EDeR：用于探索事件间依赖关系的数据集

RuiqiLi♠, **PatrikHaslum**♠, **LeyangCui**♡♠ 澳大利亚国立大学♡ 腾讯人工智能实验室{*ruiqi.li, patrik.haslum*}*@anu.edu.au leyangcui@tencent.com*

arXiv:2304.01612v1 [cs.CL] 4 Apr 2023

摘要

关系抽取是自然语言处理（NLP）和信息检索（IR）研究中的一项核心任务。我们认为，迄今为止，自然语言处理或信息检索研究尚未探索的一种重要关系类型，是一个事件作为另一个事件的论元——必需论元或可选论元。我们介绍人工标注的 **事件依存关系**数据集（EDeR），该数据集提供了这种依赖关系。标注是在 OntoNotes 数据集中的一批文档上完成的，这还带来了一项额外优势：它可以与该数据集已有的、彼此正交的标注相结合。我们研究了用于预测事件依存关系的基线方法，其中最佳方法在论元／非论元二分类任务上取得了 82.61% 的准确率。我们表明，识别这种关系能够提高事件抽取（语义角色标注）的准确性，并能改进依赖于此的下游任务，例如共指消解。此外，我们还证明，将样本划分为必需论元、可选论元或非论元的三分类任务更具挑战性。

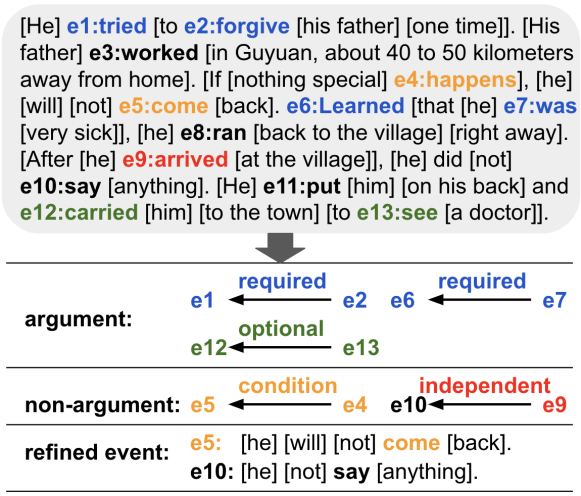


图 1：事件依存关系示例。上方：源文本。（事件谓词以粗体标记，论元跨度以方括号标记。）下方：事件依存关系和细化事件。论元：具有依存关系的事件对，以及该关系是必需还是可选；非论元：不存在依存关系的事件对，以及右侧事件是左侧事件的条件还是独立于左侧事件；细化事件：非论元事件对中的左侧事件，其论元跨度经过细化。

1 引言

事件对于帮助人工智能智能体理解和感知世界至关重要，因为它们提供了关于发生了什么以及涉及哪些实体的信息。事件由谓词（即动词）和论元构成，其中谓词表示事件的动作，论元则表示谓词的主语、宾语等（Levin 等人，1999年；Rappaport Hovav 等人，2010年）。表示并提取事件之间的依存关系，以识别语义关系（例如，指出事件谓词的宾语是什么）和句法关系（例如，指出事件谓词的关系从句是什么），是人工智能研究者感兴趣的一项任务。这类事件依存关系还提供了有益信息，可帮助一些信息检索（IR）和自然语言处理（NLP）中的下游任务。

先前在信息检索和自然语言处理领域开展的关系抽取工作主要集中于研究实体之间的关系（Miwa and Sasaki，2014年；Huguet Cabot and Navigli，2021年；Chen et al.，2020年；Ma et al.，2022年），而非事件。在相对有限的研究事件间关系的工作中，所考虑的关系类型包括因果关系（Mirza and Tonelli，2016年；Mariko et al.，2020年，2022年；Yang et al.，2022年）、时- 间关系（Bethard，2013年；Laokulrat et al.，2013年；Cham- bers，2013年；Chambers et al.，2014年；Mirza and Tonelli，2016年）以及层级关系（Hovy et al.，2013年；Glavaš et al.，2014年；Glavaš and Šnajder，2014年；

Araki 等人，2014年）。

作为补充，我们研究一个事件是另一个事件的关系**论元**，而非**独立**事件。例如，在图1中的句子“He tried to forgive his father one time.”里，$e2$ 的谓词“forgive”是他“tried”（$e1$）要采取的行动。然而，$e2$ 与 $e1$之间不存在时间、因果或层级关系；事实上，句子暗示 $e2$ （“forgive”）实际上并未发生。我们还区分了**必需论元**和**可选论元**事件。一个事件要完整且有意义，必须包含必需论元。例如，在图1中，$e2$ （“forgive”）必须充当$e1$ （“tried”）的论元；如果删除它，$e1$ 就变成了“He tried to”，既不完整也没有意义（“try to”必须是尝试做某事）。可选论元只会补充或阐明其所充当论元的事件，例如指出目的。在图1中，$e13$ 的谓词“see”就是 $e12$ （“carried”）的一个可选论元。虽然是可选的，但它具有论元事件的特性：句子并未说明该事件实际发生。相比之下，非论元事件是并列或条件事件。然而，在事件抽取中，此类非论元事件常被误标为论元（例如，被语义角色标注（SRL）系统误标）。例如，事件$e9$（“arrived”）虽然包含在事件 $e10$ （“did not say anything”）的时间修饰语中，却是独立的。论元事件关系在自然语言文本中十分常见，尤其是在新闻文章、对话等叙事文本中。我们认为，识别事件之间的这种依赖关系，有助于更准确地抽取事件及其之间的其他关系。

受此启发，我们提出了一个由人工标注的**事件依赖关系**数据集（EDeR），该数据集：（1）基于 OntoNotes（Pradhan et al.，2013年）中的一组文档提取事件依存信息，并与该数据集的正交标注相结合；（2）基于这些信息提供经过细化的语义角色标注事件表示。

我们采用基于多层级质量检查的方法，对 OntoNotes 数据集内 7 种不同体裁的 275 篇文章中的事件对进行标注。最终，我们构建了一个事件依存关系数据集（EDeR），其中包含 11,852 条高质量标注样本，并依据所提出的事件关系分类体系进行分类。该数据集可作为有效自动分类流程的基础。我们采用了多种文本分类模型，从制定启发式规则到微调基于 Transformer（Vaswani et al.，2017年）的预训练模型。二元论元／非论元分类结果为预测事件依赖关系奠定了坚实的基线。

此外，我们将 EDeR 人工标注的事件依存关系，以及性能优于基线模型所预测的事件依存关系，应用于指导当前最先进的 SRL 模型（Zhang 等人，2022年）生成事件表示。实验结果表明，我们的事件依存关系能够提升 SRL 模型在提取 EDeR 数据集所提供的更新后事件表示方面的性能。我们进一步将这些事件表示应用于下游任务：共指消解（CR）。将使用原始事件表示的当前最先进 CR 模型（Jiang 和 Cohn，2021年）的性能与使用我们更新后事件表示的模型性能进行比较，结果证明了 EDeR 数据集所提供的细化事件表示的合理性和有效性。

此外，我们还研究了三分类（即必需论元、可选论元或非论元）的基线性能。实验结果表明，对于这些语言模型而言，这是一项更具挑战性的任务。我们在以下地址发布了 EDeR 和基线系统的源代码：
https://github.com/RichieLee93/EDeR。

2 背景与相关工作

2.1 事件与事件表示

事件提及由一个动词或短语动词作为谓词，以及一组带标签的论元构成（Levin 等人，1999年；Rappaport Hovav 等人，2010年）。通过对给定句子中的词或词语片段进行语义角色标注（SRL），标记其角色（即谓词和论元），即可从文本中提取事件。OntoNotes 数据集采用 PropBank 的语义角色标签标注规范（Bonialetal.，2012年），其中论元是文本片段，论元标签分为编号论元（ARG0–ARG5）和动词的修饰语。编号论元表示动作配价所要求的论元（例如施事和受事）；修饰语则包括目的（PRP）、处所（LOC）等。示例见图2上半部分。

