arXiv:2305.03960v2 [cs.CL] 7 Aug 2023

超越基于规则的命名实体识别与关系抽取：从自然语言文本生成流程模型

Julian Neuberger、Lars Ackermann[0000−0002−6785−8998]和 Stefan Jablonski

德国拜罗伊特大学{
firstname.surname}@uni-bayreuth.de

摘要 流程感知信息系统能为企业带来诸多优势，助力日常业务活动的规划、运营和优化。然而，设计正式业务流程模型这一步骤既耗时又不可或缺，往往会阻碍这些系统发挥潜力。为应对这一挑战，从自然语言文本自动生成业务流程模型已成为一种有望加快这一环节的方案。通常需要解决两个关键子任务：从自然语言中提取与流程相关的信息，以及创建实际模型。针对第一个子任务，已有基于规则的方法，这些方法针对特定领域进行了高度优化，但难以适配相关应用。为解决这一问题，我们对现有处理流程进行了扩展，使其完全由数据驱动。我们展示了改进后的处理流程具有竞争力：它不仅消除了特征工程和规则定义所带来的大量额外开销，还能适配不同的数据集、实体和关系类型以及新的领域。此外，目前规模最大的用于第一个子任务的数据集 PET 数据集，并不包含流程描述中实体提及之间语言指代的信息。然而，将这些提及解析为单一可视化元素，对于生成高质量流程模型至关重要。我们提出了对 PET 数据集的扩展，其中纳入了语言指代信息，并提出了相应的解析方法。最后，我们对该数据集固有的挑战进行了详细分析。

关键词： 流程感知信息系统、流程抽取、命名实体识别、关系抽取、共指消解

1 引言

从自然语言流程描述中自动生成正式的业务流程模型正日益受到欢迎 [1,2,7,10,19]。例如，这一趋势的动因是手动设计此类流程模型所需的时间相对较长。在流程管理项目中，流程模型设计耗费的时间最多可占总时长的 60% [10]。相关技术

从自然语言文本自动生成流程模型旨在减少这类工作量，但为此需要解决若干子任务，这些任务可分为两个不同的阶段：（*i*）信息抽取阶段和（*ii*）流程模型生成阶段。在信息抽取阶段，相关技术会识别流程元素（例如活动、参与者、数据对象），抽取关系（例如活动之间的顺序流关系），并消解指代（例如对同一数据对象的不同提及）。流程模型生成阶段则基于这些信息创建具体的流程模型 [10, 13, 19]。信息抽取阶段最先进水平的技术存在两个核心问题，我们将在下文简要讨论。

核心问题 1 现有方法大多基于规则，即使用以领域知识为基础、人工制定的规则 [2, 10, 21]。对于专门构建这些规则的数据集，基于规则的系统通常能达到显著的精确率和召回率。然而，这些方法*a)* 需要投入大量人力来捕捉语言细微差异，*b)* 需要深厚的技术知识以及对目标领域的了解，并且*c)* 难以适应底层数据的细微变化，导致所需规则的数量不可接受地增加 [26]。采用机器学习可以克服这些缺点，尤其是深度学习方法已被证明能大幅减少所需的工作量和领域知识 [15]。然而，深度学习方法通常需要大量数据才能稳定训练 [14], ，而业务流程建模研究领域目前无法提供此类数据 [12]。采用表达能力较弱的机器学习模型，可以在这一困境中取得折中，因为它们只需少几个数量级的数据便能稳定训练。

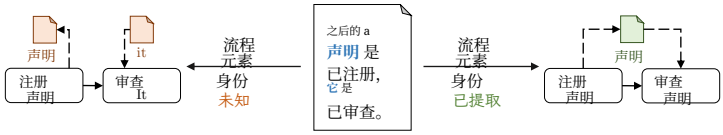


图 1：有无解析流程元素身份时信息抽取阶段的差异示例。根据提及（右图）解析流程元素身份可以生成正确的数据流程；若不解析（左图），数据流程则会断开。

核心问题 2 现有方法的适用范围过于狭窄 [18]。这包括无法捕获足够信息以生成完整流程模型的系统，以及对输入文本结构提出不切实际假设的系统。尤其值得注意的是，目前用于信息抽取阶段的最大数据集（PET [7]）未包含有关流程元素提及之间语言指代的信息。要构建高质量的流程模型，消解同一流程元素各次提及之间的指代至关重要。例如，请看 1 中所示的例子。对于

对于人类读者而言，很明显，“一项声明”和“它”指的是同一个声明实例。为了自动提取能够编码这一知识的流程模型，系统需要将两个提及“一项声明”和“它”消解为同一个实体。如果没有这一步，提取出的流程模型至少会出现两个问题：*(i)* 会为声明 创建两个不同的数据对象，因此，模型无法正确表达登记活动和审查活动处理的是同一份数据；以及*(ii)* 其中一个创建的数据对象会被标记为它，因为系统无法识别它 是对一项声明的指代。尽管声明 这一示例仅关注数据视角，但对于组织层面的流程元素（例如参与者）而言，实体消解同样不可或缺。在这种情况下，需要能够创建包含单一施事者类型的流程模型，以对应声明官员的两处提及，例如，在业务流程模型和符号（BPMN）中将其表示为同一条泳道。总之，可以说，实体消解是正确表达数据和参与者之间关系的首要前提。基于这两个核心问题，我们提出了三个主要研究问题。

RQ1 对于 PET 等规模相对较小的数据集，机器学习模型在精确率和召回率方面能否与基于规则的方法相媲美？**RQ2** 预训练的共指消解方法能否优于朴素的词语匹配，因此可用作解决流程元素提及之间语言指代问题的基线？**RQ3** 对于相同的数据集，深度学习方法能否以与基于规则的方法相当的精确率和召回率提取流程信息，并超越表达能力较弱的机器学习模型？

我们的工作提出了一种改进的处理流程，用以解决这两个问题，具体内容将在第 4 节详细介绍。我们提出了一种基于既有机器学习方法的关系抽取方案。此外，我们扩展了 PET，加入有关流程元素提及身份的信息，并提供了一种基线方法，用于根据流程元素提及解析流程元素身份。我们将该处理流程与 PET 上当前最先进的信息抽取方法进行比较，结果表明，在六种关系类型中，有五种的表现优于该方法， F_1 分数绝对提升了 6%。

