$ 为一个池化函数，将这些上下文相关词向量映射为一个个

固定长度的向量 H 。 H 由一个全连接神经网络映射为因果强度分数 C , W, b 为该全连接网络的参数。 σ 表示 sigmoid 函数, 为该全连接网络的激活函数, 用来对因果强度分数进行归一化, 确保其取值范围在 0 到 1 之间:

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \quad (4-11)$$

我们将基于预训练模型的因果强度计算方法使用一个函数 $CPLM$ 表示:

$$C = CPLM(S_1, S_2) \quad (4-12)$$

受 Li 等人^[49]两阶段预训练方法的启发, 我们在因果数据上对预训练语言模型进行进一步预训练, 使其不仅学习到文本分布的先验知识, 还学习到事件间因果关系的相关先验知识, 帮助其更好地建模两事件的因果关系强度, 并在需要因果知识的下游任务上取得更好的效果。具体地, 每条预训练数据包含两个事件对: S_1 与 S_2 为两个因果强度较强的事件, 例如从文本中抽取出的含有因果关系的事件; S'_1 与 S'_2 为两个因果强度较弱的事件, 例如从语料中随机采样的两个事件。预训练任务的目标为使 S_1 与 S_2 的因果强度高于 S'_1 与 S'_2 。我们使用最大边际损失作为预训练任务的损失函数:

$$\begin{aligned} L &= \max(0, M + C' - C) \\ &= \max(0, M + CPLM(S'_1, S'_2) - CPLM(S_1, S_2)) \end{aligned} \quad (4-13)$$

我们具体考虑了三种预训练语言模型结构用于因果强度计算:

(1) **BERT** BERT^[34]为基于双向 Transformer 编码器结构的预训练语言模型, 该模型的结构已在 2.3 节中详细介绍, 此处不再赘述。当输入为 S_1, S_2 两段文本时, BERT 使用特殊的符号将两段文本拼接为如下形式: [CLS] S_1 [SEP] S_2 [SEP], 其中[CLS]符号标记文本的开始, [SEP]符号标记文本的结尾。我们以[CLS]符号对应的上下文相关向量表示作为所有单词上下文相关向量的池化结果:

$$Pooler_{BERT}(H_{[CLS]}, \dots, H_{[SEP]}) = H_{[CLS]} \quad (4-14)$$

(2) **RoBERTa** RoBERTa 模型由 Liu 等人^[35]于 2019 年提出, 是对 BERT 模型的改进方法。RoBERTa 的模型结构与 BERT 相同, 但在预训练设置上进行了改动, 包括使用更大的 Batch 大小与更大的训练数据, 去掉了 BERT 预测下一个句子的训练任务, 使用更长的文本进行预训练, 以及在遮罩语言模型预训练任务中动态地对文本进行遮罩。此外, RoBERTa 在对文本进行分词时使用双字节编码 (Byte-Pair Encoding), 并且去掉了 BERT 中的 Segment Embedding。

对于 RoBERTa 模型, 我们采用 Kavumba 等人^[50]的方法, 将 S_1, S_2 两段文本拼接为如下形式: <s> S_1 S_2 </s>, 其中<s>符号标记文本的开始, </s>符号标记文本的结尾, 并以<s>符号对应的上下文相关向量作为池化结果:

$$Pooler_{RoBERTa}\left(H_{\langle s \rangle}, \dots, H_{\langle \overline{s} \rangle}\right) = H_{\langle s \rangle} \quad (4-15)$$

(3) **ALBERT** ALBERT 模型由 Lan 等人^[36]于 2020 年提出，在 BERT 的基础上引入跨层参数共享机制，使模型的每一个 Transformer 层共享相同的参数。这一改进在取得与 BERT 相近的实验结果的同时显著降低了模型参数量，并使得 ALBERT 可以在使用同样计算资源的情况下构建比 BERT 更大规模的模型。

此外，ALBERT 还对 BERT 的 Embedding 层进行分解，使得 Embedding 的向量大小远小于隐层向量大小，进一步减少了模型参数。ALBERT 还引入预测句子顺序的预训练任务，替代 BERT 中判断两句子是否相邻的训练任务。

对于 ALBERT 模型，我们采用 Kavumba 等人^[50]的方法，将 S_1, S_2 两段文本拼接为如下形式：[SEP] $S_1 S_2$ [SEP]，其中[SEP]符号同时标记文本的开始与结尾，并以第一个[SEP]符号对应的上下文相关向量作为池化结果：

$$Pooler_{ALBERT}\left(H_{[SEP]}, \dots, H'_{[SEP]}\right) = H_{[SEP]} \quad (4-16)$$

4.5 实验

4.5.1 实验数据

我们使用 Li 等人提出的 CausalBank 语料对 4.4 节中提出的模型进行预训练。CausalBank 是一个大规模的英文因果语料库，包含从 5.14TB 的 Common Crawl^[51]语料中抽取的约 3.14 亿个因果事件对，显著超过了目前所有的因果语料资源。这些因果事件对是由高质量的因果关系模板自动抽取的。Li 等人 CausalBank 中随机采样了 200 条抽取结果进行人工评价，认为 95% 以上的数据包含有意义的因果关系，表明该因果语料库具有很高的质量。

我们在 SemEval COPA^[2]因果关系推理数据集上对因果关系强度计算方法进行评价。该任务包含 1000 个因果常识推理问题，划分为 500 条开发集数据与 500 条测试集数据。在近年的研究中，该开发集也经常作为训练集使用，用于对有监督的方法进行微调。COPA 数据集中的每个问题包含一个前提与两个假设，以及对因果关系方向的提问，要求算法根据前提与提问选择最合理的假设，如图 3-1(a)所示。在 COPA 数据集中，两个假设并非是与前提构成或不构成因果关系的区别，而是与前提的因果关系合理性高低的区别，即使是错误的假设也与前提有一定的相关性，使得该任务具有挑战性。从因果关系强度的角度出发，可视为前提与正确的假设具有较高的因果关系强度，而与错误的假设具有较低的因

果关系强度，任务的目标是区分两个事件对因果关系强度的高低，因此该任务天然适合对因果关系强度计算方法进行评价。

Kavumba 等人^[52]提出 COPA 数据集中存在表面上的线索，例如正确与错误的假设选项中的词频差异，使得模型即使没有掌握因果推理的能力，也可以通过这些表面线索对正确答案作出判断。这些表面线索的存在可能使 COPA 上的实验结果无法客观地反映模型的因果推理能力。为了消除表面线索的影响，Kavumba 等人提出了 B-COPA 数据集，如图 4-1 所示。B-COPA 为 COPA 中的每条数据构造一个镜像数据，通过设置一个新的前提，使原先作为正确选项的假设变为错误选项，原先作为错误选项的假设变为正确选项，平衡了正误假设中的词频差异，消除了 COPA 数据集中潜在的表面线索。为了更客观地评价本文提出的因果强度计算方法，本文同时在 COPA 与 B-COPA 两个数据集上进行实验，并汇报实验结果。