此外，还有一些用于特定共享语义角色标注任务的公开数据集（例如 CoNLL-2005 和 CoNLL-2012），如 CoNLL 2005（Carreras andMàrquez，2005年）、CoNLL 2009（Hajič et al.，2009年）和 CoNLL 2012（Pradhanet al.，2012年）任务。CoNLL-2012 已并入 OntoNotes V5.0 数据集，因此本文其余部分将该数据集称为 OntnoNotes。

一个事件本身可以是另一个事件的论元。在经过标注的语义角色标注数据集中，我们发现这类论元事件包含在其所充当论元的事件跨度内；但反过来则不成立：包含在另一事件跨度内的事件并不总是论元事件，也可能与包含事件共享部分论元（主语或宾语）。例如，如图 2所示，事件“The man [who]works here”包含在事件“The man who works here tells me to get the hell out”中，但并非“tells”的论元。这类情况表明，这些语义角色标注数据集中的事件标注并未考虑事件间的论元事件关系。明确这一关系有助于我们选取更聚焦的论元跨度，例如在上述例子中，为谓词“tells”对应的事件选取“ The man”作为ARG0（施事）。

2.2 事件间关系

数十年来，已有多种事件-事件关系被提出。TimeBank（Pustejovsky etal.，2003年）、TB-Dense（Cassidy etal.，2014年）和MATRES（Ning etal.，2018年）数据集关注时间关系，用以揭示一个事件是否发生在另一个事件之前、之后等。CausalTimeBank（Mirza and Tonelli，2016年）、CIRA（Fischbach et al.，2021年）和 CNC（Tanet al.，2022年）等数据集也探讨了因果关系等逻辑关系。这些关系将两个事件视为与句法和语义无关：每个事件的发生与真实性彼此独立，不受其他事件影响。

此外，一些数据集提出了层级关系（即上位事件与子事件之间的关系），其中子事件必须满足两个条件：在时间上，子事件发生于上位事件期间；在空间上，子事件应包含在上位事件之中（Hovy 等人，2013年；Glavaš 等人，2014年；Glavaš 和 Šnajder，2014 年）。Araki 等人（ Araki 等人，2014年）提出了一个用于检测事件共指关系的数据集，即判断两个事件提及是否指代同一事件。这类层级关系无法处理这样的情况：某个事件（尤其是该事件的动词）需要从句补语，即事件动词的一个论元本身就是一个事件，如图1所示的“tried”和“forgive”示例。

2.3 事件关系抽取

关系抽取是信息检索研究中的一个热门分支。近期大多数研究将其作为命名实体识别（命名实体识别）的一项下游任务，着重抽取实体之间的关系，而非事件（Miwa and Sasaki，2014年；Huguet Cabot andNavigli，2021年；Chen et al.，2020年；Ma et al.，2022年）。

如前所述，事件以词跨度形式表示，这类事件表示可视为文本片段。近年来，基于Transformer（Vaswani 等人，2017）的语言模型（如 BERT、RoBERTa 和 GPT）在文本关系分类任务（例如文本蕴含检测、情感分析）上取得了显著进展（Balazs et al., 2017; Ethayarajh, 2019; Laskar et al., 2020; Lee et al., 2022）。这类语言模型在海量语料库上进行预训练，人们随后使用所选数据集针对具体任务对其进行微调（即“迁移学习”（Pan and Yang, 2010））。与仅使用传统机器学习模型或神经网络进行训练相比，这种预训练-微调过程通常能在关系抽取任务上取得更优的性能，尤其是在数据量较少时（Balazs et al., 2017; Papanikolaou et al., 2019; Seganti et al., 2021）。受此启发，我们也可以将这些最先进的语言模型用于分类事件对之间的依存关系。此外，基于不同风格的事件表示（即基于语义角色的表示（Carreras and Màrquez, 2005年; Hajič et al., 2009年; Pradhan et al., 2012, 2013）或基于谓词的表示（Pustejovsky et al., 2003年; Cassidy et al., 2014; Ning et al., 2018; Mirza and Tonelli, 2016年）），以及谓词之间的句法依存关系等其他特征，我们可以将各种输入与语言模型结合起来，探索模型在EDeR 数据集上的基线性能。

3 数据集

如上一节所述，据我们所知，目前还没有能够识别事件并标注其论元依

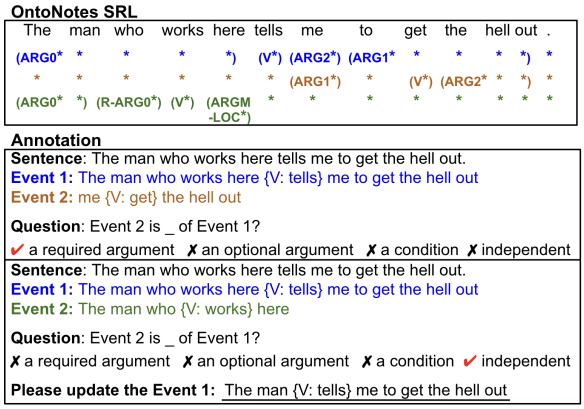


图2：上方：OntoNotes 数据集中的一个带有语义角色标签的示例句子。下方：相应的事件对，呈现给标注人员进行人工标注，并标注事件依存关系。

赖关系的自然语言文本数据集。因此，我们构建了这样一个数据集（命名为 EDeR）。鉴于 OntoNotes 提供了人工标注且采用谓词-论元格式的事件，我们从中选取部分文档，并从这些文档中提取候选事件对，其中一个事件的谓词包含在另一个事件的跨度内。人工标注者为每个事件对（即，$Event1$ 和 $Event2$）添加标签，以指明 $Event2$ 是 $Event1$的必需论元、可选论元，还是其条件，或与其独立。接下来，我们将详细介绍候选事件对的收集与预处理方法、人工标注流程以及数据集的构建。

3.1 数据收集

OntoNotes 包含采用语义角色格式的事件表示，如图2 所示。我们从七种体裁中随机抽取了 275 篇文档：广播新闻（bn）、杂志（mz）、通讯社新闻（nw）、枢轴语料库（pt）、电话对话（tc）、广播对话（bc）和网络数据（wb）。这 275 篇原始文档的数据统计见表1。各体裁抽取的文档数量，以及将其划分为训练集、开发集和测试集的方式，均遵循 OntoNotes 数据集中的原始分布。

3.1.1 候选事件对提取

由于论元是文本跨度，抽取出的事件的部分或全部可能位于另一个事件的论元中。如果e_j 的动词（即谓词）位于 e_i的某个论元中，我们称 e_j 被包含在 e_i中。这可以嵌套。被包含事件是论元事件的候选项，但不一定就是论元事件。例如，在图2上半部分，绿色事件（**e3** = {ARG0：这个男人 } {R-ARG0：他 } {V：在这里工作 } { ARGM- LOC： }）包含在蓝色事件（**e1** = {ARGO：这个在这里工作的男人 }{V：告诉 } {ARG2：我 } {ARG1：滚出去 }）的跨度内，但 e3 不是 e1 的论元；然而，橙色事件（**e2** = {ARG1：我 } {V：滚 }{ARG2：出去 }）是 e1 的论元。

