化学物质—蛋白质相互作用抽取的端到端模型：更优的分词与基于跨度的流水线策略

Xuguang Ai肯塔基大学计算机科学系美国列克星敦
xuguang.ai@uky.edu

Ramakanth Kavuluru肯塔基大学生物医学信息学部 内科学系美国列克星敦
ramakanth.kavuluru@uky.edu

arXiv:2304.01344v1 [cs.CL] 3 Apr 2023

摘要—端到端关系抽取（E2ERE）是信息抽取中的一项重要任务；随着科学文献持续呈指数级增长，这项任务对生物医学领域尤为重要。E2ERE 通常包括识别实体（即命名实体识别（NER））及其相关关系，而大多数关系抽取（RE）任务则直接假设实体已预先给定，最终只进行关系分类。由于命名实体识别中的错误可能引发滚雪球效应，导致关系抽取中出现更多错误，因此 E2ERE 本质上比单独进行关系抽取更为困难。生物医学端到端关系抽取中的一个复杂数据集是 ChemProt 数据集（BioCreative VI，2017），它用于识别科学文献中化合物与基因/蛋白质之间的关系。ChemProt 已纳入近期所有生物医学自然语言处理基准，包括 BLUE、BLURB 和 BigBio。然而，这些基准及其他独立研究通常并未以端到端方式处理该数据集，只有少数例外。本研究采用基于跨度的流水线方法，在 ChemProt 数据集上取得了新的最先进端到端关系抽取性能，相较此前最佳研究，F1分数提升了 $> 4\%$。结果表明，简单直接的细粒度分词方案有助于基于跨度的方法在端到端关系抽取中表现出色，尤其是在处理复杂命名实体时。我们的错误分析还揭示了 ChemProt 端到端关系抽取中的若干关键失效模式。

索引词—端到端关系抽取、化学物质-蛋白质关系、基于跨度的关系抽取

I. 引言

尽管在当前 COVID-19疫情期间，我们见证了生物医学研究的飞速进展，这令人惊叹，但随之而来的生命科学同行评审文献数量普遍激增，令研究人员难以持续跟进。如 Lu [1],所示，科学文献呈指数级增长，使人们通常难以掌握某一领域所有令人振奋的研究成果。为缓解这一问题，能够自动从文献中提取所报告关系信息的自然语言处理（NLP）方法日益兴起。这类关系信息中较常见的包括蛋白质-蛋白质相互作用（用于了解疾病病因学与进展）、基因-疾病关联（用于识别潜在药物靶点）、药物-疾病治疗关系（用于发现超说明书用药或评估药物再定位的潜力），以及药物-基因相互作用（用于设计靶向治疗）。对提取出的关系进行规范化并存储在结构化数据库中，可使研究人员快速检索现有研究成果，进而提出新假设并加快知识发现进程。

A. ChemProt 任务简介

为实现对自然语言处理方法的可靠基准评测，过去十年来，人们普遍致力于创建由专家标注的数据集，并将其用于共享任务，随后向更广泛的社群公开。这一广受欢迎的平台是 BioCreative 系列，它促成了 BioNLP 领域众多公开数据集的发布。作为 BioCreative VI 系列的一部分，ChemProt 抽取共享任务 [2] 是一项广受欢迎的任务，已被纳入 BLUE [3],BLURB [4], 和 BigBio [5]等知名 BioNLP 基准测试集。该任务旨在从科学文献中识别化合物与蛋白质（基因化合物[1]）之间的关系。例如，考虑以下句子：“大鼠 OMCD 中 Na+-K+- 2Cl- 协同转运蛋白 NKCC1 对 Cl- 分泌的贡献。”专家将句中的主语化学实体 Cl- 标注为宾语基因实体 Na+-K+- 2Cl- 协同转运蛋白的底物。所得关系（在 ChemProt 任务中称为“相互作用”）通常表示为三元组：(Cl-，底物，Na+-K+-2Cl- 协同转运蛋白)，其中关系标签（此处为“底物”）通常称为谓词。在 ChemProt 共享任务中，句子中主语实体和宾语实体的跨度均已给出，参赛者需要从表 I 所示的预定义谓词集合中预测二者之间化学物质-蛋白质关系的类型。（在数据集中，“底物”谓词编码为 CPR: 9。）包含非字母字符的长化学物质名称，以及相互重叠或嵌套的实体

本文报告的研究得到美国国立卫生研究院（NIH）下属美国国家医学图书馆通过资助编号 R01LM013240 的支持。本文内容完全由作者负责，不一定代表 NIH 的官方观点。

[1]由于基因编码并直接指导蛋白质的生成，基因名称与相关蛋白质名称通常相同，并且根据上下文指代其中任一实体。

表 I：ChemProt关系分　　组，依据生物语义类别划分。评估中使用的五个粗体CPR组　　.

组别	Eval	属于该组的ChemProt关系
CPR：1	N	组成部分
化学物质-蛋白质关系：2	N	调节因子 | 直接调节因子 | 间接调节因子 _
化学物质-蛋白质关系：3	Y	上调因子 | 激活剂 | 间接上调因子 _
化学物质-蛋白质关系：4	Y	下调因子 | 抑制剂 | 间接下调因子 _
化学物质-蛋白质关系：5	Y	激动剂 | 激动剂-激活剂 | 激动剂-抑制剂
化学物质-蛋白质关系：6	Y	拮抗剂
化学物质-蛋白质关系：7	N	调节剂 | 调节剂-激活剂 | 调节剂-抑制剂
化学物质-蛋白质关系：8	N	辅因子
化学物质-蛋白质关系：9	Y	底物 | 产物 | 底物产物 _ _ _
化学物质-蛋白质关系：10	N	NOT