The woman hummed to herself. What was the *cause* for this?
这位女性自言自语地哼着歌。可能的原因是？

She was in a good mood. 她心情很好。

She was nervous. 她很紧张。

(a) COPA中的原始数据

The woman trembled. What was the *cause* for this?
这位女性在颤抖。可能的原因是？

She was in a good mood. 她心情很好。

She was nervous. 她很紧张。

(b) B-COPA中的镜像数据

图 4-1 COPA 与 B-COPA 中的数据示例

4.5.2 实验设置

我们从 CausalBank 语料中随机采样了 10 万个因果事件对，用来构造预训练因果强度模型所需的数据。我们使用采样出的每个因果对作为正样本，并随机将其中的原因事件或结果事件替换为 CausalBank 语料中的任意事件，为每个正样本构造两个负样本。每个正样本分别和两个负样本配对得到两条训练数据，最终共计得到 20 万条训练数据。

我们在上述数据集上，对4.4节中提出的模型进行预训练。我们选择了不同尺寸的BERT、RoBERTa、ALBERT模型进行预训练，每组模型的参数规模以及预训练时使用的超参数如表4-1所示。最大边际损失中的超参数 M 均设置为1。

表4-1 模型尺寸与预训练时的超参数设置

模型	参数量	模型层数	隐层大小	Batch大小	学习率
CausalBERT-base-uncased	108M	12	768	16	1e-5
CausalBERT-large-uncased	334M	24	1024	4	3e-6
CausalRoBERTa-base	123M	12	768	16	1e-5
CausalRoBERTa-large	355M	24	1024	4	3e-6
CausalALBERT-base	12M	12	768	32	1e-6
CausalALBERT-large	18M	24	1024	12	1e-6
CausalALBERT-xxlarge	235M	12	4096	4	3e-6

我们在COPA与B-COPA数据集上测试本文所提出方法与基线方法的效果。为了更全面地评价各种方法的效果，并公平地与基线方法进行对比，我们提出了以下三组实验设置：

(1) 在**COPA**与**B-COPA**上微调整个模型 我们采用与Wang等人^[53]，Sap等人^[54]，Li等人^[49]及Kavumba等人^[50]的实验设置，在COPA的500条开发集数据上对模型进行微调，之后在COPA的测试集上进行测试。这一组实验设置允许我们与目前效果最好的方法进行对比。由于前人进行微调时大多使用交叉熵损失，为了进一步对比交叉熵损失与最大边际损失的效果，我们除了测试3.4节中提出的方法以外，还测试原始的BERT、RoBERTa、ALBERT模型使用最大边际损失微调的实验结果。该设置下各模型使用的超参数如表4-2所示。

(2) 不进行微调，直接在**COPA**上进行测试 我们认为，在COPA上对模型进行微调时，测试集上的实验结果实际上是对模型两个方面的综合评价：一个方面是模型先前掌握了多少因果知识，另一个方面是模型可以在微调过程中可以学到多少有效信息。Kavumba等人的实验结果显示，RoBERTa与ALBERT在不引入额外因果知识时，也能在COPA上取得极高的实验结果，表明这些模型本身具有很强的学习能力，只需在COPA开发集上进行学习即可获取该任务所需的知识。这些模型的较强学习能力会影响对其掌握多少因果知识的评价，因为模型仅依靠学习能力即可达到很高的指标，使因果知识带来的提升不明显。为了排除学习能力的影响，更准确地评价模型对因果知识的掌握，我们提出不在COPA上进行微调，直接在测试集上汇报其实验结果。这一组实

表 4-2 在 COPA 上微调时的超参数设置

模型	Batch 大小	学习率	训练轮数	最大边际损失的 M
BERT-base-uncased	64	5e-5	5	0.37
BERT-large-uncased	16	5e-6	15	0.37
RoBERTa-base	64	3e-5	10	0.37
RoBERTa-large	16	5e-6	15	0.37
ALBERT-base	64	2e-5	15	0.4
ALBERT-large	32	2e-5	15	1
ALBERT-xxlarge	8	5e-6	15	1
CausalBERT-base-uncased	64	3e-5	5	0.37
CausalBERT-large-uncased	16	5e-6	15	0.37
CausalRoBERTa-base	64	3e-5	10	0.37
CausalRoBERTa-large	16	8e-6	20	0.37
CausalALBERT-base	64	2e-5	15	0.4
CausalALBERT-large	32	2e-5	15	1
CausalALBERT-xxlarge	8	5e-6	15	1

表 4-3 在 COPA 上微调输出层时的超参数设置

模型	Batch 大小	学习率	训练轮数	最大边际损失的 M
BERT-base-uncased	64	1e-3	100	1
BERT-large-uncased	16	1e-3	100	1
RoBERTa-base	64	1e-3	100	1
RoBERTa-large	16	8e-3	100	1
ALBERT-base	1024	3e-2	20	1
ALBERT-large	512	2e-2	30	1
ALBERT-xxlarge	16	1e-3	30	1
CausalBERT-base-uncased	64	1e-2	5	0.37
CausalBERT-large-uncased	16	1e-2	10	0.37
CausalRoBERTa-base	64	1e-2	10	0.37
CausalRoBERTa-large	16	1e-2	10	0.37
CausalALBERT-base	1024	3e-2	20	1
CausalALBERT-large	512	2e-2	30	1
CausalALBERT-xxlarge	16	1e-3	30	1

验设置也允许我们更客观地与无监督的方法进行对比，例如基于统计的因果强度计算方法。

(3) 在 COPA 与 B-COPA 上微调模型输出层 第(2)组实验设置存在一个缺陷，即无法与原始的 BERT、RoBERTa、ALBERT 等模型进行对比，因为这些原始模型缺少预训练的、可直接输出因果强度分数的输出层，所以无法直接在 COPA 任务上进行测试。为了与原始模型进行对比，同时尽可能消除模型学习能力的影响，我们采用 Kavumba 等人的实验设置，只在 COPA 数据上微调模型的输出层，而固定 BERT、RoBERTa、ALBERT 模型自身的参数。这一组实验设置可以更公平地对比本文方法与原始模型对因果关系强弱的建模能力。该设置下各模型使用的超参数如表 4-3 所示。