我们选取事件对（e_i， e_j）作为待标注的候选事件对，其中 e_j 的动词包含在 e_i 的某个论元跨度内。

3.1.2 预处理

我们对选出的候选事件对应用三项筛选规则。首先，筛除其中 e_j 没有任何论元的候选事件对（e_i, e_j）。当 e_j 是表示 e_i中动词时态的情态动词时，通常会出现这种情况。其次，OntoNotes 数据集有时会将形容词误认为动词（例如 “the Matrix reloaded” 中的 “reloaded”），因此我们使用词语的词性标签将这类情况筛除。事件谓词的词性标签由斯坦福 CoreNLP 工具包（Man- ning et al., 2014）获取。第三，我们移除传递包含的事件对，即满足以下条件的事件对（e_i, e_j）：存在另一个事件 e_k，使得 e_k 的动词包含于 e_i，且 e_j 的动词包含于 e_k。在这种情况下，（e_i, e_k）和（e_k, e_j）可能是候选事件对，但（e_i, e_j）不是。例如，在“[She] {V: try } [to { V: stop } [[them] from {V: ruining } [...]]]”中，“try”和“stop”以及“stop”和“ruin”是候选事件对，但“try”和“ruin”不是。

预处理后，我们从275篇文档中获得了11,852对候选事件对，用于人工标注。

3.2 标注

3.2.1 标注说明

我们展示每个候选事件对中两个事件的完整片段，并用“{V: }”突出显示每个事件的谓词，如图 2下半部分所示。标注者可以从四个选项中选择一个，以回答这两个事件之间的关系。对于标注任务，我们将非论元情形细分为两类：条件和独立（“既非论元，也非条件”）。这使各标签的定义更容易理解，因为条件和论元事件一样，都是假设的而非实际发生的事件，出现在

	bn	mz	nw	pt	tc	bc	wb	总体
# 文档	104	7	107	31	7	3	16	275
# 训练文档	90	5	89	24	5	1	13	227
# 开发文档	7	1	9	4	1	1	2	25
# 测试文档	7	1	9	3	1	1	1	23
每篇文档的平均句子数	6.2	48.9	14.4	40.7	71.7	203.8	59.3	24.9
每篇文档的平均词数	139.7	1599.7	418.7	657.0	1089.4	4045.8	1509.3	543.4
每篇文档的平均事件数	23.6	202.1	58.3	148.7	362.1	1058.6	268.5	93.5

表 1：我们从 OntoNotes 数据集中抽取用于标注的文档统计信息（按各体裁及总体统计）：文档数量及划分到不同集合（训练集、开发集和测试集）的文档数量，以及每篇文档的平均句子数、词数和事件数。

Optional argument: Event 2 is an argument of the verb of Event 1, but Event 2 only provides additional or clarifying information about Event 1 (for example, Event 2 may indicate the purpose or goal of Event 1). Event 1 is still syntactically complete and meaningful, although it may not keep the exact same meaning, if Event 2 is removed.
For example,
Sentence: A power failure with the docking door forces Brodski to go EVA to fix it.
Event 1: Brodski {V: go} EVA to fix it.
Event 2: Brodski {V: fix} it.
Explanation: Event 2 is an optional argument, indicating the purpose of Event 1. Removing it leaves Event 1 "Brodski go(es) EVA", which is still a syntactically complete and meaningful clause.

图 3：标注者说明部分内容的截图。对于四个标签中的每一个，我们都提供其定义、关系示例，并详细解释该示例为何被标注为相应标签。截图展示的是“可选论元”标签的相关内容。

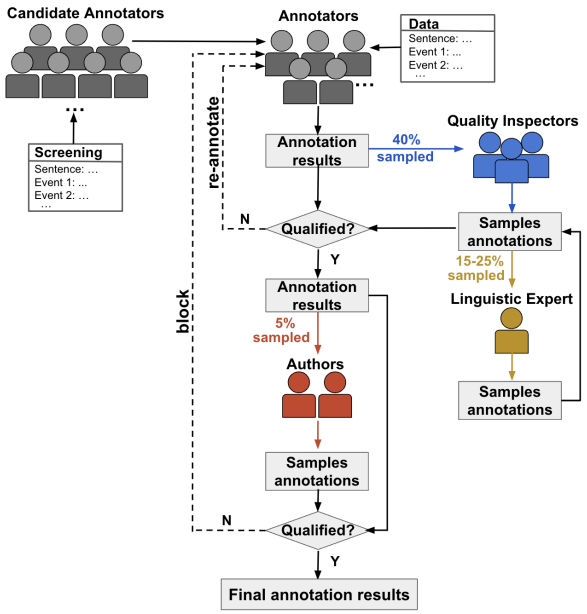


图 4：我们基于多级资格审核的标注流程示意图。

文本中。在文本中识别条件句本身就是一个有趣且活跃的研究课题（Fischbach 等人，2021年；Tan 等人，2022年），尤其是因为条件句与因果关系有关。不过，我们数据集中标记为条件的事件可能只是源文本中所有条件句的一部分，因为它们仅限于出现在包含关系的事件对中。在本文余下部分对数据集进行分析和实验时，我们将这两个非论元标签视为一类。

我们编写了标注说明，其中包括任务描述、四个标签的定义，以及每个标签的 2–3 个示例，并相应解释作出该判断的原因，如图3所示。

3.2.2 标注流程

我们采用一种基于多层级资格审核的标注流程，在多个阶段对标注者的工作进行抽样检查和修正，并将检查反馈传回标注者。图中示意了该流程4。与常用的众包和投票策略相比，该流程能让标注者在整个任务过程中不断学习并提升。尽管最终并非所有事件对都由多名参与者进行标注，但过程中采用的严格资格测试确保标注者尽其所能给出最佳答案。

通过一家数据标注机构，从中国各高校招募了 80 多名候选标注者 。他们是英语语言学专业的研究生，英语流利（通过了 TEM-8 测试[1]）。阅读说明后，候选者参加了筛选测试，测试要求他们进行标注。

[1]英语专业八级测试，是中国面向英语专业学生的最高级别测试，用于衡量英语总体水平。

50 个示例也由作者进行了标注，并与语言专家交叉核对。通过同一家数据标注机构聘用了具有丰富自然语言处理相关标注经验的语言专家。测试集中至少有 85% 的问题回答正确的 16 名候选人被选为标注者。在这 16 名标注者中，准确率最高的三人被选为质量检查员（QIs）。经过进一步培训后，语言专家讲解了标注指南以及他们在筛选测试中错误标注的案例，标注者开始进行标注。

待标注的事件对平均分成 3 个子集，并分 3 个阶段依次发放给标注者，每个阶段间隔一周。每个阶段中，标注者提交标注后，质量检查员会随机抽取并标注 40% 的案例。专家会抽查质量检查员审查过的案例中的 15–25%，并核查质量检查员的标注结果。如果发现任何错误标注，专家会将这些案例退回给质量检查员，并附上专家的标注及相应说明，帮助质量检查员理解问题，避免在下一阶段犯同样的错误。这种专家监督能够控制质量检查员的工作质量，进而避免质量检查员提供不可靠的标签误导标注者。语言学专家审查质量检查员的标注后，如果某位标注者的答案与相应质量检查员的答案一致率低于 95%，就会要求该标注者重新标注不一致的案例。反复进行这一流程，直至一致率达到 95%。

两位作者还会监控标注的质量。在每个阶段，我们从每位标注员仅标注一次的结果中抽取5%，并对这些案例进行标注，以检查其准确率。如果某位标注员达到的准确率低于85%，则丢弃该标注员的工作，并将其从后续阶段中移除。只有一位标注员未通过这项测试集，并在第一阶段后被移除。不计作者检查的案例，共有4771个案例（占全部案例的40.2%）至少接受了四次标注。

3.2.3 事件表示修订

如图 2中的第二个标注示例所示，一个事件（谓词）可以被包含在另一个事件中，但实际上并不是它的论元。对于标注者标为相互独立的事件对，我们会进一步要求他们修订包含事件（“事件1”），删除其中被包含事件（“事件2”）的片段。不过，被包含事件通常与包含事件共享主语或宾语，因此标注者会在事件1中保留所有共享部分，并删除事件2中其余的片段。图 2展示了一个例子：在第二种情况下，事件2“在这里 {V: 工作} 的那个人”并非事件1的论元或条件，但与事件1共享主语“那个人”。修订后的事件1为“那个人 {V: 告诉} 我滚出去”。结合原始事件表示中的论元标签，修订后的事件表示为 {ARG0：那个人 } {V：告诉 } {论元2：我 } {论元1：滚出去 }。对事件表示进行的这项人工修订与事件依存关系的标注同步进行，并且也由质量检查员和语言专家审查。