本文其余部分结构如下：第2 节将形式化定义流程模型抽取任务。第3 节将讨论与本文相关工作的差异。我们在第6 节对 PET 数据集和抽取方法进行全面研究，该研究基于第5节介绍的严谨实验设置。第7节简要总结各研究问题的答案。我们的实验源代码和扩展数据集均已公开[1]，从而为后续有针对性的研究奠定基础。

[1] 如获接收，文章将在 GitHub 上发布；在此之前 可在此处获取

2 任务描述

自然语言处理（*NLP*）是一门旨在利用自然语言输入数据的学科，涵盖众多子领域。其中一个子领域是从人类可读文本中进行信息抽取。下文将流程元素的抽取及其间关系的抽取描述为信息抽取的三个子问题： 命名实体识别（NER）、关系抽取（RE）和实体消解（ER）。随后，我们将结合第5.1节所述的扩展 PET 数据集，详细介绍这三个子问题。每项任务均假设输入文本已完成预处理，即已进行分词。

命名实体识别是指从一组实体 [15] 中提取词元片段，并且每个片段恰好对应其中一个实体的任务。传统上，命名实体识别只考虑专有名词的提取，但如今命名实体的定义取决于具体领域 [23]。在流程领域，命名实体是流程相关事实，例如参与者（如 首席执行官与Max）或活动（如批准与批准行为）。PET 数据集定义了七种流程相关事实，旨在为从自然语言文本生成流程模型这一任务提供通用模式 [6]。形式上，命名实体识别任务是从给定的词元列表 T 中提取一组三元组 M，使得对于每个三元组$m = (i_s, i_e, t_e) \in M$，索引 i_s 和 i_e 分别表示该片段在 T 中的起始词元和结束词元，而 t_e 指实体类型。在本文中，我们将三元组 m称为实体的提及。如果提取出的提及对应的三元组与数据集中给出的真实标注三元组列表完全匹配，则认为该提及正确。

实体消解 从给定的提及集合中提取唯一实体集合 M。这一步可以视为对同一流程元素的各个提及之间的指代进行消解。正如图 1所示，这是在后续生成有用业务流程模型时至关重要的信息。形式上，实体消解任务定义为寻找一组非空的 提及簇 E，使得每个提及 $m \in M$ 都恰好分配给一个簇 $e \in E$。这些簇称为 实体。为区分命名实体识别中实体一词的用法与实体消解中的用法，以下将命名实体识别的结果称为 提及，将实体消解的结果称为 实体。当且仅当预测实体所包含的提及集合与数据集定义的真实标注完全一致时，才认为该预测正确。实体消解本身是以下任务的超集： 回指消解，即向后指代的代词；共指消解；以及 后指消解，即向前指代的代词 [24]。尽管这些子领域之间存在细微差异和重叠，本文聚焦于共指消解。虽然加入后指消解和回指消解可能有所帮助，但对于我们计划采用的基线而言，这并不足以证明其带来的额外复杂性是合理的，因此不在本文讨论范围内。因此，在后续章节中提到实体消解任务时，我们均指共指消解。PET 数据集仅包含两种与实体身份相关的实体类型：参与者，指自然人、部门、组织或人工代理；以及 活动数据，即某个 活动 [6] 所使用的对象或数据。更多详情请参见第 5.1 节。

关系抽取是一项识别实体对之间一组语义关系R 的任务。当前文献区分了全局关系抽取和提及级关系抽取 [16]。全局关系抽取的任务是从文本中提取构成某种关系的实体对列表，且不需要任何额外信息。另一方面，提及级关系抽取方法接收一对实体提及及其所在句子，并需要预测二者之间的关系。PET 数据集包含提及级的关系信息，使我们的方法能够在局部层面进行学习。PET 数据集定义了六种关系类型，例如流程，它表示行为元素之间的执行顺序 [6]。每种关系都由一个三元组$r = (m_h, m_t, t_r)$正式定义，其中 m_h 是头实体提及或关系的来源，m_t 是尾实体提及或目标，t_r 是语义关系的类型。根据这一定义，关系具有方向性，即对于 $m_h \neq m_t$，$(m_h, m_t, t_r) \neq (m_t, m_h, t_r)$。当且仅当预测的关系元组 $r \in R$ 所对应的三元组与数据集提供的真实标注三元组列表中的某个三元组完全匹配时，该预测才被视为正确。

3 相关工作

本文建立在 [7] 所介绍的工作之上，因此与其密切相关，因为我们使用、扩展并分析了相关数据。此外，我们采用了其命名实体识别方法，并将我们提出的关系抽取方法与其处理流程进行比较。他们缺少对数据更深入的描述，尤其是对影响预测性能的重要数据特征的描述，包括但不限于：关系类型与其论元类型之间的相关性，以及数据的语言变异性。此外，他们的处理流程实现并未公开，阻碍了后续研究与开发。

本研究提出并分析的基线方法与若干相关方法有所关联。 [21], 提出了一种标注方法，该方法基于文本流程描述的依存树表示进行基于规则的模式匹配，随后用于生成事件日志。借助成熟的流程挖掘技术，可以据此抽取形式化流程模型。尽管该方法取得了当前最先进的结果，但它采用的标注方案与 PET 所用方案不同，因此无法直接比较。 [10]提出了一种能够抽取业务流程模型和符号（BPMN）形式化流程模型的处理流程，因此受限于这种流程表示语言。 [2],提出的方法也有相同的局限性，该方法利用 Declare 语言抽取流程模型。PET 采用不同的标注方案，因此无法直接比较。 [18] 提出了一种用于实体分类和关系抽取的神经网络方法，但假设相关文本片段已预先抽取出来。这项任务要容易得多，因为从自然语言中区分相关流程信息与冗余、多余和偶然出现的信息，本身就是一项艰巨任务。 [3]提出了一种高效的深度学习方法，使用形式化语义表示