因为 COPA 的测试集规模很小，只包含 500 条数据，我们对每组设置选取 5 个不同的随机种子进行实验，并汇报其平均结果，以平衡随机性带来的结果波动。

4.5.3 实验结果

我们首先汇报在 COPA 上微调整个模型的实验结果，并与前人采用同样设置的结果进行对比，如表 4-4 所示。对比表中 Wang 等人^[53]，Sap 等人^[54]与 Kavumba 等人^[50]在原始 BERT、RoBERTa、ALBERT 模型上使用交叉熵损失的实验结果，以及我们汇报的使用最大边际损失的实验结果，可以得出最大边际损失在 COPA 任务上取得了更好的效果，这与 Li 等人的结论是一致的。进一步对比基于原始模型的实验结果与基于 Causal 模型的实验结果，可以发现除了 RoBERTa-large 与 ALBERT-xxlarge 两组模型外，Causal 模型都取得了比原始模型更好的结果，表明本文提出的方法可以使模型学习到丰富的因果知识，并在 COPA 任务上带来提升。上述两组模型未取得明显提升的原因可能是微调过程中原模型过强的学习能力掩盖了因果知识带来的提升。B-COPA 上的实验结论与 COPA 一致。

我们进一步将 Causal 模型直接在 COPA 测试集上进行测试，并与基于统计的因果强度计算方法进行对比，结果如表 4-5 所示。因为 COPA 与 B-COPA 数据集测试集完全一致，只有开发集存在不同，因此这组设置仅汇报 COPA 的实验结果。在排除了模型学习能力的影响后，我们可以更准确地评价模型对因果强度的建模能力。从实验结果中可以得出，基于统计的因果强度计算方法是一个很强的基线方法，在不进行微调时，CausalBERT-base 的结果低于基于统计的方法，CausalBERT-large 也只能取得与其持平的结果。CausalRoBERTa 模型与

表 4-4 在 COPA 上微调的实验结果

模型	损失函数	COPA 结果	B-COPA 结果
BERT-large (Wang 等, 2019)	交叉熵	70.6	-
BERT-large (Sap 等, 2019)	交叉熵	75	-
BERT (Li 等, 2019)	最大边际	75.4	-
BERT-large (Kavumba 等, 2019)	交叉熵	76.5	74.5
RoBERTa-large (Kavumba 等, 2019)	交叉熵	87.7	89
ALBERT-xxlarge (Kavumba 等, 2019)	交叉熵	92.1	92.3
BERT-base-uncased	最大边际	74.5	76.3
BERT-large-uncased	最大边际	77.8	80
RoBERTa-base	最大边际	80.5	81.3
RoBERTa-large	最大边际	90.3	90.2
ALBERT-base	最大边际	70.6	74.8
ALBERT-large	最大边际	80.1	79.7
ALBERT-xxlarge	最大边际	92.8	92.7
CausalBERT-base-uncased	最大边际	78.6	78.6
CausalBERT-large-uncased	最大边际	79.3	80.6
CausalRoBERTa-base	最大边际	85.4	83.8
CausalRoBERTa-large	最大边际	90.9	90.5
CausalALBERT-base	最大边际	75.2	75.8
CausalALBERT-large	最大边际	82.1	81.5
CausalALBERT-xxlarge	最大边际	92.6	93.5

CausalALBERT 模型的实验结果都超过了基于统计的方法，CausalALBERT-xxlarge 的实验结果达到了惊人的 86.4%，甚至超越了大部分微调后的实验结果，表明其确实在预训练过程中学习了大量因果知识。与进行微调的实验结果相比，CausalRoBERTa-large 的实验结果降低了 13.1 个百分点，而 CausalALBERT-xxlarge 仅降低了 7.1 个百分点，表明 ALBERT 模型具有更强的在预训练过程中学习因果知识的能力。

最后，我们进行了只微调模型输出层的实验，结果如表 4-6 所示。在相同的模型上，我们的实验结果超过了 Kavumba 等人的实验结果，这可能归功于最大边际损失相对于交叉熵损失的优势。对比原始模型与 Causal 模型，可以发现各种类型及尺寸的模型上都取得了提升，表明排除了学习能力的影响后，预训练过程中模型所学习的因果知识的作用得以真正体现，并为实验结果带来提升。

表 4-5 直接在 COPA 上测试的实验结果

模型	损失函数	COPA 结果
BigramPMI (Goodwin 等, 2012)	-	63.4
PMI (Gordon 等, 2011)	-	65.4
PMI+Connectives (Luo 等, 2016)	-	70.2
PMI+Con.+Phrase (Sasaki 等, 2017)	-	71.4
CausalBERT-base-uncased	最大边际	67.8
CausalBERT-large-uncased	最大边际	70.2
CausalRoBERTa-base	最大边际	74
CausalRoBERTa-large	最大边际	77.8
CausalALBERT-base	最大边际	62
CausalALBERT-large	最大边际	68.2
CausalALBERT-xxlarge	最大边际	86.4

表 4-6 在 COPA 上微调输出层的实验结果

模型	损失函数	COPA 结果	B-COPA 结果
BERT-large (Kavumba 等, 2019)	交叉熵	71.7	70.5
RoBERTa-large (Kavumba 等, 2019)	交叉熵	76.4	76.7
BERT-base-uncased	最大边际	71.4	73.9
BERT-large-uncased	最大边际	74.4	74.3
RoBERTa-base	最大边际	74	73.4
RoBERTa-large	最大边际	79.7	78.8
ALBERT-base	最大边际	69	72
ALBERT-large	最大边际	73.6	73.4
ALBERT-xxlarge	最大边际	82.5	83.7
CausalBERT-base-uncased	最大边际	72.3	75.1
CausalBERT-large-uncased	最大边际	76	75.8
CausalRoBERTa-base	最大边际	77.2	76.6
CausalRoBERTa-large	最大边际	85.4	85.5
CausalALBERT-base	最大边际	69.8	73
CausalALBERT-large	最大边际	75.2	76.8
CausalALBERT-xxlarge	最大边际	90.7	91.4