对于被标注为条件的事件，被包含事件通常不会与包含事件共享主语或宾语。我们观察到，通常可以自动修订包含事件，因此在这种情况下没有要求标注者这样做。如果条件事件 e_j 及其包含事件 e_i 在句子中以以下子序列之一出现：(1) $e_i\ s\ e_j$ 或(2)$s\ e_j\ e_i$ 或(3)$e_j,\ e_i$，其中 s 是信号词/短语“如果”、“每当”、“只要”、“在 [] 条件下”、“除非”或“假如”，那么我们首先移除 e_i 中位于e_j跨度内的所有词语，其次移除作为e_j前件且与其相邻的任何信号词/短语。在少数情况下，如果在第二步中未检测到信号词或短语，我们会手动检查并修订包含事件。

3.3 分析

表2 展示了标注数据集中按各标签分类的事件对分布，以及训练集、开发集和测试集子集中的事件对数量。

分布存在偏差并不令人意外，其中多数（略高于 75%）事件对被标注为“论元”。我们选取进行标注的事件对中，事件之间的谓词位于包含该事件的跨度内这一关系，是它们构成依存关系的必要条件。在被标注为非论元（条件或独立）的 2901 个事件对中，标注者和我们的自动化方法对 2490 个不同事件的表示进行了改进。

对于人工标注的数据集，始终存在

	论元		非论元		总体
	必选、可选条件		独立		
训练集	4096	2837	335	1861	9129 (77%)
dev	635	421	41	355	1452 (12.3%)
test	594	368	70	239	1271 (10.7%)
总体	5325 (44.9%)	3626 (30.6%)	446 (3.8%)	2455 (20.7%)	11852

表 2：标注的 EDeR 数据集中，事件对在不同标签以及训练集、开发集和测试子集中的分布。

标注实例数量与标注质量之间的权衡（Kryscin-ski et al.，2019年；Cui et al.，2020年）。我们数据集的规模受到标注方法的限制，但与许多其他人工标注的事件关系推理数据集相比，其规模相当或更大（Pustejovskyetal.，2003年a；Glavaš et al.，2014年；Chambers etal.，2014b；Mirza and Tonelli，2016b；Ning et al.，2018年）。我们对最终标注集随机抽取了5%进行核验，结果发现其中至少95%是正确的。这使我们确信，多层迭代式标注流程能够产出高质量的标注结果。

4 实验1：事件依赖关系抽取

我们首先将该任务视为二元分类问题。假设给定一个句子 X，其中包含两个事件 e_i 和 e_j，模型需要判断 e_j 是否为 e_i 的论元事件。我们采用典型的分类评估标准，包括精确率、召回率和准确率。与许多现有研究一样（Gunasekaraand Nejadgholi，2018年；Zhu et al.，2022年；Chen et al.，2022年），我们还报告 ROC_AUC 的结果（Bradley，1997）。该指标用于衡量随机选取的正样本（即论元样本）排名高于随机选取的负样本（即非论元样本）的概率。ROC_AUC 越高，预测性能越好。

4.1 基线模型

我们评估了基于启发式规则的方法，以及多个预训练语言模型，包括判别式模型 BERT（Devlin 等人，2019年）、RoBERTa（Zhuang等人，2021年）、DistilBERT（Sanh 等人，2019年）、XLNet（Yang 等人，2019年）；以及一个自回归生成模型 GPT-2（Rad- ford 等人，2019年）。

基于启发式规则的方法。我们设计了一种无监督方法，用于检测两个事件之间是否存在依存关系。规则如下：只要满足其中任意一条，被包含事件e_j就是包含事件的论元 e_i：(i) 从 e_i 的谓词到e_j 的谓词的句法依存关系是从句补语（*ccomp* 或*xcomp*）或从句主语（*csubj*）。(ii) 从e_j 的谓词到 e_i 的谓词的句法依存关系是系词（*cop*）。(iii) e_j 全部包含在 e_i 的一个论元中，且该论元被标记为 ARGM-PRP（“目的”）或 ARGM-PNC（“目的而非原因”）。句法依存关系通过使用斯坦福 CoreNLP 工具包进行依存句法分析获得（Manning et al.，2014年）。

4.2 输入变体

除了各种基线模型外，我们还探讨了不同类型输入的影响。

标记谓词句子输入受 MarkedBERT（Boualili 等人，2020年）在信息检索研究中取得成功的启发，我们在句子中添加特殊标记，以突出谓词词元。例如，给定句子“The man who works here tells me to get the hell out.”以及两个事件谓词“tells”和“get”，标记谓词输入为“The man who works here [V1]tells [\V1]meto [V2] get [\V2]the hell out.”。

事件-事件跨度输入 我们还创建了两个事件的跨度样式输入。以上述同一示例为例，事件-事件跨度样式输入为“‘The man who works here {V: tells} me to get t he hell out. [SEP]me {V: get} the hell out”。其中，我们添加了一个特殊词元“[SEP]”作为分隔符。

事件-事件-SRL 输入为进一步检查语义角色标注标签对帮助分类事件-事件关系的影响，我们添加了包含事件中被包含事件谓词所在论元的论元标签。我们还添加了一个特殊词元“[语义角色标注]”，以供模型识别。

事件-事件-SRL-DEP 输入 基于事件-事件-SRL 输入，我们还添加了两个事件谓词的句法依存关系信息。指定特殊词元“[DEP]”供模型识别。

4.3 实现设置

为保证比较的公平性，所有模型均在一台配备 64 个中央处理器、126GB 随机存取存储器和一块 24GB 随机存取存储器的 RTX-3090 图形处理器的计算机上训练。我们将优化方法设为 AdamW（Loshchilov 和 Hutter，2017年），其

输入	模型	精确率(%)	召回率(%)	准确率(%)	ROC_AUC(%)
句子+谓词	基于启发式规则	**97.67 (1)**	34.93	50.08	66.16
事件-事件跨度	DistilBERT（Sanh 等人，2019年）	82.75	90.23	78.35	65.73
	BERT（Devlin 等人，2019年）	87.53	85.34	79.69	70.61
	RoBERTa（Zhuang 等人，2021年）	84.71	87.53	78.58	69.09
	XLNet（Yang 等人，2019年）	83.61	89.60	78.82	62.38
	GPT-2（Radford 等人，2019年）	85.91	86.80	79.21	69.16
事件-事件语义角色标注	DistilBERT（Sanh 等人，2019年）	84.75	87.84	78.82	67.45
	BERT（Devlin 等人，2019年）	85.96	87.21	79.53	71.37
	RoBERTa（Zhuang 等人，2021年）	82.89	**91.16 (3)**	80.06	72.71
	XLNet（Yang 等人，2019年）	82.00	**93.76 (1)**	79.69	64.74
	GPT-2（Radford 等人，2019年）	86.60	85.97	79.29	72.20
事件-事件-SRL-依存	DistilBERT（Sanh 等人，2019年）	85.48	87.53	79.29	69.09
	BERT（Devlin 等人，2019年）	85.26	89.60	80.39	72.61
	RoBERTa（Zhuang 等人，2021年）	83.21	**91.16 (3)**	80.37	71.18
	XLNet (Yang 等人, 2019年)	83.11	91.06	79.21	67.37
	GPT-2(Radford 等人,2019年)	85.89	87.94	79.92	71.41
标记谓词句	DistilBERT (Sanh 等人, 2019年)	82.13	**93.14 (2)**	79.45	71.72
	BERT（Devlin 等人，2019年）	**91.30 (2)**	85.14	**82.61 (1)**	**74.69 (3)**
	RoBERTa（Zhuang 等人，2021年）	**89.87 (3)**	85.85	**81.97 (2)**	**75.40 (1)**
	XLNet（Yang 等人，2019年）	88.22	86.38	**80.94 (3)**	**75.17 (2)**
	GPT-2（Radford 等人，2019年）	85.26	89.60	80.39	70.61