会增加 ChemProt 中命名实体识别（NER）的难度；长句和复杂的句法结构也会妨碍关系抽取（RE）。

2017 年最初提出的 ChemProt 共享任务及随后几年采用该任务的许多研究（包括最新的 BioNLP 基准测试集 BLUE 、 BLURB 和 BigBio ）都假设化学物质和蛋白质名称已在文本中被识别出来。也就是说，实体在句子中的确切位置作为输入的一部分被提供，任务归结为预测哪些化学物质-蛋白质对参与了相互作用。虽然这种非端到端设置对于在实体跨度已知时隔离并评估正确预测相互作用的能力很重要，但对于终端用户应用而言，更现实的设置是仅提供原始输入。要自动解析文献以获取新的相互作用，就不能依赖预先识别好的实体。因此，近年来出现了将命名实体识别纳入关系抽取流程的端到端关系抽取方法。目前，采用端到端设置处理 ChemProt 的研究还不多，这正是我们稿件的重点。

B. ChemProt 任务的先前研究

如前所述，先前有几项针对 ChemProt 的研究将其建模为关系分类，因此并非端到端。研究人员尝试了卷积/循环神经网络 [6] 和树结构长短期记忆神经网络（长短期记忆网络） [7]。我们也参与了这项共享任务，并使用支持向量机集成和神经模型，在当时取得了最佳性能 [8],; 随后 Sun 等人 [9]进一步提升了这一成绩。最近，Choi 等人 [10]采用校准方法，并使用上下文化语言模型进行自训练，进一步改进了先前的结果。

在少数尝试在 ChemProt 上进行端到端关系抽取的工作中，Luo 等人 [11] 采用了序列标注方法，使用带有条件随机场（CRF）层的 BiLSTM，并加入额外的上下文化特征来识别实体，再通过规则抽取关系。Zuo 和 Zhang [12] 开发了一种基于跨度的方法 SpanMB BERT：该方法考虑了最大长度范围内的所有跨度表示，随后确定跨度类型，并识别类型有效的跨度之间可能存在的关系。

他们还采用了另一种基于跨度的模型 DyGIE++ [13]，通过一种新颖的跨度图传播方法构建跨度表示，从而联合建模命名实体识别和关系抽取；该方法利用不同的启发式规则在跨度上构建图结构。Sun等人 [14]最近也在 ChemProt 上开展了一项端到端研究。然而，仔细阅读其稿件后，我们发现他们似乎无意中只在至少包含一个关系的测试句子上评估了模型。正如 Taille 等人 [15], 所指出的，这会抬高最终的性能，因为不包含任何关系的句子也可能产生假阳性。事实上，ChemProt 测试集中有 83% 的句子没有产生任何关系。

C. 我们的贡献

我们认为，持续开发用于生物医学的端到端关系抽取模型十分重要，尤其是针对现有的公开数据集。我们采用基于跨度的关系抽取方法，在 ChemProt 数据集上开展了这项工作，具体贡献如下。

- 虽然基于跨度的方法有助于处理 ChemProt 中常见的重叠和嵌套实体，但无论命名实体识别模型多么复杂，分词都会显著影响哪些实体能够被识别。例如，**Na+-K+-2Cl- 协同转运蛋白** 跨度（见第 I-A 节）中，除了编码为基因名称的完整跨度外，“Na+”和“K+”也是标注的化学实体。常用的 ScispaCy 生物医学分词器 [16] 只会输出两个词元：“Na+–K+–2Cl-”和“协同转运蛋白”。NLTK 分词器 [17] 会生成七个词元：“Na”、“+”、“K”、“+-”、“2Cl”、“-”和“协同转运蛋白”；由这些词元组成的所有跨度仍然无法涵盖标注跨度“Na+”和“K+”。我们采用一种更简单、粒度更细的分词器，以确保在预处理阶段不会遗漏许多实体。
- 流水线模型通常包含一个命名实体识别模型和一个独立的关系抽取模型，前者的输出会传递给后者。由于流水线中的错误会产生滚雪球效应，命名实体识别错误会导致关系抽取性能大幅下降，因此这类方法在端到端关系抽取中已不再受青睐，研究人员转而更多关注联合抽取方法 [18]–[20]。然而，Zhong和Chen [21]

指出，精心设计的流水线，尤其是使用实体类型标记的设计，其效果可能优于联合模型。我们采用他们的 PURE 方法 [21]，结合不同的关系上下文表示和我们的分词方案，在 ChemProt 任务上取得了新的最先进性能。

- 我们分析了不同类型的端到端关系抽取错误：由命名实体识别错误导致的错误，以及与 ChemProt 相互作用类型本身的复杂性有关的错误。

尽管数据集的创建者提供了 ChemProt[2]，，以便其他研究人员能够进行公平的端到端比较，但我们还是将基于自身分词结果的精确跨度连同相关的预处理和建模代码公开发布：https://github.com/bionlproc/end-to-end-ChemProt。

II. 方法

A. *ChemProt* 数据集与任务

ChemProt 语料库共包含 4,966 篇 PubMed 摘要：1,020 篇训练文档、612 篇开发文档和 800 篇测试文档。实体的金标准标注包括精确的字符起始和结束位置，以及实体类型：化学物质或基因。化学物质实体和基因实体之间可以建立表 I 所示的十种关系类型之一。然而，由于其他关系类型的研究价值不高，任务评估仅使用其中五种（CPR:3、CPR:4、CPR:5、CPR:6 和 CPR:9） [2]。ChemProt 训练集、开发集和测试集中的实体与关系数量见表 II。原始标注任务的输入句子中，相关基因和化学物质名称均已标注。预期输出是句子中所有可能的化学物质-基因对之间的关系。在端到端设置中，我们将原始任务扩展为还需识别实体跨度、确定实体类型并抽取关系。总体而言，该任务在句子级别进行。也就是说，可以假设输入句子中会直接写出这些实体。不过，同时也会提供包含该输入句子的完整 PubMed 摘要。只有当该句子不足以明确判断关系时，才建议标注人员参考完整摘要。因此，模型也可以结合输入句子查看完整摘要。

B. 预处理

句子切分和分词是任何自然语言处理模型的第一步，在这里也至关重要。我们使用 Stanza [22] 程序进行句子切分。在第 I-C 节中，我们展示了 ChemProt 数据集需要更细粒度的分词，而这通常无法通过自然语言处理软件中常见的默认分词器自然实现。这主要是因为化学物质名称和基因名称较为复杂，可能由字母数字字符与特殊符号混合组成。我们采用基于空格和