作为中间特征。由于他们只解决命名实体识别问题，因此我们无法将其方法与我们的方案进行比较。

由于流程抽取与命名实体识别、实体关系抽取和关系抽取这一自然语言处理组合任务之间存在密切联系，因此有多种方法可能能够解决流程抽取任务 [8, 9, 11, 22]。[4] 相关研究探讨了多种针对小型文档中的命名实体识别、实体关系抽取和关系抽取联合任务而构建的方法。将这些方法应用于业务流程管理领域，需要对 PET 的大型文档进行适当拆分，并处理长距离关系，而这超出了本文的研究范围。然而，我们选择了 Jerex [9], 因为 [8]并且 [11] 仅将提及预测为其文本表征（表层形式），这意味着包含这些提及的文本跨度可能存在歧义，因此无法确定词元索引。这不符合我们对提及的定义（第2节），也妨碍了对预测结果的评估。

4 流程信息抽取方法

下面我们简要概述 *NER*、*ER* 和 *RE* 三个处理流程步骤的实现。我们提出的整个处理流程如图 2 所示。下文将把这一处理流程的实现称为*Ours* 。由于预处理步骤和业务流程模型的实际合成都不在本文讨论范围内，因此不作详细介绍。

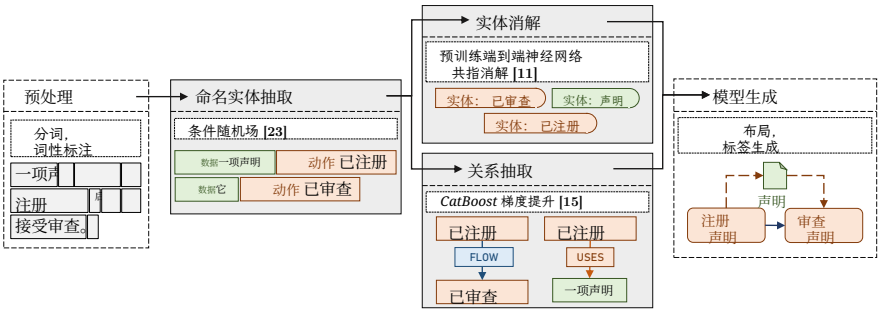


图 2：我们提出的扩展抽取流水线概览。

NER 步骤 与 [7] 中的实现完全相同。该方法基于条件随机场（*CRF*），这是一种用于对观测序列进行标注的强大技术，此处的观测序列为文本中的词元 [25]。给定以这种方式标注的词元序列后，我们随后解析提及，其中每个提及都包含一组词元索引以及预测得到的流程元素类型。

我们实现了两个用于**实体关系抽取步骤**的模块，即朴素实体识别方法，以及一种基于预训练端到端神经共指消解的方法，如 [13]所述，并在 spaCy[2]中实现。朴素实体识别方法，它】【。

[2] 详见 https://explosion.ai/blog/coref 。

我们将其称为朴素实体解析，它会迭代地选择具有相同命名实体识别标签的最佳匹配提及。两个提及之间的匹配程度根据重叠百分比计算，即共享词元数占词元总数的比例。根据该分数对提及对进行排序后，朴素实体解析方法将提及合并为簇。如果选定的某个提及已经属于一个簇，则另一个提及也会被添加到该簇中。如果选定的两个提及都属于某个簇，则将这两个簇合并。重复这一过程，直到剩余的匹配对的重叠程度均低于某个阈值 o。我们通过优化来选择最优的重叠阈值，并选择了 $o = 0.5$ 预训练端到端神经共指消解模块，我们从现在起将其称为神经实体解析，用于预测文本中的共指片段，即相互指代的文本片段。该模块在不使用任何领域知识的情况下完成这一任务，即不依赖于前序步骤中提取的流程元素提及相关知识。随后，我们将这些预测结果与提及进行对齐。如果满足以下条件之一，我们便舍弃相应的预测结果：(1) 对应的文本片段根本不是一个提及； 对应的文本片段与某个提及的文本重叠程度未达到特定百分比 α_m； 与预测文本片段对应的提及未被标记为该实体其他提及中的多数标签；或者 此前接受的预测文本片段所占比例未达到特定比例 α_c 。我们采用网格搜索方法优化这些参数，选择 $\alpha_c = 0.5$ 和 $\alpha_m = 0.5$。图3展示了这一过程的一个简单示例。

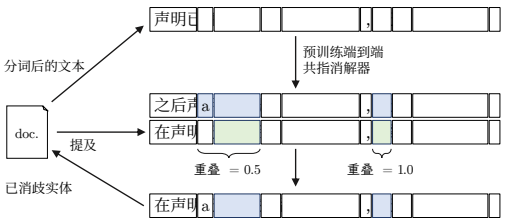


图 3：基于预训练端到端神经共指消解器的实体关系抽取方法示例。预测的共指文本片段一项声明和它被接受，并解析为包含提及声明和它的实体，因为这两个文本片段与相应提及文本的重叠度均至少为 50%。

最后，**关系提取步骤**使用 CatBoost（这是一种用于对数值数据和类别数据进行分类的梯度提升技术） [17]，提取提及之间的关系。在下文中，我们将此模块简称为$BoostRelEx$ 。对于关系中每种头部提及与尾部提及的组合，我们都会构建一组特征，其中包含它们的标签、它们之间相隔的词元数和句子数，以及作为上下文的若干c 个相邻提及标签。随后将这组特征输入模型，由模型预测其所属类别。类别包括关系标签集合，以及一个额外的无关系标签，以便模型能够预测两者之间不存在关系。