表3：Co 基于不同方法和输入组合的测试集性能比较，适用于 t he 事件依存关系 依赖关系抽取任务。每列中表现最佳的前三项结果已突出显示。

在微调每个模型时采用的初始学习率为 1e−5，批次大小为 4，训练轮数为 4。选择在开发集上 F1 分数最高的模型。

4.4 结果与分析

使用各种输入时模型的性能如表3所示。我们观察到，与仅将事件-事件跨度作为输入相比，加入事件谓词的句法依存和（语义）论元标签信息，能够提升大多数模型的性能。然而，将标记谓词的单句作为输入，比这些事件-事件形式的输入更有效。以标记谓词句作为输入时，BERT 和 RoBERTa 模型在准确率（82.61% 和 81.97%）、精确率（91.30% 和 89.87%）以及 ROC_AUC（74.69% 和 75.40%）方面均优于其他模型。与判别模型相比，GPT-2 的表现不尽如人意，说明这种生成模型的推理能力仍有提升空间。具体而言，尽管基于启发式规则的方法的召回率为 34.93%，远低于其他方法，但其精确率最高，达到 97.67%。这表明，我们的事件依存关系与谓词的句法/语义依存关系之间存在潜在相关性。作为一种无监督方法，它有助于准确预测两个未见事件之间是否存在依存关系。The

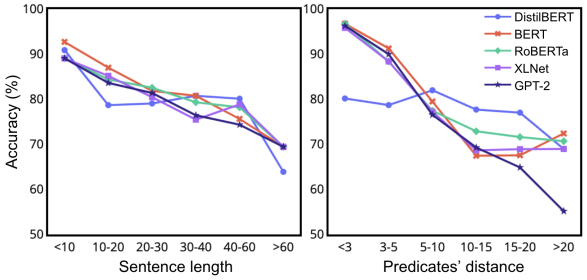


图5：基线模型在不同句子长度范围（左）和不同谓词距离范围（右）内的性能（准确率）。

基线模型的最高准确率达到 82.61%。这表明，我们的 EDeR 数据集足以基于语言模型的预训练与微调机制，训练出能够识别事件依存关系的优秀预测器。

4.4.1 句子长度和谓词距离的影响

图 5（左）展示了以标记谓词句作为输入时，基线模型在不同句子长度下的性能。随着句子长度增加，几乎所有模型的性能也随之下降，这与人类的直觉一致。较长的句子可能包含更复杂的结构，这些结构对模型来说难以捕捉。同样，如图5（右）所示，当

句中两个谓词之间的距离（即它们之间的词数）增加时，模型的准确率也会下降。这也表明，对模型而言，检测相距较远的谓词之间的依存关系颇具挑战性。

5 实验 2：下游任务的收益

在本节中，我们展示了 EDeR 可用于提升下游任务的性能。

5.1 指南 事件表示提取

事件表示抽取被视为语义角色标注（SRL）任务：给定事件的谓词及其所在句子，识别该事件的论元（包括跨度和标签）。尽管最初基于 OntoNotes 的语义角色标签，我们修订后的任务则基于 EDeR 更新后的语义角色标签。

实际上，我们选择在 OntoNotes 上训练的当前最先进的语义角色标注系统 CRFSRL（Zhang 等人，2022年）作为基线。如图6中 CRFSRL 部分所示，该系统以一个标记了事件谓词（p）的句子作为输入，并检测所有与p具有句法依存关系且作为其后代的词。随后，系统基于每个后代与句中其他词之间的句法依存关系，为每个后代构建潜在子树。这些子树就是论元跨度。

在图 6中的 CRFSRL 示例里，其事件表示输出包含一个非论元事件“The man who works here”。因此，我们将事件依存信息整合到 CRFSRL 中，对其进行改进。改进后的系统称为事件依存 CRFSRL（ED-CRFSRL）。如图6中的 ED-CRFSRL 部分所示，根据事件依存关系预测，事件（谓词）“works”不是“tells”的论元，因此会剪除所有以“works”为根节点的子树。由此，我们的 ED-CRFSRL 系统生成了正确的事件表示“The man tells me to get the hell out”。

我们根据 CoNLL-2012 共享任务提供的官方脚本评估语义角色标注系统的性能[2]。我们首先使用 OntoNotes 子集中（275 篇抽样文档）的原始事件表示，考察 CRFSRL 系统的性能。该系统命名为 CRFSRL-O。CRFSRL-O 系统在测试集上的精确率、召回率和 F1 分数分别为 77.07%、81.03% 和 79.00%。在此基础上

[2]https://www.cs.upc.edu/ srlconll/

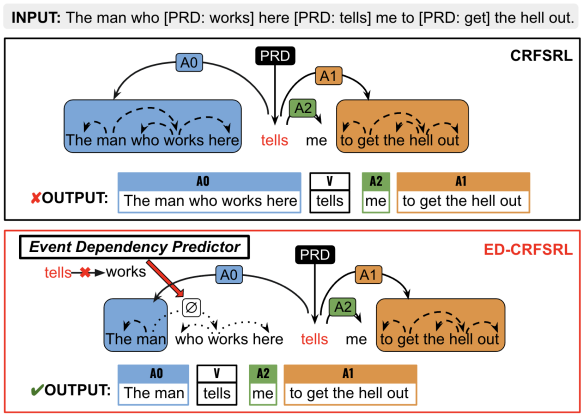


图6：CRFSRL 系统（Zhang 等人，2022年）与我们的 ED-CRFSRL 系统生成的不同事件表示示例。

我们使用 EDeR 中所有更新事件训练 CRFSRL 系统。如表4 第一行所示，CRFSRL 的精确率为 74.19%，召回率为 78.45%，F1 分数为 76.26%，总体低于 CRFSRL-O 的性能。这表明，EDeR 中更新后的事件表示对 CRFSRL 系统而言更难建模。不过，使用基线模型（通常是实验 1 中准确率最高的基线模型）预测的事件依存关系的 ED-CRFSRL（P），以及使用 EDeR 标注的事件依存信息的 ED-CRFSRL（G），其性能均优于 CRFSRL。尤其是 ED-CRFSRL（G），其精确率为 76.09%，召回率为 78.81%，F1 分数为 77.43%。

具体而言，仅针对细化事件时，性能差距更为显著。CRFSRL 在细化事件上的性能低于其在所有更新事件上的表现，这表明它处理细化事件的能力不及处理其他未细化事件。借助基线模型预测的或人工标注的事件依存关系信息，ED-CRFSRL 的性能有了大幅提升（例如，精确率分别提高 13.8% 和 20.77%，F1 分数分别提高 7.44% 和 11.18%），相较于未使用这些信息的 CRFSRL 模型。

所有结果均证明，事件依存信息对于提升语义角色标注系统在修订事件表示抽取任务中的表现至关重要，尤其是在抽取细化事件方面。

	精确率	召回率	F1 分数
所有更新事件			
CRFSRL(Zhang 等人，2022年)	74.19	78.45	76.26
ED-CRFSRL（P）	74.35	78.53	76.38
ED-CRFSRL（G）	**76.09**	**78.81**	**77.43**
仅细化事件			
CRFSRL（Zhang 等人，2022年）	67.34	65.73	66.53
ED-CRFSRL（P）	81.14	67.97	73.97
ED-CRFSRL（G）	**88.11**	**69.51**	**77.71**