表 II：ChemProt 数据集中的实体和关系数量。

数据集	训练集	开发集	Test
化学物质	13,017	8,004	10,810
GENE	12,735	7,563	10,018
化学物质-蛋白质关系：3	768	550	665
化学物质-蛋白质关系：4	2,254	1,094	1,661
化学物质-蛋白质关系：5	173	116	195
化学物质-蛋白质关系：6	235	199	293
化学物质-蛋白质关系：9	727	457	644

特殊符号的分词方法，使用 Python [23]中的标准正则表达式包。具体而言，完整的正则表达式为

`"[A-Za-zα-ωA-Ω]+|\d+|[^\s]".`

请注意，该正则表达式会将字母或希腊字符组成的序列作为单个词元捕获，并以类似方式处理数字序列，同时将所有其他非空格字符视为单独的词元。例如，使用此正则表达式时，我们的示例“Na+-K+-2Cl- 协同转运蛋白”将被分词为“Na”、“+”、“-”、“K”、“+”、“-”、“2”、“Cl”、“-”和“协同转运蛋白”。与 SciSpacy 和 NLTK 分词器不同，这十个词元能够完整捕获所有金标准跨度，包括完整字符串，以及实体“Na+”（由“Na”和“+”组合而成）和“K+”（由“K”和“+”组合而成）。请注意，这与将每个字符都作为单独词元的极端分词方式有何不同；后者可能会产生过多的候选跨度。

我们注意到，Wadden 等人 [13] 使用 SciSpacy 方法预处理 ChemProt 数据集，但称在分词过程中丢失了约 10% 的命名实体和 20% 的关系。[3] Zuo 和 Zhang [12] 使用 SciSpacy 词元，随后按“-”和“/”符号拆分，并披露丢失了 2% 的实体和关系。我们的分词方法在训练集和开发数据集中仅遗漏了实体总数的 0.4% 和关系总数的 1.37%，优于默认策略。但即使采用我们的细粒度方法，仍有一些实体无法识别。例如，考虑 ChemProt 字符串“KITD816V”；按照我们的启发式方法，它将被分词为“KITD”、“816”和“V”；然而，这些词元无法通过任何组合方式构成该字符串的金标准跨度“KIT”和“D816V”。

我们还发现，这种预处理方法有时有助于发现训练数据集中的明显标注错误。例如，在句子片段“……用维甲酸和12-O-十四烷酰佛波醇-13-乙酸酯诱导分化”中，给出的金标准实体片段明显有误，为“tinoic acid a”。正如我们手动检查时所发现的那样，我们采用的分词会漏掉这一点；实际上，这是一处错误标注，正确的标注应为“retinoic acid”。因此，我们修正了金标准实体的起始和结束字符位置，以反映这一新的片段。我们在

[2]https://biocreative.bioinformatics.udel.edu/news/corpora/chemprot-corpus-biocreative-vi/

[3]Wadden 等人的预处理脚本和结果可在线获取：https: // github.com/dwadden/dygiepp。

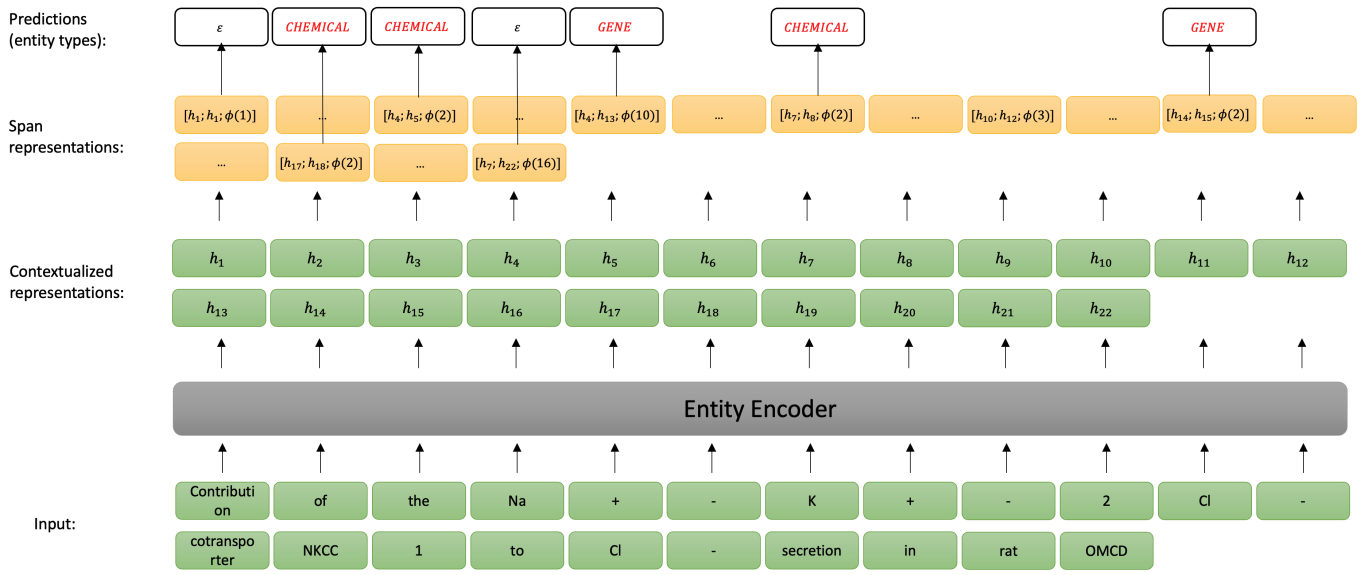

图 1：基于跨度的命名实体识别模型示例，使用一个 ChemProt 句子进行说明。对于黄色的跨度表示，将跨度边界词元嵌入（ h_i）与每个跨度长度对应的特殊词元嵌入 $\phi(l)$ 拼接。

文档键为 = 23194825的训练摘要中发现了这一问题，其中48个实体有47个因偏差一两个空格而被错误标注，我们已全部修正。除这一具体示例外，训练集或开发集中的其他内容均未作修正。