两个提及之间的关系。在训练过程中，每次迭代中，我们都会将每个包含关系的提及组合恰好输入模型一次，同时输入一定数量的负例。这些负例仅由提及组合构成，其中对应的实体之间不存在关系。这一概念称为负采样，非常重要，因为不存在关系的提及组合（44,708）远多于存在关系的组合（1,916）。如果不进行负采样，我们的关系抽取模块的精确率会极低，如图 4所示。对于每个正样本，我们会 r_n 随机抽取的负例。增加 r_n 会提升模型预测给定提及对之间是否存在关系的准确性，这一指标称为精确率 P。由于模型学会了必须拒绝某些提及组合，它也不可避免地会拒绝一些正确的组合。因此，模型会漏掉更多实际上存在关系的提及组合，从而导致召回率 R降低。两项得分的调和平均数 R 和 P 能够很好地反映模型的性能。我们将在第5.3节中更详细地讨论这一指标。我们对$BoostRelEx$ 模块进行 $i = 1000$ 次迭代训练。这是整个处理流程中计算量最大的步骤，在 Intel i9-9900K 处理器 @ 3.60GHz 上耗时约 25 分钟，负采样率为 $r_n = 40$ ，上下文大小为 $c = 2$。采样率 $r_n \geq 40$ 能够显著提升结果质量。

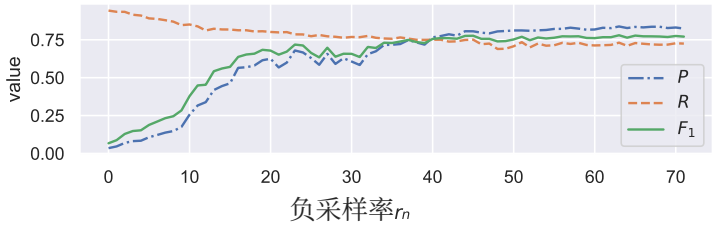


图 4：不同负采样率 r_n 下指标 P、 R 和 F_1 的值。

5 实验设置

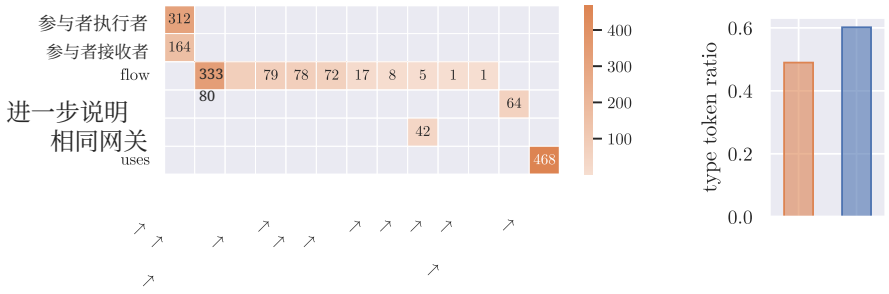
以下内容将介绍对原始 PET 数据集的扩展，并附上数据集统计信息（第 5.1 节）。为实现实证评估，第5.3 节将介绍最适用于该任务和具体数据集的性能度量指标。

5.1 数据集

PET 数据集在 [7], 中进行了详细介绍，以下我们将仅讨论与我们的扩展和分析直接相关的数据集方面。

PET 共包含 45 份文档，包含七种实体类型和六种关系类型。为便于完成第2节所述的实体消解任务，我们将每个流程元素的提及分配到一个簇中[3]。这样共得到 163 个包含两个或更多提及的簇，其中有 75 个活动数据提及簇，以及 88 个施事者提及簇。所有其他实体类型的提及，以及其余的活动数据和施事者提及，都属于仅包含单个提及的簇。

我们将实体内距离定义为实体中每个提及到其他提及的最小距离的最大值。这一指标给出了抽取方法为将两个提及识别为同一实体的一部分而必须推理的最大跨度。对所有实体取平均后，活动数据元素的该指标为31.93个词元，施事者元素的该指标为54.84个词元。对于施事者而言，指称提及之间的距离明显更长，这表明它们可能更难抽取。如图6 的c)和d)所示，我们的实验似乎支持这一观点，但要得出确切的解释可能还需要进一步分析。



(a) (b)

图5：5a 展示了按论元类型汇总的关系数量，论元类型以头部 → 尾部表示。图中仅展示至少存在一个关系的组合。5b 展示了包含至少两个提及的提及簇的平均类符形符比。

直观而言，当对某个实体的提及文本非常相似时，解析这些提及之间的指代会更容易。例如，考虑两个实体，它们各自都由两处提及组成。一个实体的提及为“一项声明”和“该声明”，另一个实体的提及为“声明人”和“一名申请人”。解析第一个实体应该容易得多，因为它的提及包含共同的文本。因此，计算给定类型实体的词汇多样性，可以让我们

[3] 所有簇均由两名专家定义；如果他们的初始标注存在差异，则由第三名专家协助处理。

预测在无错误的情况下提取它们有多困难。类符形符比（TTR）可用于衡量给定输入文本的词汇多样性 [20]。其计算方法是唯一词元数与词元总数之比。比率较高意味着短语非常多样，而比率较低则表示文本非常统一。我们选取所有至少包含两个提及的实体，并计算每个实体的 TTR。例如，由以下三个提及组成的实体：“一项声明”、“该声明”和“它”，其 TTR 因此为$TTR = \frac{4}{5} = 0.8$。然后，我们按实体类型分别计算这些 TTR 值的平均值。平均而言，活动数据提及簇的类符形符比高于施事者，如图5b所示。这一结果使我们推测施事者应该更容易解析。我们的实验支持这一观点，如图6 c）。

图 5a 展示了关系类型根据其论元类型的分布情况。对于类型为 参与者执行者、参与者接收者、相同网关和 流的关系，了解其论元类型并不能作为区分特征。对于这些情况，数据驱动方法（例如本文提出的方法）非常有用，因为它能根据数据自动推断复杂规则，从而节省大量人工工作。相较之下，对于某些关系类型，其类型可直接根据论元类型推断，例如所有头部论元为 活动 、尾部为 活动数据 元素的关系，类型都是 使用。考虑到领域知识，这并不令人意外，因为在 *PET* 中，活动数据 仅与活动搭配使用。因此，预测这些类型的关系，关键更多在于检测它们（召回率），而非正确地将其分类（精确率）。