表 4：对于修订事件抽取任务，CRFSRL 系统（Zhang 等人，2022年）、使用基线模型预测的事件依存信息的 ED-CRFSRL 系统（ED-CRFSRL（P）），以及使用 EDeR 标注的事件依存信息的 ED-CRFSRL 系统（ED-CRFSRL（G））之间的性能比较。**所有更新事件**指 EDeR 测试集中的所有事件（包括细化事件和非细化事件）的性能，而**仅细化事件**指仅针对细化事件表示的性能。

5.2 改进共指消解

为了进一步探究 EDeR 提供的精炼事件表示是否比 OntoNotes 提供的原始版本更合理，我们将其应用于一项下游任务：共指消解。

给定一段文本，共指消解（CR）的目标是识别所有指向同一实体的提及。我们选用一个当前最先进的共指消解模型 coref-HGAT（Jiang 和 Cohn，2021年），该模型最初基于 OntoNotes 的共指标注进行训练和测试。coref-HGAT 模型在异构图中编码事件语义角色信息和词语句法依存信息，并将这些图融入上下文化嵌入，以识别提及之间的共指链接。如图7 所示，在原始事件表示（来自 OntoNotes）中，图中错误地将事件“当她看到他们时”链接为事件“哭了”的论元。coref-HGAT 模型错误地预测提及“这个女孩”与“她”共指。然而，根据 EDeR 的精炼事件表示，更新后的图中“她”和“这个女孩”之间不存在语义链接。模型正确识别出提及“她”和“这个女孩”实际上并不互相共指。

由于 EDeR 基于 OntoNotes 的一个子集（275 份抽样文档），我们使用该子集中的 CF标注开展这项共指消解任务。评估仍采用官方的 CoNLL-2012 评估脚本。

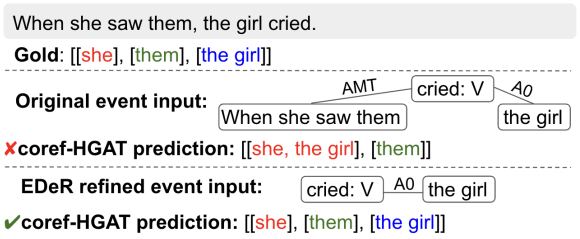


图 7：coref-HGAT 模型分别以 OntoNotes 中的原始事件表示和我们通过 EDeR 改进的事件表示为输入时，得到的不同共指消解结果示例。

输入	精确率	召回率	F1 分数
OntoNotes 事件表示（G）	67.73	65.41	66.50
EDeR 事件表示（G）	**69.14**	**65.69**	**67.40**
OntoNotes 事件表示（P）	64.82	63.79	64.28
EDeR 事件表征（P）	**67.79**	**64.42**	**65.96**

表 5：使用 OntoNotes 中经标注（G）和预测（P）的原始事件表示，以及 EDeR 中更新后的事件表征时，coref-HGAT（Jiang 和 Cohn，2021年）模型共指消解性能的比较。

首先，将 EDeR 标注的更新后事件表示作为 coref-HGAT 模型的输入时，与原始标注的事件表示（来自 OntoNotes 抽取的 275 篇文档）相比，其测试集精确率、召回率和 F1 分数分别提高了 1.41%、0.28% 和 0.9%，如表5前两行所示。这些结果证明，经过 EDeR 修订的事件表征更有利于共指消解任务，进一步证实了我们修订事件表示的合理性和有效性。

此外，我们还使用 CRFSRL-O 预测的原始事件表示，以及第5.1节中的 ED-CRFSRL（P）预测的更新事件表示，考察 coref-HGAT 模型的预测性能。如表5 最后两行所示，使用预测的更新事件表示时，coref-HGAT 模型的精确率、召回率和 F1 分数均高于使用预测的原始事件表示时，分别高出 2.97%、0.63% 和 1.68%。尽管模型预测存在错误，使用预测的更新事件表示时，coref-HGAT 模型的性能仍与使用 OntoNotes 标注的事件表示所得结果相当。这些结果表明，可以通过替换

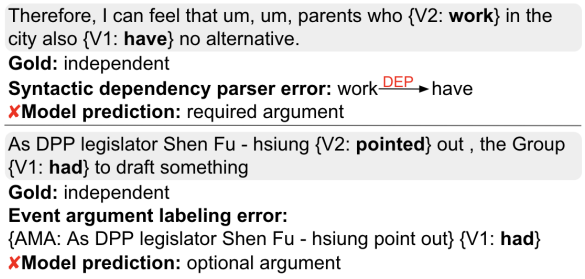


图8：细粒度事件依赖关系抽取案例研究。

其带有标注的原始事件表示输入，以及我们可预测的更新后事件表示。

6 实验3：细粒度事件依存关系抽取

如前所述，对于我们的 EDeR 数据集，论元标签还可以进一步细分为两个细粒度类别：必需论元和可选论元。我们将所介绍的基线模型及输入应用于这项三分类任务。

基线语言模型采用不同类型输入时的预测结果如表6所示。首先，在这项任务中，事件-事件-SRL-依存输入的总体性能优于其他输入。除第 4.4 节讨论的句法依存关系所提供的有用信息外，另一个可能的原因是，在 EDeR 数据集中，被标注为必需论元的事件中，超过 95.79% 的事件谓词位于其包含事件的编号论元（ARG0–ARG5）中。其次，与实验 1 的结果类似，BERT 和 RoBERTa 模型仍然表现更佳。具体而言，RoBERTa 的精确率、召回率、F1 分数和准确率均为最佳，体现出其更强的关系推理能力。第三，与论元/非论元二元分类相比，总体性能有所下降，最高准确率为 70.79%。这表明，对于我们当前的基线模型而言，三分类任务更具挑战性。

6.1 案例研究

我们从基于 Event-Event-SRL-DEP 的 RoBERTa 模型中识别出两种现象，用于细粒度事件依存关系抽取：

6.1.1 句法依存分析错误

一些错误的分类结果可能是由误导性的句法依存分析结果造成的。如图所示，在图 8中，现成的依存句法分析器错误地将句法依存关系赋予了两个事件谓词“work”和“have”，这可能会影响模型的预测。

6.1.2 事件论元标注错误

此外，如图8 中的第二个案例所示，OntoNotes 为事件谓词“had”提供的语义标签有误（例如，“the group”被遗漏，未标注为 ARG0）。这可能会妨碍模型准确预测事件依赖关系类型。

7 结论

我们介绍了 EDeR，这是一个由人工标注的高质量事件依存关系数据集。我们详细介绍了数据集的构建过程，并对其进行了分析。多种具有竞争力的基线模型的实验结果，以及该数据集在下游任务中的后续应用，证明了该数据集的实用价值。此外，基线模型在预测论元属于必选、可选和非论元的三分类任务上的性能，也体现了该数据集的挑战性。这将为我们未来的工作指明方向，我们也希望它能推动信息检索和自然语言处理领域的后续研究。我们将 EDeR 数据集及相关基线模型的代码开源。

参考文献

JunAraki、ZhengzhongLiu、Eduard Hovy 和 Teruko Mitamura。2014年。检测用于事件共指消解的子事件结构。载于第九届语言资源与评测国际会议论文集（*LREC'14*），第4553– 4558页，雷克雅未克，冰岛。欧洲语言资源协会（ELRA）。

Jorge Balazs、Edison Marrese-Taylor、Pablo Loyola 和 Yutaka Matsuo。2017年。通过注意力机制优化用于文本蕴含识别的原始句子表示。载于第二届自然语言处理向量空间表示评估研讨会论文集，第51–55页，丹麦哥本哈根。计算语言学协会。

Steven Bethard。2013年。ClearTK-TimeML：一种用于 TempEval 2013年的最小主义方法。载于 第二届词汇与计算语义联合会议（**SEM*），第*2*卷：第七届国际语义评测研讨会（*SemEval 2013*年）论文集，第10–14页，美国佐治亚州亚特兰大。计算语言学协会。