C. PURE 方法：命名实体识别模型和关系模型

如引言部分所述，我们采用普林斯顿大学关系抽取（PURE）方法 [21]（并进行了一些变体）。该方法由两个模型组成，形成一条流水线：命名实体识别模型的输出会传递给关系模型。PURE 的命名实体识别模型基于跨度，会逐一考察输入词元序列中所有可能的跨度（长度不超过固定值），并为其标注实体类型（包括类似于 IOB标注方案中 O 标签的空类型）。PURE 使用上下文化语言模型（例如本研究中使用的 SciBERT [24]或 PubMedBERT [4] ）处理输入序列，并获取每个词元的上下文化嵌入。对于每个跨度，将跨度首尾词元的嵌入与表示跨度长度的特殊词元拼接，作为softmax层的输入，以预测实体类型。图 1 用一个 ChemProt 句子示例展示了该基于跨度的命名实体识别模型：“Na+-K+-2Cl- 协同转运蛋白 NKCC1 对大鼠 OMCD 中 Cl- 分泌的贡献”。例如，$[h_1; h_2; \phi(2)]$ 是候选实体“Contribution of”的跨度表示。图中还显示了以下金标准实体：“Na+”（化学物质）、“Na+-K+-2Cl- 协同转运蛋白”（基因）、“K+”（化学物质）、“NKCC1”（化学物质）和“Cl-”（化学物质）。

识别出化学物质和蛋白质后，关系模型会接收相应的跨度，并预测每个化学物质-蛋白质对之间是否存在特定关系。一个自然一种方法是直接从实体模型中提取实体跨度嵌入，并将它们组合起来预测关系（例如，将它们拼接后输入 softmax 层）。请注意，这意味着无论评估哪个化学物质-蛋白质对以判断其是否存在潜在关系，实体跨度嵌入都不会发生变化。Zhong和Chen [21] 认为，实体表示应针对所考虑的每个特定实体对进行调整。为此，他们引入了所谓的“实体标记”，即放置在实体跨度两侧的特殊起始和结束标记。在 ChemProt 任务中，化学物质始终是主语，基因则是宾语。因此，我们有四个特殊标签 ［S:CHEM］、［\S:CHEM］、［O:GENE］和 ［\O:GENE］，它们在修改后的表示中包围相应的实体跨度。对于候选化学物质-蛋白质对（“Cl-”，“Na+-K+-2Cl- 协同转运蛋白”），前述示例句子输入关系模型时所用的修改后输入序列为：“Contribution of the［O:GENE］Na+-K+-2Cl-cotransporter［\O:GENE］NKCC1 to ［S:CHEM］ Cl-［\S:CHEM］secretion in rat OMCD”[4]。接下来，将该序列输入预训练语言模型，以获取所有词元（包括实体标记词元）的上下文化嵌入。其思路是使用这些特殊词元的嵌入作为其所包围实体的表示，以预测它们之间是否存在潜在关系。在这种设置下，考虑到类BERT语言模型中注意力模块的工作方式，很容易看出，同一句子中的同一化学物质（蛋白质）跨度的表示，会根据构成候选对时所考虑的蛋白质（化学物质）跨度而变化。在 PURE 中，这些实体标记嵌入会被传递到一个两层前馈

[4]为方便起见，我们不显示此句中各词对应的子词 WordPiece 词元。

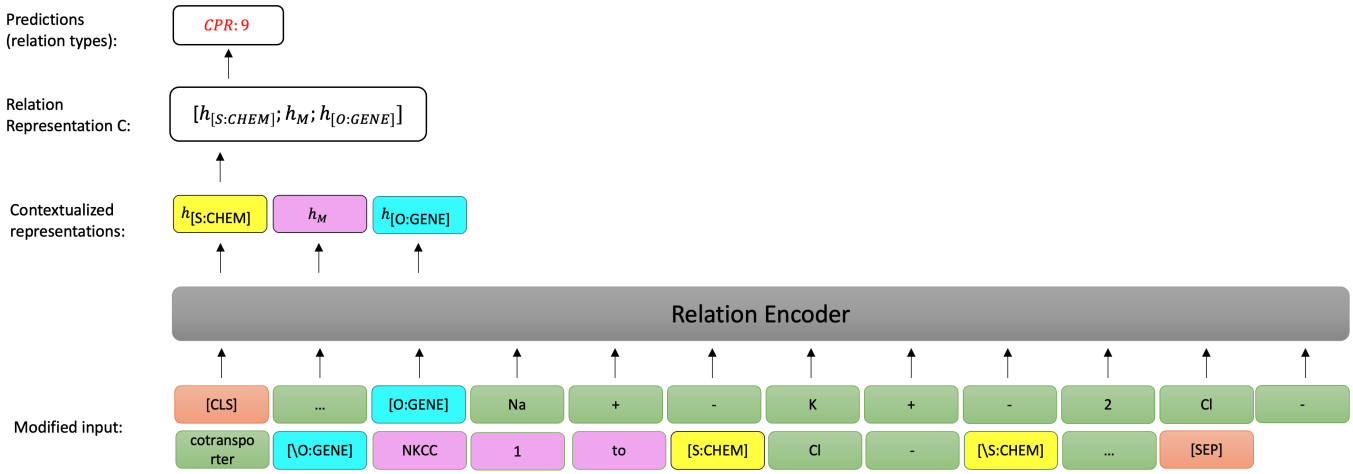

图 2：关系模型及一个示例片段，其中候选主体“Cl-”（化学物质）和客体“Na+-K+- 2Cl- 协同转运蛋白”（基因）由相应的实体标记括起。此处，关系表示由起始实体标记的嵌入以及两者之间文本的表示组成（h_M）。

网络中，该网络采用 ReLU 激活函数，随后接一个 softmax 层来预测关系类型。（我们使用 null ϵ 类表示不存在关系，同时使用表 I 中的 CPR: 3、4、5、6 和 9 类。）图 2 展示了这种实体标记方案：在输入预训练语言模型之前，用实体标记包围候选主语和宾语跨度。