5.2 对比方法

我们将提出的处理流程与 [7], 中介绍并扩展了 ER 模块的基线进行比较。该处理流程与图 2中展示的我们的处理流程非常相似，但它使用基于规则的关系抽取器，而非$BoostRelEx$ 模块；我们将该抽取器称为 $RuleRelEx$。这些规则定义于 [7], ，但据我们所知，目前没有公开实现；我们的代码是首个向社区提供的可执行版本。共有六条规则，按顺序应用于文档。这意味着规则 1 的优先级高于规则 3 等规则；规则 3 依赖于这一点，因为它需要此前抽取的 流关系信息。下文将该处理流程称为 $Bellan\ +$ 实体关系抽取 。

要回答 **RQ3** ，需要一种能够抽取提及、实体和关系的深度学习方法。$Jerex$ [9] 适用于这项任务，因为它是一种联合训练的端到端深度学习方法，并有望减少错误传播的影响。Jerex 以未经分词的原始文本作为输入，对其进行分词，并预测提及、实体及其之间的关系。对于 $DocRed$ 数据集 [27], 而言，它达到了最先进水平。该数据集是一个大型基准数据集，用于从文档中抽取提及、实体和关系——其任务描述与我们在第 2 节中给出的非常相似。此外，Jerex 能够抽取提及在输入文本中的确切位置，

而竞争方法只能抽取提及的文本[4]。虽然在只关注流程元素文本的应用中，这一缺陷可能不那么重要，但对于业务流程生成任务而言，也就是为业务流程模型和符号流程模型中的活动生成人类可读且内容丰富的标签，则需要预测的 活动 [10] 周围的文本。

5.3 评估

对于现有 [7]基线方法以及我们所做工作的性能评估，我们采用了 中的评估策略。这意味着我们对整个处理流程进行五折交叉验证，并对各个模块的得分取平均值。模块在预测过程中产生的错误会沿处理流程向下游传播，甚至可能加重，因为下游模块也会因此产生错误。为了单独评估某个模块的性能，我们将真实标注数据而非预测结果作为输入。由此共形成五种不同情景，相关结果将在第 6 节中详细讨论。这些情景分别为：**(S1)** 使用 模块的预测结果进行提及项抽取 的实体消解，以及 **(S2)** 使用真实标注提及项进行实体消解。此外，关系抽取 **(S3)** 使用处理流程迄今预测出的实体，**(S4)** 使用基于真实标注提及项进行实体消解 时预测出的实体，最后是 **(S5)** 使用真实标注实体。

在每种情况下，我们都使用 F_1 分数作为度量，因为它反映了任务的目标：尽可能多地找到预期的提及、实体和关系（召回率R），同时不牺牲类型或存在性预测的精确率 P。随后将 F_1计算为 P与 R的调和平均数，即 $F_1 = \frac{2 \cdot P \cdot R}{P+R}$。由于每项预测任务中都包含多个类别，因此必须汇总 F_1、 P和 R 。在第 6 节中，我们始终采用微平均策略，该策略计算 P和 R 时不考虑预测所属的类别。这种策略更偏重样本数量多的类别，因为这些类别的高分可能会掩盖样本数量少的类别中的低分。如果对此有所顾虑，可以采用宏平均策略，分别计算每个类别的 P、 R和 F_1 ，然后再取平均。我们认为，最有用的做法是不论类型如何，都尽可能多地找出流程元素；换言之，找出 90% 的所有活动和仅 10% 的所有与网关，比找出 50% 的所有元素更好，因为在 PET [7]中有 501 个活动，而与网关只有 8 个。因此，微 F_1 分数更适合这项任务。

根据第2节中的任务描述，我们采用以下匹配策略。若一个提及包含的词元与对应的真实标签完全相同，且标签也相同，则将其计为正确预测。若一个实体包含的提及与真实标签完全相同，则将其计为正确预测。最后，若一个关系的两个论元及其标签均与真实标签匹配，则将其计为正确预测。因此，例如，提及“*the*”中缺少一个“*the claim*”，就会导致该提及预测错误，以及所有引用该提及的实体和关系也都预测错误。

[4] 例如，请参见相关讨论 https://github.com/Babelscape/rebel/issues/57

对此产生的影响称为错误传播，这也是我们选择若干场景来单独评估模块，或使用一定程度的真实值输入进行评估的原因，例如(**S4**)。用户或许能够接受精确度稍低的预测，尤其是在只遗漏了限定词等无关紧要的词元时。调查用户如何评价不同精确率水平的重要性不属于本文的讨论范围，将作为未来工作的一部分。

6 结果

以下部分报告了上一节中定义的实验和场景的结果5。基于这些结果，本节回答第1节中提出的研究问题。在第6.1 节中，我们报告实体关系抽取步骤的结果，并比较朴素方法与基于预训练端到端神经共指消解的方法；比较对象既包括单独运行的模块（场景(**1**)），也包括基于命名实体识别模块预测结果的情况（场景(**2**)）。第6.2节展示了在端到端处理流程设置（场景(**3**)）和单独运行（场景(**5**)）时，关系抽取步骤实验的结果。最后，我们在6.3节讨论影响关系抽取结果质量的若干因素，例如错误传播的影响（场景(**4**) 至(**6**)）。

6.1 实体消解性能

我们计算 F_1 所有至少包含两条提及的提及簇的分数，因为解决只含一条提及的簇是轻而易举的。图6 d）展示了两种方法之间的差异。总体而言，朴素版本达到 $F_1 = 0.26$，而我们提出的预训练方法表现显著优于它，达到$F_1 = 0.52$。这一巨大差异源于我们采用了精确匹配：簇中只要有一条提及缺失或多余，就会导致整个预测错误。按照设计，朴素方法无法解决回指和前指，即代词的向后指代和向前指代。这意味着，每个至少包含一个回指或前指的实体都会被错误预测。与关系提取步骤类似，使用命名实体识别步骤的结果会大幅降低性能。根据实验结果，我们得出结论：朴素实体识别方法并不可行，而采用神经方法则能显著提升性能。如果对预训练模型进行微调能否提高准确率，将会是一个值得研究的问题。此外，我们目前使用的是任务无关模型；也可以将命名实体识别步骤中提取的提及信息整合到实体关系抽取中，而不是像目前这样使用任务无关模型。这些考量目前不在本文的讨论范围之内，因为本文针对实体关系抽取的研究旨在弥合当前机器学习领域最先进水平与下游方法需求之间的差距；前者聚焦于从自然语言文本中提取业务流程模型的数据（PET）。本节的讨论引出我们对**RQ2**的回答：我们提出的预训练共指消解方法显著优于朴素文本匹配，并且在从自然语言文本生成业务流程模型的场景下，是一种有用的实体提及消解基线。