输入	模型	精确率（%）	召回率（%）	F1 分数（%）	准确率（%）
事件-事件跨度	DistilBERT（Sanh 等人，2019年）	61.05	60.97	60.90	64.80
	BERT（Devlin 等人，2019年）	65.33	65.64	64.92	69.13
	RoBERTa（Zhuang 等人，2021年）	65.63	65.56	65.59	69.06
	XLNet（Yang 等人，2019年）	62.52	62.46	62.47	66.46
	GPT-2（Radford 等人，2019年）	61.70	61.49	61.49	65.43
事件-事件语义角色标注	DistilBERT（Sanh 等人，2019年）	63.74	60.90	61.41	65.98
	BERT（Devlin 等人，2019年）	65.68	65.89	65.77	69.21
	RoBERTa（Zhuang 等人，2021年）	**66.08（3）**	66.44	66.08	**69.69（3）**
	XLNet（Yang 等人，2019年）	65.71	66.04	65.28	69.21
	GPT-2(Radford 等人, 2019年)	63.04	62.70	62.73	66.46
事件-事件-SRL-依存	DistilBERT (Sanh 等人, 2019年)	63.81	62.82	63.15	66.77
	BERT (Devlin 等人, 2019年)	**66.21 (2)**	66.13	**66.12 (3)**	**69.76 (2)**
	RoBERTa（Zhuang 等人，2021年）	**67.21 (1)**	**67.31 (1)**	**67.15 (1)**	**70.79 (1)**
	XLNet（Yang 等人，2019年）	66.01	**66.21 (3)**	66.08	69.37
	GPT-2（Radford 等人，2019年）	62.86	62.99	62.82	66.54
标记谓词句 DistilBERT（Sanh 等人，2019年）		62.18	60.99	61.40	64.96
	BERT（Devlin 等人，2019年）	65.91	**66.67 (2)**	**66.13 (2)**	69.29
	RoBERTa (Zhuang 等人，2021年)	65.83	66.17	65.89	69.53
	XLNet (Yang 等人，2019年)	64.83	65.69	64.78	68.11
	GPT-2(Radford 等人，2019年)	60.70	60.88	60.74	64.33

表6：细粒度事件依赖关系分类任务的性能比较。每列中排名前三的结果已突出显示。精确率、召回率和 F1 分数均为宏平均值。

Claire Bonial、Jena Hwang、Julia Bonn、Kathryn Conger、Olga Babko-Malaya 和 Martha Palmer。2012年。《英语PropBank标注指南》。计算语言与教育研究中心 认知科学研究所 科罗拉多大学博尔德分校，48。

Lila Boualili、Jose G. Moreno 和 Mohand Boughanem。2020年。Markedbert：将传统信息检索线索融入预训练语言模型以进行段落检索。载于 第*43*届国际*ACM*信息检索研究与发展会议论文集，SIGIR '20，页码 1977–1980，纽约州纽约市，美国。计算机协会。

Andrew P Bradley。1997年。机器学习算法评估中ROC曲线下面积的使用。模式识别，30(7):1145–1159。

Xavier Carreras 和 Lluís Màrquez。2005年。CoNLL-2005共享任务简介：语义角色标注。载于第九届计算自然语言学习会议论文集（*CoNLL-2005*），第152–164页，密歇根州安娜堡。计算语言学协会。

Taylor Cassidy、Bill McDowell、Nathanel Chambers 和 Steven Bethard。2014年。密集事件排序的标注框架。技术报告，卡内基梅隆大学，宾夕法尼亚州匹兹堡。

Nathanael Chambers。2013年。Navytime：从原始文本中进行事件与时间排序。技术报告，海军学院，马里兰州安纳波利斯。

NathanaelChambers、Taylor Cassidy、Bill McDowell 和 Steven Bethard。2014年a。采用多遍架构的密集事件排序。《计算语言学协会汇刊》，2：273–284。

NathanaelChambers、TaylorCassidy、BillMcDowell 和 Steven Bethard。2014b。采用多遍架构的密集事件排序。《计算语言学协会汇刊》，2：273–284。

Miao Chen、Ganhui Lan、FangDu 和 VictorLobanov。2020年。在命名实体识别和关系抽取任务中使用预训练Transformer进行联合学习，以支持临床分析。载于第三届临床自然语言处理研讨会论文集，第234–242页，线上。计算语言学协会。

Yanda Chen、Ruiqi Zhong、Sheng Zha、George Karypis 和 He He。2022年。通过语言模型上下文内调优实现元学习。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第719–730页，爱尔兰都柏林。计算语言学协会。

Leyang Cui、Yu Wu、Shujie Liu、Yue Zhang 和 Ming Zhou。2020年。MuTual：一个用于多轮对话推理的数据集。载于计算语言学协会第*58*届年会论文集，第1406–1416页，线上。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向变换器预训练。载于*2019*年计算语言学协会北美分会年会论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Kawin Ethayarajh。2019年。上下文化词语表示的上下文依赖性有多强？比较 BERT、ELMo 和 GPT-2 嵌入的几何结构。载于*2019*年自然语言处理实证方法会议暨第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第55–65页，中国香港。计算语言学协会。

Jannik Fischbach、Julian Frattini、Arjen Spaans、Max- imilian Kummeth、Andreas Vogelsang、Daniel Mendez 和 Michael Unterkalmsteiner。2021年。需求工件中因果关系的自动检测：CIRA 方法。载于需求工程：软件质量基础国际工作会议，第 19–36 页，施普林格国际出版社。施普林格国际出版社。

Goran Glavaš 和 Jan Šnajder。2014年。从新闻报道中构建连贯的事件层级。载于*Text Graphs-9* 论文集：自然语言处理图方法研讨会，第 34–38 页，多哈，卡塔尔。计算语言学协会。

GoranGlavaš、Jan Šnajder、Marie-FrancineMoens 和 Parisa Kordjamshidi。2014年。HiEve：一个用于从新闻报道中提取事件层级的语料库。载于第九届语言资源与评测国际会议论文集（*LREC'14*），第 3678–3683 页，雷克雅未克，冰岛。欧洲语言资源协会（ELRA）。

Isuru Gunasekara 和 Isar Nejadgholi。2018年。在线内容多标签毒性识别的标准文本分类实践综述。载于第二届线上滥用语言研讨会论文集（*ALW2*），第 21–25 页，布鲁塞尔，比利时。计算语言学协会。

Jan Haji˘c、Massimiliano Ciaramita、Richard Johansson、Daisuke Kawahara、Maria Antònia Martí、Lluís Màrquez、Adam Meyers、Joakim Nivre、Sebastian Padó、Jan Št˘epánek、Pavel Stra˘nák、Mihai Surdeanu、Nianwen Xue 和 Yi Zhang。2009年。CoNLL-2009 共享任务：多种语言中的句法和语义依存关系。载于第十三届计算自然语言学习会议（*CoNLL 2009*）论文集：共享任务，第 1–18 页，博尔德，科罗拉多州。计算语言学协会。

Eduard Hovy、Teruko Mitamura、Felisa Verdejo、Jun Araki 和 Andrew Philpot。2013年。事件并不简单：同一性、非同一性和准同一性。载于事件研讨会：定义、检测、共指和表示，第 21–28 页，亚特兰大，佐治亚州。计算语言学协会。

Pere-Lluís Huguet Cabot 和 Roberto Navigli。2021年。REBEL：通过端到端语言生成进行关系抽取。载于《计算语言学协会计算语言学实证研究：*EMNLP 2021*》，第 2370–2381 页，希尔顿黑德岛，多米尼加共和国。计算语言学协会。

Fan Jiang 和 Trevor Cohn。2021年。利用异构图注意力网络，在共指消解中融入句法和语义。载于《计算语言学协会北美分会 *2021* 年会议论文集：人类语言技术》，第 1584–1591 页，线上。计算语言学协会。