实体模型使用多类交叉熵损失进行训练，依据的是每个跨度（长度最多为 L）对应的金标准标签的估计概率。关系模型在假设金标准实体已知的情况下（采用流水线方法），使用多类交叉熵损失进行训练，依据的是每个候选化学物质-蛋白质对对应的金标准关系类型的概率估计。在测试时，实体模型识别出的化学物质和蛋白质会传递给关系模型，以推断相互作用。

D. 关系表示

在第 II-C 节中，我们说明了实体标记嵌入用于预测化学物质与蛋白质之间的关系类型。不过，我们没有详细说明具体如何实现这一点。本质上，需要基于实体标记嵌入构建关系表示。根据原始 PURE 论文 [21]，，实体起始位置上下文嵌入 [S:CHEM] 和 [O:GENE],的拼接 $r_A = [h_{[\text{S:CHEM}]} : h_{[\text{O:GENE}]}]$ 被输入前馈层以预测关系。不过，我们也可以考虑其他证据，包括 $h_{[\backslash\text{S:CHEM}]}$ 和 $h_{[\backslash\text{O:GENE}]}$实体结束标记词元。此外，$h_{[CLS]}$以及两个实体之间的词元也可能为最终预测提供信息。受先前工作的启发（ [25] 和 [26]），我们考虑了表 III 所示的不同关系表示，其中 h_M 指化学物质跨度与蛋白质跨度之间所有词元的上下文化嵌入的平均值。 h_M 设为当化学物质跨度与蛋白质跨度之间没有词元时，设为0 向量。

表 III：用于预测关系类型的关系表示。

符号说明	基于实体标记的关系表示
r_A	$[h_{[\text{S:CHEM}]} : h_{[\text{O:GENE}]}]$
r_B	$[h_{[CLS]} : r_A]$
r_C	$[h_{[\text{S:CHEM}]} : h_M : h_{[\text{O:GENE}]}]$
r_D	$[h_{[CLS]} : r_C]$
r_E	$[h_{[\text{S:CHEM}]} : h_{[\backslash\text{S:CHEM}]} : h_M : h_{[\text{O:GENE}]} : h_{[\backslash\text{O:GENE}]}]$
r_F	$[h_{[CLS]} : r_E]$

E. 跨句上下文

正如Zhong和Chen [21],所证明的，跨句上下文有助于语言模型更好地执行端到端关系抽取，尤其是在涉及代词实体时。如前所述，ChemProt 任务主要被设计为句子级任务，但在需要时，允许标注人员查看完整摘要。因此，尽管在许多情况下可能不需要跨句信号，但它在某些情况下可能会有所帮助。为适应此类情况，我们将每个输入句子向左和向右扩展固定数量的词，该数量由超参数 W_{NER}（用于命名实体识别）和 W_{RE}（用于关系抽取）指定。

F. 评估指标

我们使用精确率、召回率和 F 分数来评估命名实体识别和端到端关系抽取的表现。对于命名实体识别，只有实体边界和实体类型均正确，预测实体才被视为正确。对于关系抽取，只有主客体实体的边界和关系类型均正确，预测关系才被视为正确。请注意，分词本身就会导致性能下降（如第 II-B 节所述），因为有些实体（及其相关关系）无法恢复。

对于 ChemProt 测试集，我们的分词方法会导致 0.33% 的实体和 0.98% 的关系丢失。在计算召回率时，我们将这些遗漏的实体和关系视为假阴性，此外还包括由可恢复的实体和关系产生的假阴性。

III. 实验与结果

A. 模型配置与超参数

在我们的实验中，我们将 ChemProt 中原始的训练集和开发集合并，创建了一个合并后的训练数据集，并从中选取 20% 作为新的开发数据集来调整超参数。这种数据集划分方式与其他研究一致 [11], [14],，因为从不使用测试集。作为实体模型和关系模型编码器的预训练语言模型，我们使用了 SciBERT、PubMedBERT（摘要）和 PubMedBERT（摘要+PMC）。SciBERT 使用科学文本进行训练：18% 的论文来自计算机科学领域，82% 来自广义生物医学领域。PubMedBERT（摘要）使用来自 PubMed 的摘要从头开始训练（采用自定义词汇表）；PubMedBERT（摘要+PMC）则使用来自 PubMed 的摘要和来自 PubMedCentral 的全文文章进行训练。

在所有实验中，我们将实体模型训练 50 轮，将关系模型训练 10 轮。命名实体识别和关系抽取的批次大小均为 16。我们根据训练数据进行经验评估，设定上下文窗口大小参数 $W_{NER} = 300$ 和 $W_{RE} = 100$，以及最大跨度长度 $L = 16$。其余所有模型设置（包括学习率）均与 PURE 模型代码 [21]相同。由 16 个以上词元构成的实体非常少；由于模型无法捕捉这些实体，因此会将其计为假阴性。报告的性能是使用 5 个不同随机种子进行多次模型运行后计算得出的平均值。

B. 主要结果

无论使用哪种基础语言模型，不同随机种子下的命名实体识别表现都非常相近，F 分数介于 90.3 至 91.2 之间[5]。在关系抽取方面，SciBERT 关系模型在表 III 所列的不同关系表示下，F1分数介于 65.8 至 66.1 之间。采用 PubMedBERT（摘要）和 PubMedBERT（摘要+PMC）的关系模型在不同关系表示下表现相近；其中，PubMedBERT（摘要+PMC）的关系抽取 F 值最高，为 68.8，而 PubMedBERT（摘要）的最高 F 值为 69.0。我们认为，SciBERT 表现不够理想，是因为它并非专门针对生物医学文本。由于 PubMedBERT（摘要）略优于 PubMedBERT（摘要+PMC），我们仅在表 IV 中展示以 PubMedBERT（摘要）为基础模型时，关系模型在不同关系表示下的表现。从表中可以看出，r_A、r_B，以及 r_C 的 F 值相同，但精确率和召回率的组合略有不同；r_C 通过牺牲召回率来稍微提高精确率，而 r_A 的召回率最高。纳入结束标签似乎帮助不大（r_E 和 r_F）。我们在表 VI 中将最佳结果与此前在 ChemProt 上的端到端关系抽取工作进行了比较，结果显示，命名实体识别 F 值比此前最佳结果提高了 2%，关系抽取 F 值提高了 4% 以上。