4.6 本章小结

本章介绍了基于统计的因果关系强度计算方法，并提出了基于预训练模型的因果关系强度计算方法。基于预训练模型的因果强度计算方法在大规模因果数据上对 BERT、RoBERTa、ALBERT 等模型进行进一步预训练，使其学习建模因果关系强度的能力。事件间的因果关系强度并不是 0-1 二值的，而是分布在 0 与 1 之间的连续实数，因此本章采用最大边际损失而非传统的交叉熵损失训练模型，使模型输出的因果强度分数尽可能均匀地分布在 0 与 1 之间，而不会集中于 0 与 1 附近。我们在 COPA 因果推理数据集上验证了两种方法的效果，并探索了基于预训练模型的方法在 BERT、RoBERTa、ALBERT 三种模型结构上的效果。实验结果显示，基于预训练模型的方法结合 ALBERT 在各项实验设置中都取得了最佳的效果，表明该方法能够有效地使模型学习到丰富的因果知识，并准确地对因果强度进行建模。

第5章 金融事理图谱构建系统的设计与实现

5.1 引言

在前面几章的内容中，我们分别就事件因果关系抽取、因果关系强度计算、事件表示学习等事理图谱构建过程中的几个关键技术进行了研究。本文第二章探索了基于序列标注模型的事件因果关系抽取方法，该方法是一种端到端的方法，能够同时进行因果关系的识别与事件文本的抽取，相比早期研究中先抽取事件、再判断因果关系的流水线方法在错误级联问题与运行效率上有了较大改进。具体地，首先探索了基于预训练语言模型的有监督因果事件抽取方法，该方法能够在训练数据充足的情况下取得令人满意的效果。此外，本章还探索了基于噪声模型的半监督因果事件抽取方法，充分利用大规模无标注数据进一步提升模型性能，以缓解该任务上有标注数据稀缺的问题。中、英文两个因果抽取数据集上的实验结果验证了以上方法的有效性。

第三章探索了常识信息增强的事件表示学习方法。事件是事理图谱的核心元素，准确地建模事件语义对于事理图谱的构建以及事理图谱在下游任务中的应用都有着非常重要的作用。传统的事件表示学习方法只考虑了事件元素文本的字面信息，而忽略了事件背后隐含的领域信息与常识信息，这些信息包括事件主客体间的实体关系、事件施事者的意图以及事件的情感极性信息等。本章在张量神经网络基础上，提出了融合实体关系、意图、情感信息的事件表示联合学习框架，通过构造三个额外的训练目标，使事件表示模型在训练过程中学习到实体关系、意图、情感等常识知识。事件相似度、脚本事件预测、股市预测等多个下游任务上的实验结果验证了该方法的效果。基于本章提出的事件表示方法，我们能够更准确地建模事件语义，以对文本中抽取出的具体事件进行泛化以抽象，得到更具有一般性的事件演化逻辑知识。

第四章探索了数据驱动的因果关系强度计算方法。因果关系是事理图谱中一种重要的事件关系，而因果边上的权重是事理图谱中一种重要的事理知识，能够作为特征为下游任务提供有帮助的信息。本章首先探索了基于统计的因果强度计算方法，利用词对间的共现信息计算事件对的因果强度。这一方法受到事件稀疏性的制约，因此本章继续探索了基于预训练语言模型的因果强度计算方法。在大规模因果数据上对语言模型进行进一步预训练使其学习了丰富的因果知识，在COPA因果推理数据集上的实验结果显著超过了基线方法，表明其

具有很强的建模因果强度的能力。这一系列方法为准确计算事理图谱中的因果强度提供了基础。

基于上述研究成果，本章提出了一个面向金融领域的事理图谱构建系统，并在大规模金融领域语料上构建了具有百万级别节点与关系的事理图谱。本章下文将详细介绍我们的事理图谱构建系统框架，并对构建好的事理图谱内容进行展示。

5.2 系统介绍与结构设计

我们基于第二至第四章中介绍的事件因果关系抽取、事件表示学习、因果关系强度计算等关键技术，设计了如图 5-1 所示的事理图谱构建流程，并实现了面向金融领域的事理图谱构建系统，下面具体介绍该系统中的每个模块。

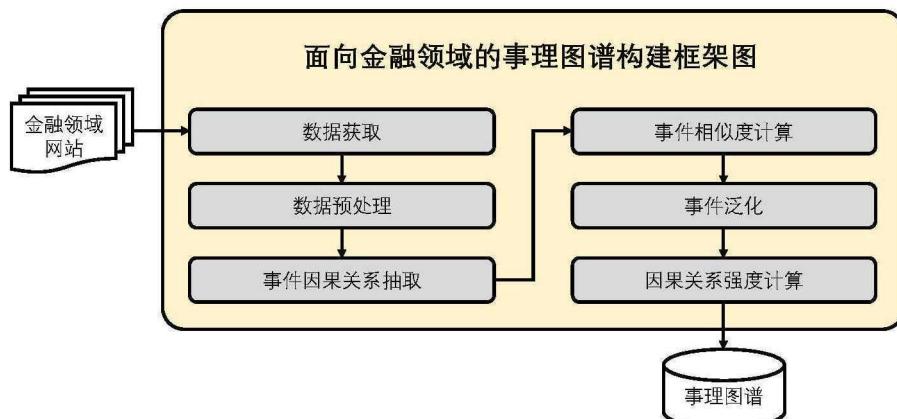


图 5-1 面向金融领域的事理图谱构建流程

(1) 数据获取 我们基于 Scrapy 工具包^[61]编写了爬虫，从腾讯、网易、股吧、和讯等十余个网站的金融板块爬取了新闻、研究报告等金融领域的文本，目前共收集了 25G 的语料数据，包含 1200 多万篇文档，作为事理图谱的数据来源。该数据获取模块目前仍在持续运行，不断地从互联网上获取新的金融领域文本，对事理图谱数据进行补充。

(2) 数据预处理 我们首先对互联网上爬取的金融语料进行清洗，以去除其中的标记语言与链接等。之后，我们对语料中的文本进行归一化，其中包括将除小数点外的标点符号统一为中文，将连续的空白字符合并为一个，将全角的数字及英文字符统一为半角等。我们进一步使用 LTP^[62]自然语言处理工具包对语料进行预处理，使用 LTP 的分句工具将原始语料切分为句子，并对其进行分词、词性标注与依存句法分析。预处理步骤得到的结果与语料一起保存在文件中，共后续处理步骤使用。

(3) 事件因果关系抽取 我们使用第二章提出的基于噪声模型的半监督

学习方法训练事件因果关系抽取模型，并将训练好的模型应用于事理图谱的构建。爬取好的金融领域文本经预处理及分词后输入事件因果关系抽取模型，模型对输入单词序列生成符合 BIO 标注规范的标签序列，最后由该标签序列解码得到原因事件短语与结果事件短语。我们将抽取出的原因、结果事件与因果对所在的上下文文本一起保存在事理图谱中。