Wojciech Kryscinski、Nitish Shirish Keskar、Bryan Mc- Cann、Caiming Xiong 和 Richard Socher。2019年。神经文本摘要：批判性评估。载于*2019*年自然语言处理实证方法会议与第九届自然语言处理国际联合会议论文集（*EMNLP-IJCNLP*），第 540–551 页，中国香港。计算语言学协会。

Natsuda Laokulrat、Makoto Miwa、Yoshimasa Tsu- ruoka 和 Takashi Chikaya ma。2013年。UTTime：使用深层句法特征进行时间关系分类。载于第二届词汇与计算语义联合会议（**SEM*），第 *2* 卷：第七届语义评估国际研讨会（*SemEval 2013*）论文集，第 88–92 页，美国佐治亚州亚特兰大。计算语言学协会。

Md Tahmid Rahman Laskar、Jimmy Xiangji Huang 和 Enamul Hoque。2020年。用于答案选择任务中句子相似度建模的上下文化嵌入式 Transformer 编码器。载于第十二届语言资源与评测会议论文集，第 5505–5514 页，法国马赛。欧洲语言资源协会。

Seonghyeon Lee、Dongha Lee、Seongbo Jang 和 Hwanjo Yu。2022年。通过基于最优传输的对比式句子学习，实现可解释的语义文本相似度。载于计算语言学协会第 *60* 届年会论文集（第 *1* 卷：长篇论文），第 5969–5979 页，爱尔兰都柏林。计算语言学协会。

Beth Levin 等人。1999年。宾语性：事件结构视角。*CLS* 论文集，35(1):223–247。

Ilya Loshchilov 和 Frank Hutter。2017年。解耦权重衰减正则化。

Youmi Ma、Tatsuya Hiraoka 和 Naoaki Okazaki。2022年。基于表格标注、使用卷积神经网络的实体与关系联合抽取。载于第六届自然语言处理结构化预测研讨会论文集，第 11–21 页，爱尔兰都柏林。计算语言学协会。

Christopher Manning、Mihai Surdeanu、John Bauer、Jenny Finkel、Steven Bethard 和 David McClosky。2014 年。Stanford CoreNLP 自然语言处理工具包。载于计算语言学协会第*52*届年会论文集：系统演示，第55–60页，马里兰州巴尔的摩。计算语言学协会。

Dominique Mariko、Hanna Abi-Akl、Kim Trottier 和 Mahmoud El-Haj。2022年。金融因果关系抽取共享任务（FinCausal 2022）。载于 *LREC2022* 第四届金融叙事处理研讨会论文集，第105–107页，法国马赛。欧洲语言资源协会。

Dominique Mariko、Estelle Labidurie、Yagmur Ozturk、Hanna Abi Akl 和 Hugues de Mazancourt。2020年。面向金融因果共享任务的数据处理与标注方案。

Paramita Mirza 和 Sara Tonelli。2016年。CATENA：从自然语言文本中抽取因果与时间关系。载于*COLING 2016* 论文集：第 *26* 届国际计算语言学会议：技术论文，第 64–75 页，日本大阪。COLING 2016 组委会。

Paramita Mirza 和 Sara Tonelli。2016b。CATENA：从自然语言文本中抽取因果与时间关系。载于*COLING 2016* 论文集：第 *26* 届国际计算语言学会议：技术论文，第 64–75 页，日本大阪。COLING 2016 组委会。

Makoto Miwa 和 Yutaka Sasaki。2014年。使用表格表示对实体和关系抽取进行联合建模。载于*2014* 年自然语言处理实证方法会议论文集（*EMNLP*），第1858–1869页，多哈，卡塔尔。计算语言学协会。

Qiang Ning、Hao Wu 和 Dan Roth。2018年。事件时间关系的多轴标注方案。载于第*56*届计算语言学协会年会论文集（第*1*卷：长篇论文），第1318–1328页，澳大利亚墨尔本。计算语言学协会。

Sinno Jialin Pan 和 Qiang Yang。2010年。迁移学习综述。*IEEE* 知识与数据工程汇刊，22(10):1345–1359。

Yannis Papanikolaou、Ian Roberts 和 Andrea Pier- leoni。2019年。用于无监督关系抽取的深度双向 Transformer。载于第二届低资源自然语言处理深度学习方法研讨会论文集（*DeepLo 2019*），第67–75页，中国香港。计算语言学协会。

Sameer Pradhan、Alessandro Moschitti、Nianwen Xue、Hwee Tou Ng、Anders Björkelund、Olga Uryupina、Yuchen Zhang 和 Zhi Zhong。2013年。基于 OntoNotes 实现稳健的语言分析。载于第十七届计算自然语言学习会议论文集，第 143–152 页，保加利亚索非亚。计算语言学协会。

SameerPradhan、Alessandro Moschitti、Nianwen Xue、Olga Uryupina 和 Yuchen Zhang。2012年。CoNLL-2012年共享任务：对 OntoNotes 中多语言、无限制的共指进行建模。载于*EMNLP* 与 *CoNLL* 联合会议——共享任务，第 1–40 页，韩国济州岛。计算语言学协会。

James Pustejovsky、Patrick Hanks、Roser Sauri、Andrew See、Robert Gaizauskas、Andrea Setzer、Dragomir Radev、Beth Sundheim、David Day、Lisa Ferro 等。2003年a。TimeBank语料库。载于语料库语言学论文集，第 2003 卷，第 40 页，英国兰开斯特。语料库语言学。

James Pustejovsky、Patrick Hanks、Roser Saurí、Andrew See、Rob Gaizauskas、Andrea Setzer、Dragomir Radev、Beth Sundheim、David Day、Lisa Ferro 和 Marcia Lazo。2003年。TimeBank 语料库。语料库语言学会议论文集，2003:40。

AlecRadford、JeffreyWu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019年。语言模型是无监督的多任务学习器。*OpenAI* 博客，1(8):9。

Malka Rappaport Hovav、Edit Doron 和 Ivy Sichel。2010年。词汇语义学、句法与事件结构。牛津大学出版社，英国。

Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。2019年。DistilBERT，BERT 的蒸馏版本：更小、更快、更便宜且更轻量。

Alessandro Seganti、Klaudia Firlag、Helena Skowron- ska、Michal Satlawa 和 Piotr Andruszkiewicz。2021年。多语言实体与关系抽取数据集和模型。载于 计算语言学协会欧洲分会第*16*届会议论文集：主要卷，*EACL 2021*，线上，*2021*年*4*月*19*日至*23*日，线上。计算语言学协会。

Fiona AntingTan, Ali Hürriyeto˘glu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza 和 Tiancheng Hu。2022年。因果新闻语料库：标注新闻事件句中的因果关系。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 IlliaPolosukhin。2017年。注意力就是你所需要的一切。

Linyi Yang、Zhen Wang、Yuxiang Wu、Jie Yang 和 YueZhang。2022年。迈向细粒度因果推理与问答。

Zhilin Yang、Zihang Dai、Yiming Yang、Jaime Carbonell、Ruslan Salakhutdinov 和 Quoc V.Le。2019年。XLNet：面向语言理解的广义自回归预训练。

Yu Zhang、Qingrong Xia、Shilin Zhou、Yong Jiang、Guohong Fu 和 Min Zhang。2022年。将语义角色标注视为依存句法分析：探索论元内部的潜在树结构。载于 论文集

第*29*届国际计算语言学会议，第4212–4227页，韩国庆州。国际计算语言学委员会。

Jieming Zhu、Quanyu Dai、Liangcai Su、Rong Ma、Jinyang Liu、Guohao Cai、Xi Xiao 和 Rui Zhang。2022年。Bars：面向推荐系统的开放基准测试。

Liu Zhuang、Lin Wayne、Shi Ya 和 Zhao Jun。2021年。一种通过后训练实现稳健优化的 BERT 预训练方法。载于第*20*届中国计算语言学大会论文集，第1218–1227页，中国呼和浩特。中国中文信息学会。