[5]本文报告的所有性能（包括表格中的结果）均四舍五入至小数点后一位。

表 IV：使用 PubMedBERT（摘要）编码器时，不同关系表示的结果。

关系表示	P	R	F
r_A	69.9	**68.3**	**69.0**
r_B	70.3	67.8	**69.0**
r_C	**70.8**	67.2	**69.0**
r_D	69.9	67.6	68.7
r_E	70.2	66.9	68.5
r_F	70.4	66.4	68.3

C. 额外上下文的消融实验

我们进行了一项简单实验，移除由 $W_{NER} = 300$ and$W_{RE} = 100$ 指定的句子边界之外的额外上下文，以观察性能是否会下降。我们对 r_C 模型进行了这项消融实验，结果见表 V，其中首行包含命名实体识别和关系抽取模型中的额外上下文。移除关系抽取上下文后，F 值仅下降 0.4%（第 2 行）。移除命名实体识别和关系抽取上下文后，命名实体识别性能下降 0.5%。与包含完整上下文的模型相比，移除命名实体识别和关系抽取上下文后，F 值下降 0.8%。考虑到这些下降幅度均为 < 1%，我们得出结论：额外上下文可能并未带来显著的性能提升。这与 PURE 论文中观察到的小幅提升相似 [21]。

表 V：使用 PubMed- BERT（摘要）模型 r_C 进行命名实体识别和关系抽取时，有无额外上下文的性能。

上下文窗口大小	NER			RE		
	P	R	F	P	R	F
$W_{NER} = 300, W_{RE} = 100$	91.0	90.9	91.0	70.8	67.2	69.0
$W_{NER} = 300, W_{RE} = 0$	91.0	90.9	91.0	70.5	66.8	68.6
$W_{NER} = 0, W_{RE} = 0$	90.5	90.4	90.4	69.2	67.3	68.2

IV. 错误分析

本节主要关注 E2ERE 流水线中的关系错误。尽管命名实体识别的 F 值超过 90，但命名实体识别的假阳性和假阴性导致了近 40% 的关系抽取错误。这令人惊讶，因为这意味着命名实体识别中 10% 的错误率导致了五分之二的关系抽取错误。可能是少数漏检的

表 VI：ChemPro 不同端到端关系抽取方法的比较 t.

模型	NER			RE		
	P	R	F	P	R	F
注意力双向长短期记忆网络-条件随机场+ELMo [11]	82.5	79.8	81.1	59.5	51.2	55.1
DyGIE++ [12]	89.7	87.6	88.7	65.4	60.5	62.9
SpanMB BERT [12]	89.3	88.3	88.8	68.0	61.5	64.6
我们的方法 （$_{TC}$）	**91.0**	**90.9**	**91.0**	**70.8**	**67.2**	**69.0**

实体参与了多个金标准关系，或者少数错误识别的实体导致了许多关系假阳性。遗憾的是，部分匹配会同时导致假阳性和假阴性（在命名实体识别和关系抽取阶段均如此）。例如，考虑以下句子片段——“由于该化合物仍具有良好的 **乙酰胆碱酯酶**抑制活性，且其六氢色烯[4,3-b]吡咯基团令人联想到毒扁豆碱的 **六氢吡咯并[2,3-b]吲哚**（3），.....”。这里，“乙酰胆碱酯酶”是金标准蛋白质，而“毒扁豆碱的六氢吡咯并[2,3-b]吲哚”是金标准化学物质。然而，我们预测出的化学物质只是其中一个子串：**六氢吡咯并[2,3-b]吲哚**。这导致金标准化学物质跨度出现假阴性，同时被标注为化学物质的子串也构成假阳性。在这个具体例子中，由于化学物质假阴性，模型漏掉了“乙酰胆碱酯酶”与“毒扁豆碱的六氢吡咯并[2,3- b]吲哚”之间的金标准化学物质—蛋白质关系，从而造成关系假阴性。然而，模型实际上预测了“乙酰胆碱酯酶”与部分匹配子串“六氢吡咯并[2,3-b]吲哚”之间的新关系，导致关系假阳性。因此，单个实体的部分匹配造成了两个命名实体识别错误和两个关系抽取错误。该实体特别长（按我们的预处理方法计算，包含 11 个词元）且结构复杂，命名实体识别模型可能因此漏检。此外，也有一些简短的缩写实体被命名实体识别模型漏检，尤其是当它们与常见词相似或属于同形异义词时。

接下来，我们讨论并非由命名实体识别错误导致的错误。这些错误是关系抽取模型特有的错误，其中化学物质和蛋白质均已被命名实体识别模型正确标注。我们先看假阴性：超过80%的错误是模型在金标准标签属于五种有效关系之一时，错误地预测了 a ϵ (null) 关系。因此，这类假阴性并非源于对两种不同关系类型的混淆，而是因为模型根本未能检测到任何相互作用。我们计算了各关系类型中此类假阴性错误所占的比例，如图3所示。显然，对于四分之一的 CPR:9（金标准底物或产物关系）关系，都会发生这种情况。例如，在以下句子中，我们发现一个假阴性：其金标准 CPR:9 关系连接了加粗的实体——“本研究的假设是，大鼠多巴胺转运体配体作为正性强化物的效力和效能存在差异，这些差异与其作为 DA 摄取抑制剂的效力和效能有关。”我们推测，这里的底物关系并未被明确陈述，而是通过化学物质与蛋白质之间的间接关系，隐含在一个冗长复杂的句子中。我们认为，这类包含细微表达差异的长句可能导致了这些 CPR:9 假阴性。