(4) 事件相似度计算 我们为金融语料中抽取出的具体事件计算语义相似度，为后续的事件泛化步骤提供基础。我们结合杰卡德相似度与事件表示余弦相似度度量事件对的相似程度。杰卡德相似度将两个事件视为词袋，计算两个事件的交集大小除以其并集大小作为事件相似度。事件表示余弦相似度首先使用第四章中训练的模型将事件映射为一个向量表示，之后计算两事件向量的余弦相似度。我们保留相似度高于阈值（0.4）的事件相似关系，以无向边的形式保存在事理图谱中。

(5) 事件泛化 我们基于事件相似度的计算结果对抽取好的具体事件进行泛化。我们首先对事件的可泛化性进行判断，若一个事件相似度大于 0.4 的相似事件数不足 3 个，则认为这个事件缺少一般性而不对其进行泛化。对于有较多相似事件的具体事件，我们提取这些事件中的公共成分作为泛化后的事件。具体地，我们筛选出在 5 个以上相似事件中出现过的单词，并判断公共词中是否包含至少一个动词或形容词词性的词，若是，我们将其作为一个泛化后的事件。这里，我们认为一些形容词（如“低房价”中的“低”）体现了事物处于某种状态，因此认为形容词也可以指示一个事件的发生。我们首先进行对事件的泛化，之后若两个具体事件间有一条因果关系边相连，则在其泛化后的事件间也加入一条泛化的因果关系边，进而对因果关系也进行了泛化。

(6) 因果关系强度计算 我们使用第四章提出的数据驱动的因果关系强度计算方法，为泛化后的因果事件对计算因果关系强度。我们综合使用基于统计的因果关系强度计算方法与基于预训练模型的因果强度计算方法，在基于统计的方法中，基于两事件中所有单词对之间的共现信息计算因果强度；在基于预训练模型的方法中，将两个事件的文本拼接后输入模型，由模型输出两事件的因果强度分数。计算好的因果强度作为因果关系边上的权重记录在事理图谱中。

5.3 系统展示

我们为构建好的金融事理图谱搭建了可视化系统进行展示。系统主页面如图 5-2 所示，支持对图谱中泛化后的事件以及因果关系的查询。页面左上方为

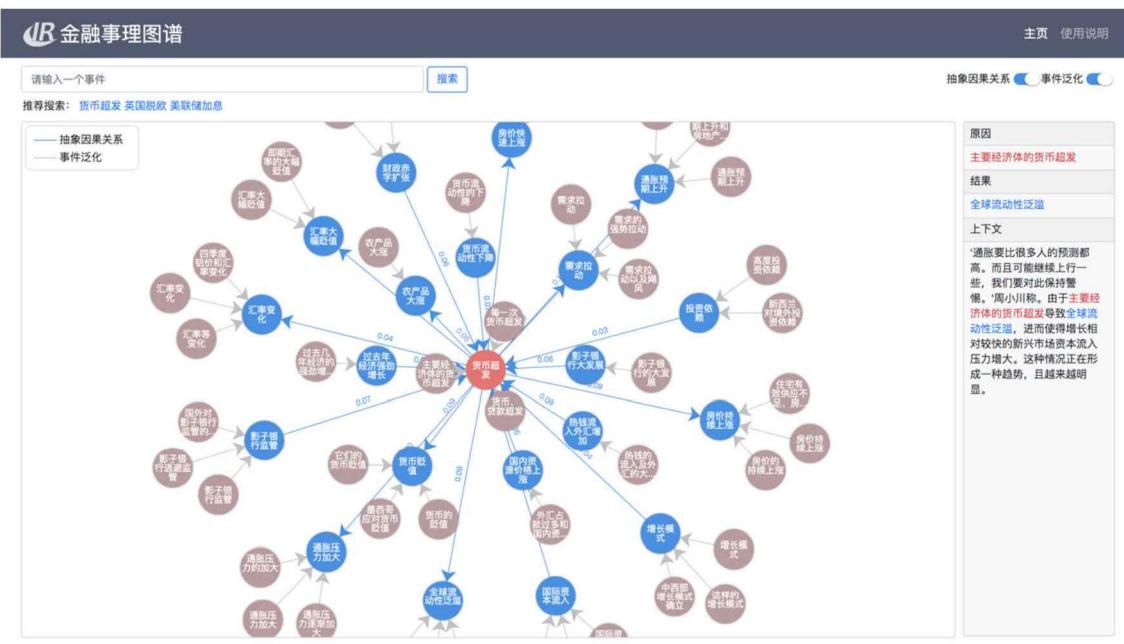


图 5-2 泛化事件查询页面

事件搜索框，用户可在其中输入感兴趣的文本，并点击“搜索”按钮在事理图谱中搜索相关的事件。搜索功能支持模糊匹配，用户输入的文本不一定要与图谱中的事件文本完全一致，也能够返回搜索结果。搜索结果以可视化的形式展示在页面中央的图谱中。图谱中央为用户搜索的事件节点，以红色高亮显示。系统会同时为该事件查询多个与其构成因果关系的事件，以蓝色的节点显示在图谱上，这些节点与中心节点间以蓝色的因果边相连，边由原因节点指向结果节点，边上显示了两事件节点的因果强度数值。当用户用鼠标划过因果边时，图谱右侧的信息栏会显示出该因果事件对的详细信息，包括完整的原因、结果事件文本以及该事件对在语料原文中的上下文，上下文中原因、结果事件分别以红色、蓝色高亮显示。用户还可以双击图谱中的蓝色因果事件节点，对图谱进行进一步扩展，在页面上显示更多的事件与因果关系。图谱中每个蓝色泛化事件节点的周围还有多个灰色的具体事件节点环绕，并存在灰色的事件泛化边将具体事件与泛化后的事件相连。页面右上角的开关可以控制图谱中各种类型边的显示与隐藏，帮助用户更清晰地查看所关注的内容。

此外，系统中还存在一个用于查询具体事件及因果关系的页面，如图 5-3 所示。该页面中展示的红色中心节点与蓝色的因果事件节点均为具体事件，便于更直观地查看事件因果关系抽取结果。