6.2 关系抽取性能

我们提出的*BoostRelEx* 步骤在实验中明显优于*RuleRelEx*，分别高出 [7] 、$F_1 = 0.10$、$P = 0.04$和 $R = 0.16$ 。图6对此进行了可视化展示，而表1b 列出了具体数值。*BoostRelEx* 显著受益于命名实体识别步骤中的正确预测；当我们提出的处理流程端到端运行时，性能大幅下降，这一点足以证明这一点；Bellan 的 + 实体关系抽取也呈现出类似情况。尽管我们的处理流程在实验中仍优于*Bellan* + 实体关系抽取，但优势大幅缩小，差距分别为 $F_1 = 0.01$、$R = 0.02$，且召回率相同。性能大幅下降的原因之一是我们采用了精确匹配策略。提及缺失、多余或分类错误，都会在关系提取步骤中导致错误，因为只有所有涉及的提及都正确时，关系才会被视为正确（参见第5.3节）。

鉴于错误传播对*BoostRelEx*的影响显著，采用联合训练的端到端模型似乎顺理成章。在第5.2 节中，我们介绍了*Jerex*作为一个有潜力的候选方案。然而，根据我们的实验结果，*Jerex* 无法与我们的处理流程竞争，表现也显著更差，差距分别为 $F_1 = 0.11$、$P = 0.14$和 $R = 0.02$。我们认为，这源于PET规模较小，以及Jerex拥有大量可训练参数。因此，对于研究问题**RQ3** ，我们的回答是否。

	P	R	F_1
Jerex [9]	0.20	0.27	0.22
Bellan [7] + 实体关系抽取	0.32	0.29	0.30
Ours	**0.34**	0.29	**0.31**

(a)

	P	R	F_1
基线 [7] + 实体关系抽取	0.79	0.66	0.72
Ours	**0.83**	**0.82**	**0.82**

(b)

表 1：1a：Jerex、PET 基线以及我们提出的增强型处理流程的整体性能。1b：我们提出的机器学习关系抽取模块与基于规则的基线关系抽取模块各自独立运行时的性能。

图 7 按关系类型细分了 F_1 得分。根据这些结果，我们得出结论：数据集 PET 尚不适合以监督方式训练深度学习模型。目前可用的数据量不足以实现稳定收敛，因此无法构建实用的模型。为了缓解数据量不足的问题，有必要进一步研究预训练模型（如大语言模型）的应用。这些模型利用大量未标注数据来学习自然语言的结构和组成。随后，可以通过零样本设置（从未针对该任务进行过显式训练）、少样本设置（利用少量特定任务数据进行微调），或将其组合到新模型中（用于从自然语言文本中提取有用的特征）来加以应用。在[5] 文中，作者探讨了预训练大语言模型和上下文学习在提取流程相关事实与关系方面的可行性，研究结果表明，在数据量较少的情况下，这些方法有望应用于业务流程模型生成任务。

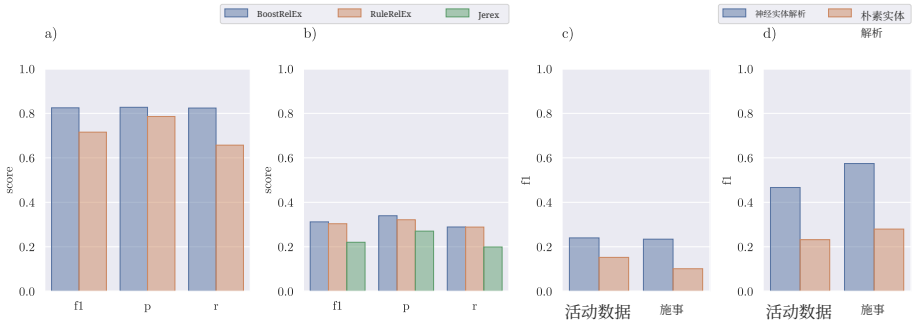


图6：a) 展示了*BoostRelEx* 与*RuleRelEx*的比较。b) 展示了我们提出的*Ours*处理流程和*Bellan+* 实体关系抽取端到端运行的性能。c) 使用命名实体识别步骤的结果比较朴素实体解析 和神经实体解析 的性能。d) 展示了与 c) 相同的比较，但基于真实提及。

我们在这项工作中提出的改进，很大一部分来自对 参与者接收者 和 参与者执行者，以及 使用关系的更好抽取。*BoostRelEx* 在抽取相同网关 关系类型时明显逊于 ***RuleRelEx*** 。造成这一现象的一个可能原因是，*RuleRelEx*使用了有关已抽取的 流程 关系的信息（参见第5.2节），而我们的机器学习方法无法做到这一点，因为它会一次性抽取所有关系。为关系类型定义抽取顺序将违背最初使用我们方法的目的：这会使方法与数据集紧密耦合，因而无法轻易应用于其他数据集。总体性能并未因此受到很大影响，因为 相同网关关系的示例非常少。不过，仍需进一步研究有助于正确抽取相同网关 的特征，以及能够学习更复杂规则的可能训练技术。例如，提及中的关键短语的同义词和上位词就是很有前景的特征。此外，对模型进行多轮处理，每次都细化其预测结果，可能有助于预测具有互斥性的关系，例如 相同网关 关系在 PET 中所体现的情况。

6.3 性能分析

逐步降低输入到*BoostRelEx* 和*RuleRelEx* 步骤的质量，会导致性能逐渐变差，这清楚地表明了错误传播（参见第5.3节）。使用数据集中的真实提及项，但采用神经实体解析步骤预测的实体，会导致 F_1 分数下降约。20，其中*BoostRelEx* 下降。12，其中*RuleRelEx*。若进一步在下游引入错误，使用命名实体识别模块，即端到端运行完整流水线，则 F_1 会下降。51，其中*BoostRelEx* 下降。42，其中*RuleRelEx*。图7 展示了各关系类型的这种性能退化情况。