剩余一小部分关系抽取假阴性错误，是指模型预测的非空关系与金标准关系不同的情况；此时会同时产生假阴性和假阳性。我们统计了所有此类错误，发现它们只发生在某些特定的关系类型对之间。图4展示了我们的结果，其中最常被混淆的关系是 CPR:3（上调剂或激活剂）和 CPR:4（下调剂或抑制剂）。Sun 等人 [9]在其 DS-LSTM 模型中也发现了这类错误。来看下面这个句子，其金标准关系为 CPR:3：“纹状体多巴胺和**酪氨酸羟化酶**含量的降低，在接受**EHT**治疗后也没有那么明显。”模型将其预测为 CPR:4。单独来看，“降低”和“不那么明显”这两个表述可能表示抑制剂相互作用，但它们共同构成的双重否定似乎表明存在激活剂关联。由于上调和下调在调节相关术语的用法上相似，模型很难区分二者，尤其是在使用如示例所示的复杂结构时。我们还分析了各关系类型产生的假阳性的相对比例，发现被预测为 CPR:9（底物或产物关系）的关系中，有四分之一实际上是假阳性（图5）。

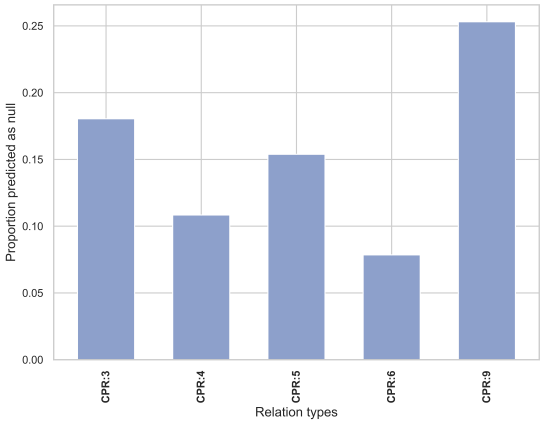

图3：被预测为空关系的金标准关系比例

五、结论

化学物质-蛋白质相互作用是药物作用机制的核心，关系到治疗潜力和副作用。

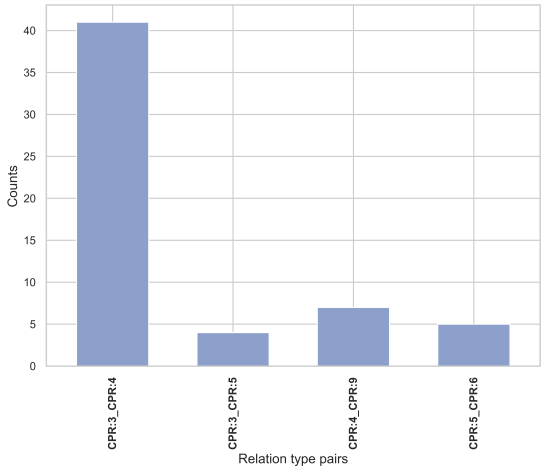

图 4：模型经常混淆的关系类型对数量

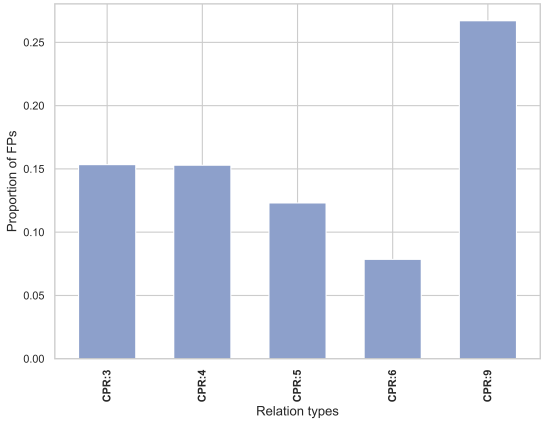

图 5：各关系类型预测中假阳性的比例

ChemProt共享任务旨在为这一重要任务建立一个自然语言处理基准，自推出以来，大多数尝试都将其视为关系分类问题，其中实体已经完成标注。很少有尝试以端到端的方式处理 ChemProt 抽取任务。在本文中，我们采用基于跨度的流水线方法，并在关系抽取步骤中额外使用实体标记，从而在 ChemProt 的端到端关系抽取方面超越了先前的最佳性能。我们还采用了一种细粒度分词方案，在保留从标准自然语言处理软件包中的默认分词器所能提取的实体数量的同时，还具备提取更多实体的能力。我们的改进幅度相当显著（F 值提升 4.4%），因此不可能仅仅源于更好的分词方案：此前最佳结果所采用的分词方案因分词损失了 2.0% 的关系，而我们仅损失 1.37%。消融实验表明，额外的句子上下文使性能提升了 <1%。

尽管我们的模型相比此前的最佳分数有了相当显著的提升，但最终 F 值仍为 < 70%，距离 90% 仍差 20 个百分点；通常达到这一水平的模型才会被认为强大，并接近人类水平的表现。错误分析显示，较长的实体跨度可能会损害命名实体识别的性能，而目前该项性能为 91%。由于 40% 的关系抽取错误源于命名实体识别错误，命名实体识别性能最后阶段的提升可能会大幅提高端到端关系抽取的分数。与其他类型相比，底物关系类型（CPR:9）涉及许多假阴性和假阳性，需要更仔细地考察，以便为该类型设计定制策略。尽管细粒度分词有助于命名实体识别步骤，但由于每个实体中的词元过多，可能无法像较少但更长的词元那样捕捉跨度的语义表示，因此可能会损害关系抽取模型。命名实体识别步骤完成后，恢复使用更简单的分词方案，可能有助于关系抽取模型更好地利用基础语言模型中的语义先验。基于生成式方法的模型，例如 BioGPT [27] (https://huggingface.co/docs/transformers/ model doc/biogpt) 和 BioMedLM (https://huggingface.co/ stanford-crfm/BioMedLM)，也可以进行调整以提升端到端性能，值得进一步探索。