本系统还制作了使用说明页面，用户点击主页面右上方的“使用说明”链接即可跳转。使用说明页面中使用可动的 GIF 图片展示了本系统的使用方法，

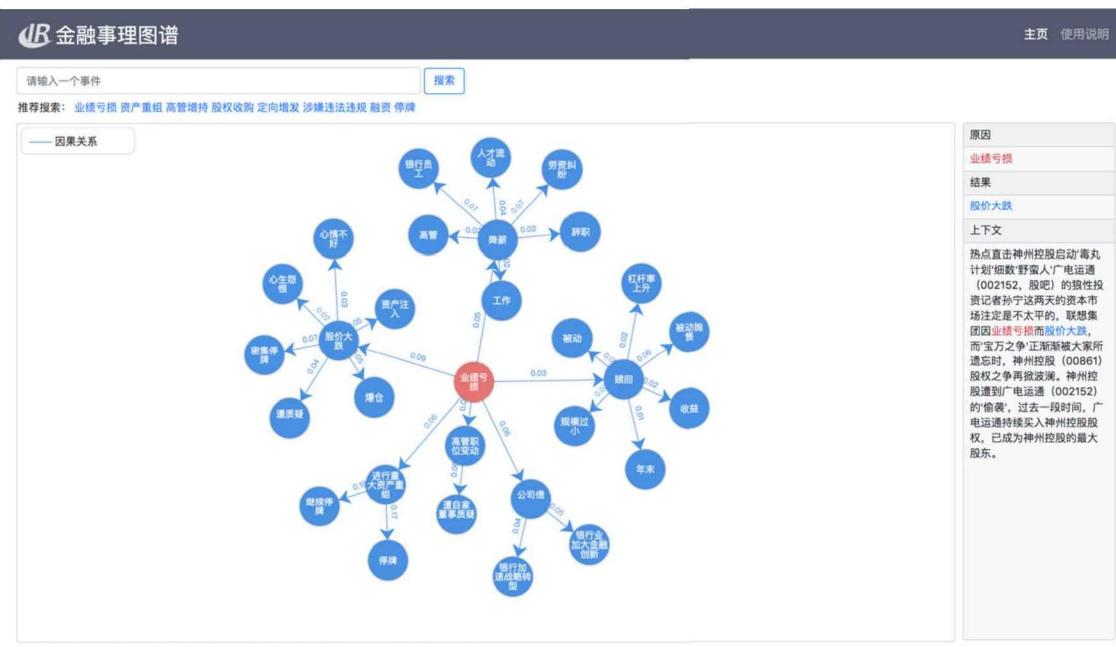


图 5-3 具体事件查询页面

以帮助用户较快地熟悉本系统的使用方法，提高系统的易用性。

5.4 本章小结

本章基于第二、三、四章的研究内容，设计并实现了面向金融领域的事理图谱构建系统，该系统包括事件因果关系抽取、事件相似度计算、事件泛化、因果关系强度计算等完整的事理图谱构建流程，能够自动从大规模语料中挖掘事理知识，构建事理图谱。目前，我们已经构建了包含百万级别的事件与因果关系的金融事理图谱，并且仍在持续不断从互联网爬取新的数据，以实现事理图谱的持续更新。本章进一步开发了可视化的用户界面，使用户能够简单方便地对事理图谱数据进行访问。

结 论

人工智能对世界的认识与理解离不开事件演化规律知识，凸显了事理图谱作为事理逻辑知识库的重要性。本文从金融领域出发，探索从文本中自动挖掘事理知识并构建事理图谱的关键技术。因果关系是一种重要的事件演化关系，本文首先探索了事件因果关系抽取方法，以自动挖掘事件间的因果关系。事理图谱的构建与应用离不开对事件的理解，为了更好地建模事件的语义，本文提出了常识信息增强的事件表示学习方法，为事件学习蕴含丰富语义信息的向量表示。为了准确地建模因果关系的强弱，本文进一步提出了数据驱动的方法计算事件对的因果关系强度。最后，本文基于上述研究成果，设计并实现了面向金融领域的事理图谱构建系统，并构建了金融事理图谱。

本文针对面向金融领域的事理图谱构建关键技术研究这一课题，对端到端的事件因果关系抽取、常识信息增强的事件表示学习、数据驱动的因果关系强度计算以及金融事理图谱构建系统的设计与实现进行了研究，主要工作和贡献如下：

(1) 本文将事件因果关系抽取建模为序列标注任务，提出使用序列标注模型端到端地同时进行因果关系的识别与相关事件的抽取，并具体探索了基于预训练模型的有监督事件因果关系抽取与基于噪声模型的半监督事件因果关系抽取。本文提出了结合自训练、噪声模型与条件随机场的半监督因果关系抽取新方法，并在中、英文两个事件因果关系抽取数据集上验证了其效果，实验结果表明该方法可以有效地提高事件因果关系抽取任务的效果。

(2) 本文提出了常识信息增强的事件表示学习方法。事件表示学习将事件文本映射为低维稠密向量，为使用计算机对事件进行分析提供了基础。我们在张量神经网络的基础上，提出了融入意图、情感、实体关系等多种常识知识的多任务学习框架，使用三个额外的训练目标，使事件表示模型学习到事件背后蕴含的意图、情感、实体关系等常识信息。事件相似度、脚本事件预测、股市预测三个任务上的实验结果表明，该方法能有效地在事件表示中融入常识信息，并显著地提高与事件有关的下游任务上的效果。

(3) 本文探索了数据驱动的因果关系强度计算方法，为原因、结果事件对计算一个分数建模其因果关系的强弱，并作为事理图谱中因果关系边上的权重。本文提出了基于预训练模型的因果关系强度计算方法，通过在大规模因果数据上对模型预训练使其学习建模因果关系强度的能力，并与基于统计的因果关系强度计算方法进行了对比。COPA 因果推理数据集上的实验结果表明，基于预

训练模型的方法显著超过了基于统计的方法。本文进一步探索了 BERT、RoBERTa、ALBERT 等不同的模型结构在该方法上的效果。

(4) 基于上述研究成果，本文设计并实现了面向金融领域的事理图谱构建系统。该系统能够自动地从互联网上收集金融领域文本，从中抽取事件间的因果关系，将其泛化为事理知识，并构建事理图谱。本文基于该系统构建了包含百万级别事件与因果关系的金融事理图谱，并开发了可视化系统对图谱内容进行展示，目前系统已部署到互联网，以期为后续研究提供参考。

尽管本文在事件因果关系抽取、因果关系强度计算、事件表示学习三项事理图谱构建中的关键技术上取得了不错的结果，但事理图谱的构建仍存在许多值得研究的问题。例如，如何有效地对事件间的其他关系进行抽取，包括事件间的顺承关系与上下位关系等；再如，如何有效地对事件进行泛化，由文本中抽取的具体事件关系得到高度泛化的事件演化规律知识等。上述问题有待未来的研究继续探索与解决。