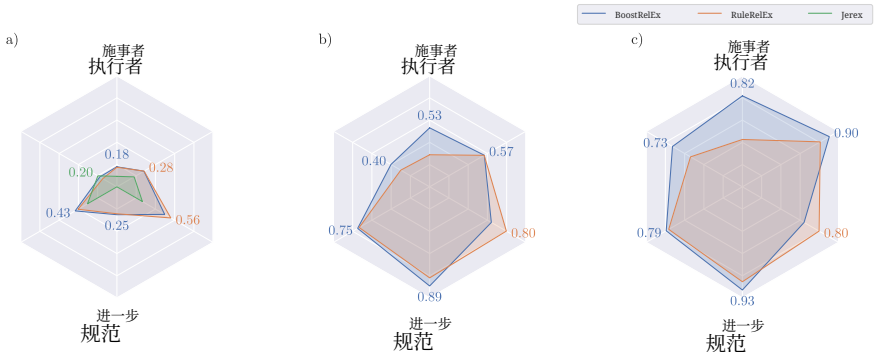


图 7：a）场景 (4) 中按关系类型划分的关系抽取结果，即完整流水线；b）场景 (5) 中使用从完美提及中解析出的实体进行关系抽取；c）场景(6) 中从完美实体进行关系抽取。

还应按照第5.3节所述，进一步研究较宽松的评估方式，以便对预测质量进行不那么保守的评估。

输入质量并不是影响关系抽取质量的唯一因素。我们发现，关系的论元之间的距离也与正确性呈负相关。关系的头实体与尾实体之间的距离越长，将其错误分类或完全无法检测到的可能性就越大。我们将关系论元的距离计算为两个实体提及之间的最小距离。图8展示了这一影响的示例。我们根据每种方法的所有预测结果创建了数据集，其元组形式为（$distance$，o），其中 $o = 1$ 表示正确预测，$o = 0$ 表示错误预测。随后，我们使用*statsmodels*[5] python 软件包对这些数据集拟合逻辑回归模型。逻辑回归模型尝试通过某个输入变量（预测变量）来预测一个结果（响应变量）。它使用逻辑回归，其形式为 $y = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$，并以使模型在给定输入 x 时尽可能准确地预测观测结果 $y = o$ 的方式选择 β_0 和β_1 。然后，我们可以使用所得曲线来讨论某种方法预测特定关系类型的能力。

对于短距离的流程 关系，*Boost-RelEx* 和*RuleRelEx*都能很好地解决。置信区间非常窄，表明拟合效果很好，因此我们认为，参数距离超过 33 个词元的关系很可能被这两种方法误分类。如果这一事实会损害所生成业务流程模型的质量，这是一个有趣的问题，但不在本文讨论范围内。

该相同网关关系经常被*Boost-RelEx* 方法错误分类，这一点在图7中已经有所体现。*BoostRelEx*似乎对该关系中论元之间的距离非常敏感，

[5]请参阅 https://www.statsmodels.org/stable/generated/statsmodels.discrete.discrete_model.Logit.html。

更常出现误分类，或者一旦词元距离超过 15 个词元，就完全无法识别示例。*RuleRelEx* 在这方面显著更为稳健，通常能够正确识别相同网关 关系，直到其论元之间的距离超过 32 个词元。对于这两种方法，拟合结果产生的置信区间都非常宽；如果能使用更大的数据集，为这种关系提供更多示例，就可以解决这一问题。

类型为进一步说明的关系可以由*BoostRelEx* 以极高的精确率和召回率抽取。图7已展示了这一点，其中F_1 的进一步说明得分为 0。93。逻辑回归拟合结果估计，论元距离与正确性之间不存在相关性。然而，对于距离超过 10 个词元的情况，置信区间非常宽，因此在有更多示例时，仍不能排除二者存在相关性的可能性。尽管*RuleRelEx* 错误预测的进一步说明关系比*Boost-RelEx*更多，但它仍能正确分类大多数情况（距离为 0–6 个词元）。因此，整体表现非常相近，如图7所示。

总之，如果能提高较长距离关系的*BoostRelEx* 和*RuleRelEx*的精确率和召回率，我们预计两者的性能都会有所提升。此外，由于我们的基于机器学习的关系抽取方法在最佳情况下优于基于规则的关系抽取方法，在最差情况下则与之相当，因此我们可以回答**RQ1**： 是。我们的深入评估表明，*BoostRelEx* 在处理长距离关系时相当稳健，只有在*RuleRelEx* 的相同网关关系上表现不如后者。不过，鉴于这种关系的样本数量较少，整体而言这一点并不重要。

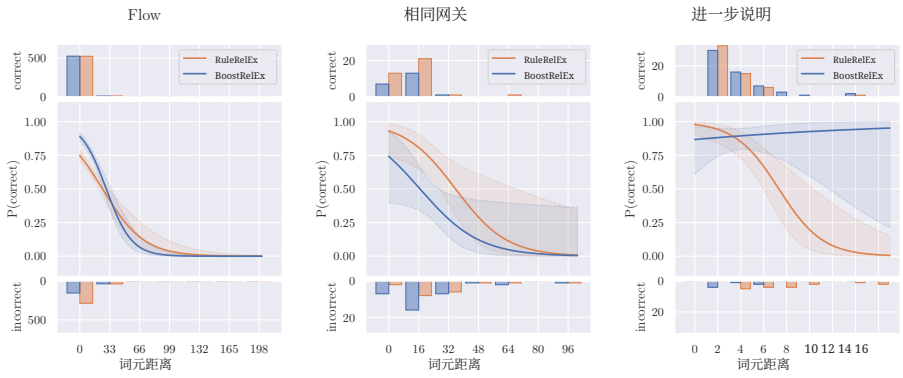


图 8：逻辑回归拟合结果，展示预测正确性与其论元之间词元距离的相关性。上方和下方的条形图分别显示正确预测（上方）和错误预测（下方）的数量。主图以实线表示拟合的逻辑回归曲线，并以透明带表示 95% 置信区间。

7 结论与未来工作

本文通过实体关系抽取拓展了业务流程信息抽取任务。我们为 PET 增添实体身份信息，并提出一种基于预训练端到端神经共指消解的抽取方法。

鉴于能够快速适应新数据、新领域或新标签集的优势，我们提出了一种基于梯度提升的新型关系抽取方法。我们展示了该方法在端到端设置下能够取得相当或更好的结果，并且在输入质量较高时显著优于基线。我们发现，PET 数据集规模尚不足以训练一种来自自然语言处理领域的先进深度学习方法 Jerex，尽管该方法在其他规模更大的相关任务基准数据集上取得了先进成果。最后，我们讨论了 PET 数据集的一些特征，这些特征会损害预测质量，例如较高的语言变异性，以及关系论元之间的距离。我们的实验证实了错误传播这一现象，即早期步骤中的错误会在后续步骤中被放大。因此，我们计划采用联合模型来抽取提及和关系，并消解流程实体，因为这些模型能够同时训练以解决这三项任务，从而减轻错误传播效应。尽管 Jerex 未能产生高质量的预测，但它及类似方法注定适用于从自然语言文本生成业务流程这一任务。因此，有必要进一步研究如何在业务流程管理这一低数据领域应用深度学习。我们计划改进实体消解模块的性能，例如纳入领域内知识，如前序步骤中的提及信息。此外，通过使用领域内数据训练预训练的神经共指消解器，对其进行微调，也是进一步提升性能的一种潜在途径。最后，最佳实践建议在评估业务流程信息抽取任务的预测质量时使用 micro F_1 分数。尽管这无疑是一项有用的指标，但我们怀疑它可能无法完全反映下游任务和用户的需求。我们计划研究替代指标，以及这些指标与人类对抽取模块的期望之间的相关性。

指代

1. Van der Aa, H.，Carmona Vargas, J.，Leopold, H.，Mendling, J.，Padr´o, L.：《在业务流程管理中应用自然语言处理的挑战与机遇》。载于：COLING（2018）2. van der Aa, H.，Di Ciccio, C.，Leopold, H.，Reijers, H.A.：《从自然语言中提取声明式流程模型》。载于：高级信息系统工程：第31届国际会议（2019）3. Ackermann, L.，Neuberger, J.，Jablonski, S.：《基于形式化意义表示的数据驱动文本流程描述标注》。载于：国际高级信息系统工程会议（2021）4. Ackermann, L.，Neuberger, J.，K¨appel, M.，Jablonski, S.：《贯通研究领域：关于联合神经关系抽取技术的实证研究》。载于：CAiSE（2023（即将出版））

5. Bellan, P., Dragoni, M., Ghidini, C.: 使用预训练语言模型辅助构建流程知识图谱。在：AIxIA 2022 论文集——人工智能进展（2022） 6. Bellan, P., Dragoni, M., Ghidini, C., van der Aa, H., Ponzetto, S.: 文本中流程模型标注指南（2022） 7. Bellan, P., Ghidini, C., Dragoni, M., Ponzetto, S.P., van der Aa, H.: 从自然语言文本中抽取流程：PET 数据集与标注指南。在：第六届 NL4AI 研讨会论文集（2022） 8. Cabot, P.L.H., Navigli, R.: Rebel：通过端到端语言生成进行关系抽取。在：EMNLP（2021） 9. Eberts, M., Ulges, A.: 一种使用多实例学习进行实体级关系抽取的端到端模型。在：ACL（2021） 10. Friedrich, F., Mendling, J., Puhlmann, F.: 从自然语言文本生成流程模型。在：CAiSE（2011） 11. Giorgi, J., Bader, G., Wang, B.: 一种用于文档级关系抽取的序列到序列方法。在：生物医学语言处理研讨会（2022） 12. K¨appel, M., Sch¨onig, S., Jablonski, S.: 利用小样本学习进行业务流程管理。《信息与软件技术》（2021） 13. Lee, K., He, L., Lewis, M., Zettlemoyer, L.: 端到端神经共指消解。在：EMNLP（2017） 14. Li, H.: 自然语言处理的深度学习：优势与挑战。《国家科学评论》（2018） 15. Li, J., Sun, A., Han, J., Li, C.: 命名实体识别深度学习综述。《IEEE 知识与数据工程汇刊》（2020） 16. Pawar, S., Palshikar, G.K., Bhattacharyya, P.: 关系抽取：综述。arXiv 预印本 arXiv:1712.05191（2017） 17. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost：利用类别特征进行无偏提升。NeurIPS（2018） 18. Qian, C., Wen, L., Kumar, A., Lin, L., Lin, L., Zong, Z., Li, S., Wang, J.: 一种通过多粒度文本分类抽取流程模型的方法。在：CAiSE（2020） 19. Quishpi, L., Carmona, J., Padr´o, L.: 从流程的文本描述中抽取注释。在：BPM 2020（2020） 20. Richards, B.: 类型/词元比：它们究竟告诉了我们什么？《儿童语言杂志》（1987） 21. S`anchez-Ferreres, J., Burattin, A., Carmona, J., Montali, M., Padr´o, L., Quishpi, L.: 释放业务流程文本描述的潜力。SoSyM（2021） 22. Sanh, V., Wolf, T., Ruder, S.: 一种从语义任务中学习嵌入的层次化多任务方法。在：AAAI 人工智能会议论文集（2019） 23. Sharnagat, R.: 命名实体识别：文献综述。印度语言技术中心（2014） 24. Sukthanker, R., Poria, S., Cambria, E., Thirunavukarasu, R.: 照应与共指消解：综述。《信息融合》（2020） 25. Wallach, H.M.: 条件随机场：简介。技术报告（CIS）（2004） 26. Waltl, B., Bonczek, G., Matthes, F.: 基于规则的信息抽取：优势、局限与前景。《Jusletter IT》（02 2018） 4 （2018） 27. Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., Sun, M.: DocRED：大型文档级关系抽取数据集。arXiv 预印本 arXiv:1906.06127（2019）