参考文献

[1] Z. Lu，《Pubmed 及其他：生物医学文献检索网络工具综述》，*Database*，第2011卷，2011年。[2] M. Krallinger、O. Rabal、S. A. Akhondi、M. P. P´erez、J. Santamar´ıa、G. P. Rodr´ıguez、G. Tsatsaronis、A. Intxaurrondo、J. A. L´opez、U. Nandal 等，《BioCreative VI 化学物质-蛋白质相互作用赛道综述》，载于第六届 *BioCreative* 挑战评测研讨会论文集，第1卷，第2017期，2017年，第141–146页。[3] Y. Peng、S. Yan 和 Z. Lu，《生物医学自然语言处理中的迁移学习：在十个基准数据集上评估 BERT 和 ELMo》，载于第*18*届生物自然语言处理研讨会及共享任务，2019年，第58–65页。[4] Y. Gu、R. Tinn、H. Cheng、M. Lucas、N. Usuyama、X. Liu、T. Naumann、J. Gao 和 H. Poon，《面向生物医学自然语言处理的领域专用语言模型预训练》，*ACM 医疗保健计算汇刊（HEALTH）*，第3卷，第1期，第1–23页，2021年。[5] J. A. Fries、L. Weber、N. Seelam、G. Altay、D. Datta、S. Garda、M. Kang、R. Su、W. Kusa、S. Cahyawijaya 等，《BigBio：以数据为中心的生物医学自然语言处理框架》，载于神经信息处理系统第三十六届会议数据集与基准赛道，2022年。[6] S. Liu、F. Shen、R. Komandur Elayavilli、Y. Wang、M. Rastegar-Mojarad、V. Chaudhary 和 H. Liu，《使用基于注意力的神经网络抽取化学物质-蛋白质关系》，*Database*，第2018卷，2018年。[7] S. Lim 和 J. Kang，《使用递归神经网络抽取化学物质-基因关系》，*Database*，第2018卷，2018年。[8] Y. Peng、A. Rios、R. Kavuluru 和 Z. Lu，《结合支持向量机集成与深度学习模型抽取化学物质-蛋白质关系》，*Database*，第2018卷，2018年。[9] C. Sun、Z. Yang、L. Luo、L. Wang、Y. Zhang、H. Lin 和 J. Wang，《利用深度上下文化词表示从生物医学文献中抽取化学物质-蛋白质相互作用的深度学习方法》，*IEEE Access*，第7卷，第151 034–151 046页，2019年。[10] D. Choi 和 H. Lee，《通过校准深度神经网络和自训练抽取化学物质-蛋白质相互作用》，载于计算语言学协会论文集：*EMNLP 2020*，2020年，第2086–2095页。[11] L. Luo、Z. Yang、M. Cao、L. Wang、Y. Zhang 和 H. Lin，《从生物医学文献中抽取生物医学实体与关系的神经网络联合学习方法》，《生物医学信息学杂志》，第103卷，文章编号103384，2020年。[12] M. Zuo 和 Y. Zhang，《用于抽取细菌生境实体及其关系的基于跨度的联合模型》，《生物信息学》，第38卷，第1期，第220–227页，2022年。

[13] D. Wadden、U. Wennberg、Y. Luan 和 H. Hajishirzi，《利用上下文化跨度表示进行实体、关系和事件抽取》，载于*2019*年自然语言处理实证方法会议和第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，2019年，第5784–5789页。[14] C. Sun、Z. Yang、L. Wang、Y. Zhang、H. Lin 和 J. Wang，《Mrc4bioer：在机器阅读理解框架中联合抽取生物医学实体与关系》，《生物医学信息学杂志》，第125卷，文章编号103956，2022年。[15] B. Taill´e、V. Guigue、G. Scoutheeten 和 P. Gallinari，《停止在端到端关系抽取中进行错误比较！》，载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，2020年，第3689–3701页。[16] M. Neumann、D. King、I. Beltagy 和 W. Ammar，《SciSpacy：用于生物医学自然语言处理的快速稳健模型》，载于第*18*届生物自然语言处理研讨会及共享任务，2019年，第319–327页。[17] E. Loper 和 S. Bird，《NLTK：自然语言工具包》，载于*ACL-02*自然语言处理与计算语言学教学有效工具和方法研讨会论文集，2002年，第63–70页。[18] M. Miwa 和 M. Bansal，《利用序列和树结构上的长短期记忆网络进行端到端关系抽取》，载于第*54*届计算语言学协会年会论文集（第*1*卷：长篇论文），2016年，第1105–1116页。[19] T. Tran 和 R. Kavuluru，《用于快速端到端关系抽取的神经度量学习》，*arXiv* 预印本 *arXiv:1905.07458*，2019年。[20] M. Eberts 和 A. Ulges，《利用多实例学习进行实体级关系抽取的端到端模型》，载于第*16*届欧洲计算语言学协会会议论文集：主会场，2021年，第3650–3660页。[21] Z. Zhong 和 D. Chen，《一种出奇简单的实体与关系抽取方法》，载于*2021*年计算语言学协会北美分会会议：人类语言技术论文集，2021年，第50–61页。[22] P. Qi、Y. Zhang、Y. Zhang、J. Bolton 和 C. D. Manning，《Stanza：面向多种人类语言的 Python 自然语言处理工具包》，载于第*58*届计算语言学协会年会论文集：系统演示，2020年，第101–108页。[23] G. Van Rossum，《Python库参考手册，3.8.2版》，*Python*软件基金会，第36卷，2020年。[24] I. Beltagy、K. Lo 和 A. Cohan，《SciBERT：面向科学文本的预训练语言模型》，载于*2019*年自然语言处理实证方法会议和第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，2019年，第3615–3620页。[25] L. B. Soares、N. Fitzgerald、J. Ling 和 T. Kwiatkowski，《匹配空白：用于关系学习的分布相似度》，载于第*57*届计算语言学协会年会论文集，2019年，第2895–2905页。[26] W. P. Hogan、M. Huang、Y. Katsis、T. Baldwin、H. Kim、Y. Vazquez-Baeza、A. Bartko 和 C. Hsu，《用于生物医学关系抽取的抽象化多实例学习（AMIL）》，载于第三届自动化知识库构建会议（*AKBC 2021*），线上，*2021*年*10*月*4*日至*8*日，D. Chen、J. Berant、A. McCallum 和 S. Singh 编，2021年。[线上]。链接：https://doi.org/10.24432/C5V30P [27] R. Luo、L. Sun、Y. Xia、T. Qin、S. Zhang、H. Poon 和 T.-Y. Liu，《BioGPT：用于生物医学文本生成与挖掘的生成式预训练变换器》，《生物信息学简报》，第23卷，第6期，2022年。