参考文献

- [1] Luo Z, Sha Y, Zhu K Q, et al. Commonsense causal reasoning between short texts[C] // Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning. 2016.
- [2] Roemmele M, Bejan C A, Gordon A S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning[C] // 2011 AAAI Spring Symposium Series. 2011.
- [3] Zhao S, Wang Q, Massung S, et al. Constructing and embedding abstract event causality networks from text snippets[C] // Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 2017 : 335-344.
- [4] Schuler K K. VerbNet: A broad-coverage, comprehensive verb lexicon[J], 2005.
- [5] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11) : 39-41.
- [6] Li Z, Zhao S, Ding X, et al. EEG: Knowledge Base for Event Evolutionary Principles and Patterns[C] // Chinese National Conference on Social Media Processing. 2017 : 40-52.
- [7] Li Z, Ding X, Liu T. Constructing narrative event evolutionary graph for script event prediction[J]. arXiv preprint arXiv:1805.05081, 2018.
- [8] 赵森栋. 基于文本的因果关系抽取与推理[D]. [S.l.] : 哈尔滨工业大学, 2018.
- [9] Asghar N. Automatic extraction of causal relations from natural language texts: a comprehensive survey[J]. arXiv preprint arXiv:1605.07895, 2016.
- [10] Grishman R. Domain modeling for language analysis[R]. [S.l.] : NEWYORKUNIV NY, 1988.
- [11] Kaplan R M, Berry-Rogghe G. Knowledge-based acquisition of causal relationships in text[J]. Knowledge Acquisition, 1991, 3(3) : 317-337.
- [12] Garcia D, others. COATIS, an NLP system to locate expressions of actions connected by causality links[C] //International Conference on Knowledge Engineering and Knowledge Management. 1997 : 347-352.
- [13] Chan K, Lam W. Extracting causation knowledge from natural language texts[J]. International Journal of Intelligent Systems, 2005, 20(3) : 327-358.
- [14] Girju R, Moldovan D I, others. Text mining for causal relations.[C] //FLAIRS conference. 2002 : 360-364.
- [15] Girju R. Automatic detection of causal relations for question answering[C] // Proceedings of the ACL 2003 workshop on Multilingual summarization and

- ques-tion answering-Volume 12. 2003 : 76-83.
- [16] Quinlan J R. C4. 5: programs for machine learning[M]. [S.I.] : Elsevier, 2014.
- [17] Chang D-S, Choi K-S. Causal relation extraction using cue phrase and lexical pair probabilities[C] // International Conference on Natural Language Processing. 2004 : 61-70.
- [18] Blanco E, Castell N, Moldovan D I. Causal relation extraction.[C] // Lrec. 2008.
- [19] Breiman L. Bagging predictors[J]. Machine learning, 1996, 24(2) : 123-140.
- [20] Girju R, Nakov P, Nastase V, et al. Semeval-2007 task 04: Classification of semantic relations between nominals[C] // Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). 2007 : 13-18.
- [21] Girju R, Beamer B, Rozovskaya A, et al. A knowledge-rich approach to identifying semantic relations between nominals[J]. Information processing & management, 2010, 46(5) : 589-610.
- [22] Sil A, Huang F, Yates A. Extracting action and event semantics from web text[C] // 2010 AAAI Fall Symposium Series. 2010.
- [23] 付剑锋, 刘宗田, 刘炜, 等. 基于层叠条件随机场的事件因果关系抽取[J]. 模式识别与人工智能, 2011, 24(4) : 567-573.
- [24] Zhao S, Liu T, Zhao S, et al. Event causality extraction based on connectives analysis[J]. Neurocomputing, 2016, 173 : 1943-1950.
- [25] De Silva T N, Zhibo X, Rui Z, et al. Causal relation identification using convolutional neural networks and knowledge based features[J]. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 2017, 11(6) : 696-701.
- [26] Kruengkrai C, Torisawa K, Hashimoto C, et al. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks[C] // Thirty-First AAAI Conference on Artificial Intelligence. 2017.
- [27] Dasgupta T, Saha R, Dey L, et al. Automatic extraction of causal relations from text using linguistically informed deep neural networks[C] // Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue. 2018 : 306-316.
- [28] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8) : 1735-1780.
- [29] Dunietz J, Carbonell J G, Levin L. DeepCx: A transition-based approach for shallow semantic parsing with complex constructional triggers[C] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language

- Processing. 2018 : 1691- 1701.
- [30] Li Z, Li Q, Zou X, et al. Causality Extraction based on Self-Attentive BiLSTM-CRF with Transferred Embeddings[J]. arXiv preprint arXiv:1904.07629, 2019.
- [31] Weber N, Balasubramanian N, Chambers N. Event representations with tensor-based compositions[C] // Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
- [32] Ding X, Zhang Y, Liu T, et al. Deep learning for event-driven stock prediction[C] // Twenty-fourth international joint conference on artificial intelligence. 2015.
- [33] Ding X, Zhang Y, Liu T, et al. Knowledge-driven event embedding for stock prediction[C] // Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers. 2016 : 2133-2142.
- [34] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [35] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
- [36] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.
- [37] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C] // Advances in neural information processing systems. 2017 : 5998-6008.
- [38] Wu Y, Schuster M, Chen Z, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv:1609.08144, 2016.
- [39] Lafferty J, McCallum A, Pereira F C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J], 2001.
- [40] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods[C] // 33rd annual meeting of the association for computational linguistics. 1995 : 189-196.
- [41] Maeireizo B, Litman D, Hwa R. Co-training for predicting emotions with spoken dialogue data[C] // Proceedings of the ACL Interactive Poster and Demonstration Sessions. 2004 : 202-205.
- [42] McClosky D, Charniak E, Johnson M. Effective self-training for parsing[C] // Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics. 2006 : 152-159.
- [43] Bekker A J, Goldberger J. Training deep neural-networks based on unreliable labels[C] //2016 IEEE International Conference on Acoustics, Speech and

- Signal Processing (ICASSP). 2016 : 2682-2686.
- [44] Goldberger J, Ben-Reuven E. Training deep neural-networks using a noise adaptation layer[J], 2016.
- [45] Paul D, Singh M, Hedderich M A, et al. Handling Noisy Labels for Robustly Learning from Self-Training Data for Low-Resource Sequence Labeling[J]. arXiv preprint arXiv:1903.12008, 2019.
- [46] Prasad R, Dinesh N, Lee A, et al. The Penn Discourse TreeBank 2.0.[C] // LREC. 2008.
- [47] Dunietz J, Levin L, Carbonell J G. The BECaSE corpus 2.0: Annotating causality and overlapping relations[C] // Proceedings of the 11th Linguistic Annotation Workshop. 2017 : 95-104.
- [48] Chen Q, Zhuo Z, Wang W. Bert for joint intent classification and slot filling[J]. arXiv preprint arXiv:1902.10909, 2019.
- [49] Li Z, Chen T, Van Durme B. Learning to Rank for Plausible Plausibility[J]. arXiv preprint arXiv:1906.02079, 2019.
- [50] Kavumba P, Inoue N, Heinzerling B, et al. Balanced COPA: Countering Superficial Cues in Causal Reasoning[J], .
- [51] Buck C, Heafield K, Van Ooyen B. N-gram Counts and Language Models from the Common Crawl.[C] // LREC : Vol 2. 2014 : 4.
- [52] Kavumba P, Inoue N, Heinzerling B, et al. When Choosing Plausible Alternatives, Clever Hans can be Clever[J]. arXiv preprint arXiv:1911.00225, 2019.
- [53] Wang A, Pruksachatkun Y, Nangia N, et al. Superglue: A stickier benchmark for general-purpose language understanding systems[C] // Advances in Neural Information Processing Systems. 2019 : 3261-3275.
- [54] Sap M, Rashkin H, Chen D, et al. Socialqa: Commonsense reasoning about social interactions[J]. arXiv preprint arXiv:1904.09728, 2019.
- [55] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
- [56] 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.
- [57] Sap M, Le Bras R, Allaway E, et al. Atomic: An atlas of machine commonsense for if-then reasoning[C] // Proceedings of the AAAI Conference on Artificial Intelligence : Vol 33. 2019 : 3027-3035.
- [58] Cambria E, Poria S, Hazarika D, et al. SenticNet5: Discovering conceptual primitives for sentiment analysis by means of context embeddings[C] // Thirty-

- Second AAAI Conference on Artificial Intelligence. 2018.
- [59] Zhang Y, Weiss D. Stack-propagation: Improved representation learning for syn-tax[J]. arXiv preprint arXiv:1603.06598, 2016.
- [60] Kartsaklis D, Sadrzadeh M. A study of entanglement in a categorical framework of natural language[J]. arXiv preprint arXiv:1405.2874, 2014.
- [61] Myers D, McGuffee J W. Choosing scrapy[J]. Journal of Computing Sciences in Colleges, 2015, 31(1) : 83-89.
- [62] Che W, Li Z, Liu T. Ltp: A chinese language technology platform[C] // Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. 2010 : 13-16.

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] Xiao Ding, Kuo Liao, Ting Liu, Zhongyang Li, Junwen Duan. Event Representation Learning Enhanced with External Commonsense Knowledge[C]. In 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). Hong Kong, China. 2019.11.
- [2] Xiao Ding, Dingkui Hao, Yuewei Zhang, Kuo Liao, Zhongyang Li, Bing Qin, Ting Liu. HIT-SCIR at SemEval-2020 Task 5: Training Pre-trained Language Model with Pseudo-labeling Data for Counterfactuals Detection. SemEval 2020.
- [3] Zhongyang Li, Xiao Ding, Kuo Liao, Jinglong Gao, Bing Qin, Ting Liu. CausalBERT: Injecting Causal Knowledge Into Pre-trained Models with Minimal Supervision[C]. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020). (在投)
- [4] 廖阔, 丁效, 秦兵, 刘挺, 黄虎杰. 事件表示学习综述[J]. 智能计算机与应用, 2020.06.

(二) 申请及已获得的专利

- [1] 丁效, 刘挺, 秦兵, 廖阔. 一种基于自训练与噪声模型的因果事件抽取方法: 中国, 202010397785.6[P]. 2020-05-12.

致 谢

光阴似箭，日月如梭，转眼间我也即将成为一名硕士毕业生，在此，谨向大学生活中每一位关心、帮助过我的老师、同学、朋友、亲人表达最衷心的感谢。

感谢刘挺教授，是您给了我进入实验室学习的机会，带我走进 SCIR 这个大家庭，使我有幸结交各位良师益友。您“以中文技术，助民族复兴”的家国情怀，一丝不苟的学术态度，引领实验室发展方向的高瞻远瞩，以及充满激情、充满乐趣的人生态度无不使我钦佩不已，是您为我树立了终生学习与奋斗的楷模。在今后的工作岗位上，我将继续铭记您和实验室的教诲，践行“友爱，力行，乐学，日新”，努力成为一个“心中有情怀，做人讲情义，生活有情趣”的人。

感谢秦兵教授，您就像实验室的母亲一般，为每一位同学带来关怀与温暖。您不仅在课堂上向我们传授学术知识，更在日常生活中向我们传授做人的道理。感谢车万翔教授，是您精彩的授课为我打开了计算语言学的大门，您永远奋战在最前沿的科研态度是我终身学习的榜样。

感谢丁效老师，是您在我两年的硕士生涯中给了我最多的指导。犹记得刚加入实验室时，我对实验室的研究方向还很陌生，感谢您愿意让这样的我参与到论文与项目工作中，使我能够在实践中迅速进入科研状态。您对我的指导不仅是学术上的，更涵盖了表达、交流、为人处事的各个方面，使我从一名不善言辞的学生成长为一位有信心应对各种场合的工程师，使我受益终生。

感谢张宇老师、刘铭老师、张伟男老师、赵妍妍老师、冯骁骋老师在我硕士期间对我的指导与帮助，老师的教诲我将永远铭记于心。

感谢实验室每一位师兄、师姐、师弟、师妹，以及与我同届并肩奋斗的战友们，感谢你们两年来对我的帮助，我也向你们学习了很多。

感谢我的室友姜庆彬同学，感谢你在六年大学生活中对我的帮助。

感谢我的女友尹莉娜同学，感谢你一直以来的陪伴与鼓励，陪我走过最痛苦最艰难的时刻。谢谢你愿意一直等我。

最后，感谢我的父母，感谢你们对我默默的支持与关注。

中国优秀硕士学位论文全文数据库 信息科技辑
Chinese Master's Theses Full-text Database Information Science and Technology

中国优秀硕士学位论文全文数据库
信息科技辑
(月刊)
2021 年第02期
2021-01-16—2021-02-15出版

计算机软件及计算机应用

面向金融领域的事理图谱构建关键技术研究

廖阔 I138-2530

共 1 条 < 1 >

主 管: 教育部

主 办: 清华控股有限公司

编辑出版: 中国学术期刊(光盘版)电子杂志社

地 址:

邮政编码: 100084

电 话:

E-mail:

发 行: 同方知网技术有限公司

发行范围: 国内外公开发行

刊 号: ISSN 1674-0246 CN 11-9144/G

订 购 处:

订购电话:

开 户 银 行:

账 号:

本期定价: