通过转移学习的零散事件提取。挑战与启示

柳青[1], **张宏明**[2*], **Elior** [Sulem1], **Dan** [Roth1]

[1]UPenn计算机与信息科学系[2] 香港科技大学计算机科学与工程系
{lyuqing, eliors, danroth @seas.upenn.edu
hzhangal@cse.ust.hk

摘要

长期以来，事件提取一直是一项具有挑战性的任务，主要通过监督方法来解决，这些方法需要昂贵的注释，并且不能扩展到新的事件本体。在这项工作中，我们通过将其表述为一组文本关联（TE）和/或问题回答（QA）查询（例如，"*一个城市被攻击*"意味着 "*有一次攻击*"）来探索零镜头事件提取的可能性，利用预先训练的TE/QA模型进行直接转移。在ACE-2005和ERE上，我们的系统取得了可接受的结果，但与有监督的方法相比仍有很大差距，这表明目前的QA和TE技术在转移到不同的领域时是失败的。为了调查差距背后的原因，我们分析了剩余的关键挑战，它们各自的影响，以及可能的改进方向。[1].

1 简介

长期以来，事件提取（EE）一直是一项重要而富有挑战性的NLP任务。图1展示了ACE-2005数据集（Walker等人，2006年）中的一个 "转让-拥有 "事件，其中*触发器*是 "*购买*"，*参数*包括 "*中国*"（买方）、"*俄罗斯*"（卖方），等等。EE的子任务包括识别和分类事件触发器及其相应的论据。

主要的方法通常需要监督（例如Lin等人，2020），这在转移到新的事件本体时既昂贵又不灵活。最近的工作（Chen et al. , 2020; Du and Cardie, 2020）指出了问题回答（QA）和EE在开发监督系统方面的联系。同时，有几项工作探索了无监督的方法。Peng等人（2016）首先尝试使用基于相似性的方法，在最小的监督下提取事件*触发器*。

*这项工作是在作者访问宾夕法尼亚大学（Uni versity of Pennsylvania）时完成的。

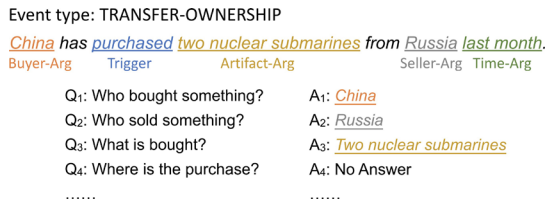

图1：ACE-2005中的一个事件例子，以及如何通过QA提取论据。

启发式方法。Huang等人（2018）和Lai等人（2020）探讨了触发器和参数提取，但设置略有不同：对一些事件类型进行训练，对未见过的事件进行测试。最近，Liu等人（2020）提出了一种基于QA的零次元参数提取方法，该方法没有处理触发器。到目前为止，还没有人提出在没有任何EE训练数据的情况下提取事件触发器和论据的方法[2].此外，现有的零点尝试的性能，特别是在论据方面，仍然远远不能令人满意，但对可能的根本原因知之甚少。

在这项工作中，我们研究了通过从文本关联（TE）和质量保证（QA）的转移学习来实现零枪击EE的可能性。请注意，给定预先训练好的TE/QA模型，提取事件可以被视为回答问题/验证关于文本的假设。例如，图1中的句子，作为前提，将包含"*有一个所有权的转让*"这一论点，从而提供了事件类型。然后，通过询问Q_1"*谁买了东西？*"，我们得到"*中国*"作为买主。同样地，Q_2，Q_3将得到卖方和艺术品，等等。

基于以上的观察，我们提出了一个直观的零散的EE方法。它不需要任何事件训练数据，但我们仍然根据开发集做出了一些设计选择。为了证明泛化水平，我们选择了具有ACE开发集的最佳模型，并在ACE与ACE开发集上对其进行评估。ERE（LDC2015E29）测试集。性能

[1]我们的代码和模型将在http://cogcomp.org/page/publication_view/943。

[2] 一个例外是Zhang等人（2021），同时完成。

计算语言学协会第59届年会论文集
和第11届自然语言处理国际联合会议（短文），第322-332页，2021年8月1-6日。©2021年计算语言学协会

在给定黄金触发跨度的情况下，我们在每个子任务上都超过了以前的零拍方法，但与有监督的方法相比，仍然不能令人满意，这揭示了在使用现成的TE/QA模型进行直接转移方面的巨大差距。为了阐明为什么会出现这种情况，我们确定了差距背后的关键挑战，并将它们分别归因于预训练模型的内在弱点、我们对它们的使用或任务本身。然后，我们用一个消减研究来剖析它们各自的影响。

我们的贡献是。(1) 我们提出了第一个基于TE/QA的事件提取系统，在没有任何事件训练数据的情况下解决了触发器和论据*的*问题；(2) 我们表明现有的TE/QA模型不能很好地支持直接的领域转移；(3) 我们对剩余的挑战、它们各自的影响以及未来研究的可能方向提供了深入的看法。

2 办法

我们的管道由两个模块组成，即触发器排除和参数提取，这两个模块都依靠预先训练好的TE/QA模型进行直接转移。

我们使用的预训练模型都是基于BERT的（Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020），包括在MNLI上训练的TE模型（Williams et al., 2018），在BoolQ上训练的Yes/No QA模型（Clark et al., 2019），以及在QAMR（Michael et al., 2018）和/或SQuAD2.0[3]（Rajpurkar et al., 2018）上训练的extrac- tive QA模型[4].TE模型，当给定一个*前提*和一个*假设时*，预测它们之间的关系（"必然"、"矛盾 "或 "中性"）。是/否QA模型将一个*上下文*和一个*是/否问题*作为输入，并返回是或否。最后，抽取式QA模型也是给定一个*上下文*，但有一个*Wh-question*，答案是上下文中的一个跨度。有了这些模型，我们设计了用于事件提取的两个模块。

2.1 触发器提取（T-Ext）

我们将触发器提取（T-Ext）表述为一个TE或Yes/No QA任务。我们只说明了TE的情况，因为其他情况只在查询格式上有区别。

争论	问题
创作	"买的是什么？"
买 方	"谁买了东西？""谁卖的东
卖 方	西？""什么东西要多少钱
价格	？"
受益人时间	"东西是为谁买的？""何时购买
地点	？""在哪里买的？"

表1：TRANSFER-OWNERSHIP事件中每个参数类型的预定义问题。

为了从语料中获得潜在的事件触发器，我们首先运行语义角色标签（SRL）作为预处理步骤。我们使用一个基于BERT的动词+名词的SRL模型[5].然后，该句子被分块为 "文本片段"，每个片段包含一个SRL谓词及其核心参数（例如：A_0, A_1, A_2）。然后，对于每一个文本片段，我们将其作为前提传递给TE模型，再加上一个格式为 "*这个文本是关于..... "的假说*。"的假设，灵感来自Yin等人（2019）。例如，BE-BORN的假说是 "*这个文本是关于某人的出生*。".然后，对于每个假设，模型返回它被前提所包含的概率。如果所有事件类型中最高的连带概率超过了阈值，我们就将相应的SRL谓词输出为这种类型的事件触发器。[6]

[3]此后简称为SQuAD。

2.2 论点提取（A-Ext）

我们将论据提取（A- Ext）的任务形式化为与预训练的提取性QA模型的QA互动序列。

给定一个输入句子和提取的触发器，我们根据事件类型的定义提出一组问题，并检索QA模型的答案作为参数预测。

考虑图1中的例子。假设T-Ext已经识别了一个触发器为 "*购买*"的TRANSFER-OWNERSHIP事件。在这种情况下，我们为当前事件类型中的每个论据类型查询了一套预定义的问题。例如，表1为TRANSFER-OWNERSHIP中的所有参数提供了一个完整的问题库。最后，为了获得参数的头部（例如 "两个核潜艇 "中的 "潜艇"），我们在AllenNLP依赖性分析器[7]，作为后处理步骤，实现了一个简单的基于启发式的头部识别。

在上述过程中，一个重要的注意事项是缺失的参数。具体来说，事件模板中的许多论据类型并不是在每个句子中都出现，例如，在图1中，没有Place论据。为了简单起见，我们把带有非空的黄金答案 "有答案"（HA）问题

[4]模型和数据集详情见附录A和B。

[5]https://github.com/CogComp/SRL-英语

[6]配置细节见附录C.2。

[7]https://demo.allennlp.org/dependency-parsing

设置	系统	钛合金	TI+TC	AI	AI+AC
划伤(监制)	Lin等人 20	78.2	74.7	59.2	56.8
擦伤（零照）	Huang等人 18[8]Zhang等人 20 我们的	55.6 **58.3** 45.5	49.1 **53.5** 41.7	**27.8** 16.3 27.0	15.8 6.3 **16.8**
黄金TI(0-shot)	Huang等人 18 Zhang等人 20 我们的	- - -	33.5 82.9 **83.7**	- - **38.9**	14.7 - **24.2**
黄金TI+TC(零距离拍摄)	Liu等人 20 我们的	- -	- -	- **44.3**	25.8 **27.4**

表2：ACE-2005的F1得分。子任务包括触发器识别（TI）、触发器分类（TC）、论点识别（AI）和论点分类（AC）。设定的定义见第3节。零枪法中的SOTA结果以黑体字显示。

在这一过程中，有两个问题被认为是 "无答案"（NA）问题，其余的是 "无答案"（NA）问题。当QA模型预判出一个空跨度或最高的非空跨度置信度低于阈值时，就被认为是输出了NA。

3 实验设置

我们在ACE-2005数据集上评估我们的系统。它的事件本体有7个类型和33个子类型，我们直接在子类型上评估T-Ext。我们使用了Lin等人（2020）的相同的训练/开发/测试分割。我们做了几个设计选择[9]我们在开发中做了几个设计选择，并在测试中报告结果，忽略了训练集。

为了证明我们的模型是如何泛化的，我们还直接评估了ERE数据集（LDC2015E29）上的最佳模型。为了适应ERE，我们为每个新的事件类型定义了一个查询。

事件提取有四个子任务。触发器识别（**TI**），触发器分类（**TC**），论据识别（**AI**），以及论据分类（**AC**）。我们在三种情况下进行实验：**从头开始**，系统执行所有子任务，没有任何黄金注释；**黄金TI**，给出黄金触发器跨度；**黄金TC**，给出黄金触发器跨度和类型。[10].

按照Ji和Grishman(2008)的说法，精确率、召回率和F1被用来评估[11].我们在头部水平上评估参数跨度，这与大多数先前的工作一致（Huang等人，2018；Wadden等人，2019；Lin等人，2020；张等人，2021）。

设置	系统	钛合金	TI+TC	AI	AI+AC
划伤(监制)	Lin等人 20	68.4	57.0	50.1	46.5
划伤 金色的TI 金色的 TI+TC（0-shot）。	我们的	39.8 - -	31.8 58.4 -	23.0 30.8 47.9	15.0 18.8 27.5

4 结果

我们报告了与几个前零点方法的比较结果（Huang等人，2018；Liu

[8]对10种事件类型进行了训练；对未见过的事件进行了测试。

[9]见附录C.2和C.3。

[10]我们没有**黄金AI**设置，因为拟议的QA- 基于A-Ext模块不能单独做AC。

[11]评估脚本改编自http://blender.cs.illinois.edu/software/oneie。

表3：ERE上的F1得分。最佳模型是在ACE dev上选择的，并直接在ERE上进行评估。

等人，2020年；Zhang等人，2021年），以及一个超视距的SOTA系统（Lin等人，2020年）。

如表2所示，在ACE测试集上，我们的系统在"黄金TI"和"黄金TI+TC"设置下的每个子任务中都优于先前的零拍方法。然而，它在"划痕"中失败了，这表明主要瓶颈在于识别准确的触发跨度。与有监督的SOTA相比，我们的系统在TI、AI和AC上仍然明显较差，特别是像其他的零射击系统。表3显示了在ERE上的结果。与ACE相比，我们的参数检测模块泛化得很好，而触发模块则不然。在黄金TI设置下，重叠事件类型的TC F1是70.4，而在新的事件类型上，它只有19.0，可能是因为ERE中新增加的事件类型有更精细的定义。例如，一个模型需要理解"一个联系人是否是当面的"，以区分MEET（当面的）、CORRESPONDENCE（非当面的）和CONTACT（不确定）。进一步的研究应该集中在如何有效地推广到新的事件上。具有微妙定义的类型。

5 分析报告

利用ACE的结果，我们现在提出了对该任务其余核心挑战的分析，以及对其个别影响的消减研究。为了进一步了解这些挑战，我们把每个挑战都归结为预训练*模型*的脆弱性（**M-错误**），我们对模型的*使用*（**U-错误**），或*任务*本*身*（**T-错误**）。

5.1 触发器提取

5.1.1 误差分析

我们首先分析了错误类型的分布。具体来说，我们手工检查了100个错误的预言，并在图2（a）中显示了计数。这里只讨论最常见的类型，其余的可以在附录E.1.1中找到。

微妙的触发（M-错误）。这是TE模型的主要内在错误（17%）。事件类型如死亡和执行，攻击和受伤，以及MEET和PHONE-WRITE是特别混乱的。

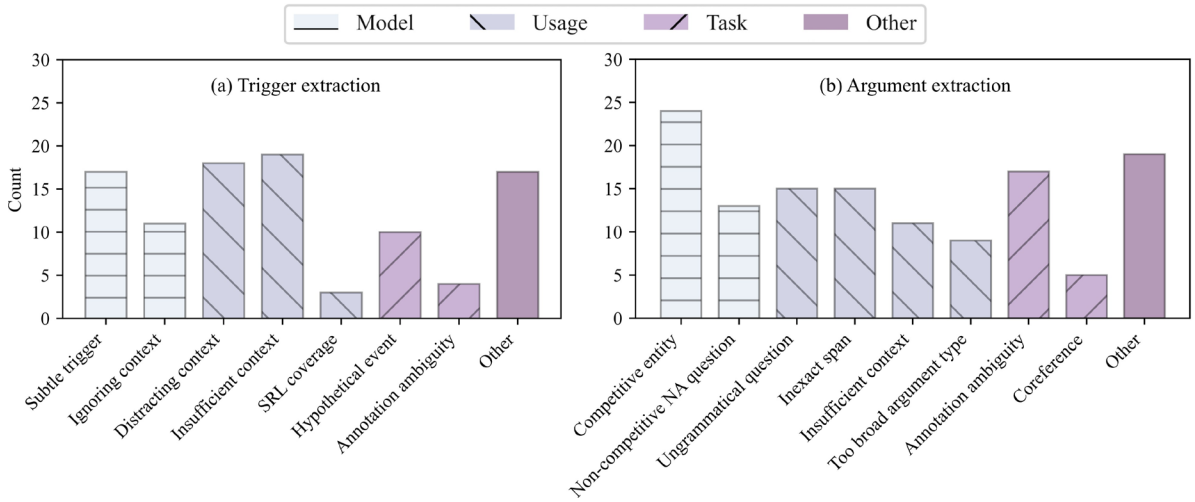

图2：100个错误预测中触发器和参数提取的错误类型。由于一个预测可能包含多种类型的错误，所以计数总和超过了100。颜色/图案表示错误类型的来源。

ing。虽然他们的定义略有不同，但该模型未能捕捉到这一层面的微妙之处。

分散注意力和不充分的背景（U-错误）。在我们对TE模型的使用中，还有两种错误类型涉及分散注意力（18%）或不充分（19%）的语境。一个分散注意力的例子是 "*该妇女的父母...发现了腐烂的尸体*"。鉴于 "腐烂 "这个词，模型预测它是一个DIE事件的触发器，因为前提是 "*身体*"。相反，不充分的语境提供的信息太少。例如，在 "*土耳其派出1000人的军队.....并表示将派出更多的军队* "这个句子中，TE模型被要求预先判断 "*派出*"的事件类型，但只看到 "*它派出更多的*军队 "作为前提，因为 "军队 "不是 "*派出*"的SRL论据的一部分。因此，该模型预测了一个TRANSFER-MONEY而不是TRANSPORT事件。

假设的事件和注释的模糊性（T-错误）。最后，有两种错误类型来自任务本身。"假设性事件"（10%）和 "注释模糊性"（4%）。假设性事件指的是 "*如果太贵，他们就不会买*"这样的句子，TE模型预测 "*买* "是一个转移-拥有事件的触发器。尽管这类事件应该按照ACE An- notation Guideline（3.4）进行注解，但这并不总是被严格遵守。其他不一致的注释情况也会导致错误，例如，在所有 "*生* "*的出现中，*触发器在某些情况下是 "*给*"，而在其他情况下是 "*生*"。

5.1.2 消融研究

我们进一步探索这两种U型错误类型，衡量它们对性能的影响，同时

控制其他因素。本节只列出了一种类型，其余的可在附录E.1.2中找到。

前提设计。为了观察**不充分和分散注意力的环境的**影响，我们选择这两种类型的所有情况，并改变前提设计。重新预测是在Gold-TI下进行的。对于不足的语境，现在的前提是en- tire句子。对于分散注意力的语境，我们采用 "最小配对前提 "策略。前提A是原始的（例如"..... *分解的身体*....."）；前提B是通过删除A中的候选触发器形成的（例如"..... *身体*....."）。然后，我们将A和B之间具有最高包含概率*差异的*事件类型作为预测。直观地说，这个差异意味着候选触发器对一个事件类型的语义*贡献*。

重新预测后，59%的错误在不充分的上下文上得到了纠正。在剩下的41%中，要么是模型仍然忽略了上下文，要么是现在较长的上下文带来了干扰。

在分散注意力的背景下，只有18%的错误被纠正。该模型仍然不能克服大多数剩余错误中的分心，这表明除了操纵前提之外，还需要一个更复杂的策略。

5.2 论据提取

5.2.1 误差分析

同样地，我们分析了100个错误的参数预测，并讨论了几个主要的错误类型。图2(b)显示了它们各自的计数。完整的解释，见附录E.2.1。

竞争性的实体和非竞争性的NA ques-

的问题（M-Error）。QA模型在 "竞争性实体"（24%）和 "非竞争性NA问题"（13%）方面具有内在的弱点。

在确定目标事件的参数时，同一类型的另一个实体，即 "com- petitive entity"，可以在上下文中共同出现。例如，句子 "*一个单位.....召开保密会议，审查欧洲的恐怖活动* "有一个MEET事件。当被问及 "*会议在哪里* "时，我们的模型回答 "*欧洲*"，而黄金答案是空的，因为 "*在欧洲*" 与 "*活动*"相连。我们发现，如果这些实体是问题所要求的类型，那么在提取性QA数据上训练的模型很容易被这些实体所欺骗。需要注意的是，竞争性实体在HA和NA问题中都可能出现。

另一种类型涉及没有任何竞争实体的NA问题。例如，考虑到 "*伊拉克军队以炮火回应* "的语境，"*何时开火* "的问题没有答案，也没有时间类型的实体来分散模型的注意力。然而，由于模型对NA问题的固有能力，它仍然可以非常有把握地给出任意的答案（例如 "*炮击*"）。

不合语法的问题（U-Error）。这种频繁出现的错误类型（15%）归因于我们对QA模型的使用。为了便于模型更好地定位目标事件，我们尽可能地在问题中模拟触发器，这有时不可避免地会使它们变得不符合语法。例如，我们对转移-所有权事件中的地点参数的问题是 "*哪里是触发点*"。只有当触发器是一个名词时，这才符合语法。因此，QA模型可能会被这种问题所迷惑。

5.2.2 消融研究

为了分离出A-Ext，我们在黄金TI+TC设置下进行消融研究。我们探讨了涉及M-Error和U-Error的四种错误类型，其中两种包括在本节中，其余的在附录E.2.2中。

预训练数据。为了研究**非专业问题的**影响，我们比较了在QAMR （He等人， 2020） 和SQuAD2.0上训练的QA模型，其中只有后者有非专业问题。结果显示，在QAMR上训练的模型大大优于在SQuAD上训练的模型（AI上+16.9；AC上+13.6）。为了解释为什么会出现这种情况，我们提出了三个假设。(1) QAMR和ACE都有一个句子。

SQuAD和ACE的HA问题类型相似，而NA问题的类型不同。(3) QAMR和ACE的每句话的答案*密度都*很高，而SQuAD则很低。我们用控制实验来测试每一个假设，但是没有一个假设能够完全解释性能差异的原因[12].

此外，我们在SQuAD的平衡样本上为HA和NA问题训练了一个二元分类器，结果是超过86的域内准确性。在ACE上，这个数字下降到57。这表明，当涉及到一个新的数据集时，QA模型甚至不能很好地区分HA和NA问题，更不用说回答它们。

问题的语法性。 为了了解**不符合语法的问题**的影响，我们手动纠正了语法错误，并使用模型重新进行预测。在所有相关的错误预测中，40%现在是正确的。其余60%的问题大多也是NA问题，证明需要的不仅仅是修复语法来解决。

6 结论

我们提出了第一个通过TE和QA的转移学习而建立的完整的零枪弹事件牵引系统。虽然QA/TE模型在标准基准（SQuAD、QAMR、MNLI）上表现非常好，但在EE数据集上使用时，它们并没有达到预期的泛化效果。我们分析了当前方法的有限成功和几个主要挑战，并为未来的改进提供了见解。

鸣谢

这项工作得到了FA8750-19-2-1004和FA8750-19-2-0201号合同的部分支持。

由美国国防高级研究计划局（DARPA）和国家情报局局长办公室（ODNI），情报高级研究项目活动（IARPA），通过IARPA合同号2019-19051600006和2019-19051600004的BETTER计划。所表达的观点是作者的观点，不反映国防部或美国政府的官方政策或立场。

我们感谢Celine Lee和Hangfeng He分别提供了SRL和QAMR模型。我们还感谢林颖、刘健、黄立夫、张浩辰和匿名审稿人的宝贵帮助和/或反馈。

语境，而SQuAD有段落。(2) SQuAD中的NA问题 "混淆 "了模型，即

[12]详情见附录E.2.2。

参考文献

陈云茂，陈同飞，Seth Ebner，Aaron Steven White，和Benjamin Van Durme。2020.阅读手册。事件提取作为定义的梳理。在*第四届NLP结构化预测研讨会上*，第74-83页。

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova.2019.Boolq：探索自然的是/否问题的惊人难度。In *Proceed- ings of the 2019 Conference of the North American Chapter of the Association for Computational Lin-guistics:Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924-2936.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.2019.BERT：用于语言理解的深度双向变换器的预训练。 在*计算语言学协会北美分会2019年会议记录*中。*Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171-4186, Minneapolis, Minnesota.Association for Computational Linguistics.

杜欣雅和克莱尔-卡迪。2020.通过回答（几乎）自然的问题进行事件提取。In *Proceed- ings of the 2020 Conference on Empirical Methods in Natural Language Processing（EMNLP）*, pages 671-683.

何杭峰，宁强，和Dan Roth。2020.Quase:问题-答案驱动的句子编码。In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss.2018.用于事件提取的零枪转移学习。In *Proceed- ings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160-2170, Melbourne, Australia.计算语言学协会。

Heng Ji和Ralph Grishman。2008.通过跨文档推理来完善事件前的牵引。In *Pro- ceedings of ACL-08:Hlt*，第254-262页。

Viet Dac Lai, Thien Huu Nguyen, and Frank Dernoncourt.2020.广泛匹配的少量学习事件检测。在*第一届叙事理解、故事线和事件联合研讨会上*，第38-45页。

Mike Lewis, Yinhan Liu, Naman Goyal, Mar- jan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer.2020.BART:用于自然语言生成、翻译和理解的去噪序列对序列预训练。在*计算语言学协会第58届年会的论文集中*，第7871-7880页，在线。计算语言学协会。

林颖，纪恒，黄飞，吴凌飞。2020.一个具有全局特征的信息提取联合神经模型.在*计算语言学协会第58届年会论文集中*，第7999-8009页。

刘健，陈玉波，刘康，毕伟，和刘晓江。2020.作为机器阅读理解的事件提取.在*2020年自然语言处理经验方法会议（EMNLP）论文集中*，第1641-1651页，在线。As- sociation for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man- dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.2019.Roberta:a robustly optimized bert pretraining ap- proach. *arXiv preprint arXiv:1907.11692*.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Da- gan, and Luke Zettlemoyer.2018.众包问题-答案的意义表述。In *Pro- ceedings of the 2018 Conference of the North Amer- ican Chapter of the Association for Computational Linguistics:Human Language Technologies, Vol-ume 2 (Short Papers)*, pages 560-568, New Orleans, Louisiana.计算语言学协会。

彭浩若，宋阳秋，和Dan Roth。2016.事件检测和协同参考与最小的监督。*在2016年自然语言处理中的经验方法会议上*，第392-402页，德克萨斯州奥斯汀。计算语言学协会。

Vasin Punyakanok, Dan Roth, and Wen-tau Yih.2008. 句法解析和推理在语义角色标记中的重要性。*计算语言学*，34（2）：257-287。

Pranav Rajpurkar, Robin Jia, and Percy Liang.2018. 知道你不知道什么。班组的无解问题。在*计算语言学协会第56届年会论文集（第2卷：短文）*中，第784-789页。

David Wadden, Ulme Wennberg, Yi Luan, and Han- naneh Hajishirzi.2019.用上下文的跨度表征提取实体、关系和事件。见*《2019年自然语言处理经验方法会议暨第九届自然语言处理国际联席会议（EMNLP-IJCNLP）论文集》*，第5784-5789页，中国香港。计算语言学协会。

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda.2006.Ace 2005多语言训练语料库。*Linguistic Data Consortium, Philadelphia*, 57:45.

Adina Williams, Nikita Nangia, and Samuel Bowman.2018.用于通过推理理解感官的广覆盖挑战语料库。In *Proceed- ings of the 2018 Conference of the North American*

计算语言学协会的分会。人类语言技术，第一卷（长篇论文），第1112-1122页，新奥尔良，路易斯安那州。计算语言学协会。

尹文鹏，Jamaal Hay，和Dan Roth。2019.零散文本分类的基准。数据集、评估和赋值方法。In *Proceedings of the 2019 Conference on Empiri- cal Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*，pages 3914-3923，Hong Kong，China.计算语言学协会。

张宏明，王浩宇，和Dan Roth。2021.无监督标签意识的事件触发和争论分类。在*计算语言学协会（ACL）第59届年会论文集发现*。

A 数据集统计

我们使用的预训练数据集包括MNLI（Williams等人，2018）、BoolQ（Clark等人，2019）、QAMR（Michael等人，2018）和SQuAD2.0（Rajpurkar等人，2018）。我们的评估数据集是ACE-2005（LDC2006T06）和ERE（LDC2015E29）。

表4显示了每个数据集中的例子数量。

数据集	火车	拓展	测试
MNLI	392,702	20,000	20,000
BoolQ	9,427	3,270	3,245
秦皇岛	73,561	27,535	26,994
SQuAD2.0	130,319	11,873	8,862
ACE-2005	17,172	923	832
ERE	-	-	2,069

表4：所有使用的数据集中的例子数量。

B 关于预训练模型的细节

我们使用三种不同的预训练表征。BERT (Devlin et al. , 2019), RoBERTa (Liu et al. , 2019)，以及BART（Lewis等人，2020）。所有的模式都是用HuggingFace Transform- ers实现的。[13].

我们使用的预训练模型检查点包括：`bert-base-uncased`（110M参数）、`bert-large-uncased`（336M参数）、`roberta-base`（125M参数）、`roberta-large`（335M参数）、`facebook/bart-base`（373M参数）、`facebook/bart-large`（406M参数）[14]。

对于TE和Yes/No QA，我们使用标准的`SequenceClassification`管道对预训练的模型进行微调。对于抽取式质量保证，我们使用`QuestionAnswering`管道对模型进行微调。[15].微调脚本改编自HuggingFace Transformers资源库中的文本分类和问题回答实例。[16].超参数值和预训练模型将通过HuggingFace模型共享服务提供。

我们在NVIDIA GeForce RTX 2080 Ti GPU上运行我们的实验，采用半精度浮点格式（FP16）和O1优化。根据任务的不同，微调需要3小时到20小时。

C 事件提取系统的细节

我们在此列出了在建立我们的事件提取系统时探索的超参数配置的完整清单。为了选择最佳配置，我们根据F1得分在开发集上进行网格搜索。

C.1 预处理

我们改编自Lin等人（2020）的预处理脚本[17].此外，我们使用几个通用的NLP工具来进一步处理文本，其中包括部分语音标签器、依赖性分析器、构词法分析器[18].

C.2 触发器提取模块

如附录B所述，我们用三种表征（BERT、RoBERTa和BART）及其基础和大型版本进行了实验。

[13] https://github.com/huggingface/变压器

[14]以上所有模型均可在https://huggingface.co/transformers/pretrained_models.html。

[15]两条管道都可以从网站上获得。https://huggingface.co/transformers/model_doc/。

[16]https://github.com/huggingface/变压器/Tree/master/examples/legacy

[17]http://blender.cs.illinois.edu/软件/oneie

[18]POS标记器来自http://www.nltk.org/；其余的来自https://demo.allennlp.org/。

预训练任务 我们有两个预训练任务选择，TE（使用MNLI作为训练数据）和Yes/No QA（使用BoolQ作为训练数据）。

前提中的SRL成分 对于每个谓语，我们只包括其本身和一些核心论据，以形成前提。我们尝试的组合包括。只有谓语；谓语、Arg0、Arg1、Arg2；谓语和所有论据。

信任度阈值 为了将一个SRL谓词识别为事件触发器，我们要求TE模型在 "Entailment "标签上的信任度分数（即QA模型在 "Yes "标签上的是/否）超过一个阈值。我们在[0.80, 0.85, 0.90, 0.95, 0.99]范围内搜索阈值。

假设的格式 我们尝试用两种策略来表述假设。

- *主题*：假设的格式是 "*此文本是关于主题*"，{}其中 "*主题*"是为每个事件类型预先定义的。例如，对于攻击，假设是 "这个文本是关于攻击的"。
- *自然的*。假设是以自然语言的形式出现的。例如，对于ATTACK，它是 "*有人被攻击了*"[19].

触发器提取的最佳配置是。
- 训练前的表示。RoBERTa-large。
- 训练前的任务。TE。
- 前提中的SRL参数：谓词、Arg0、Arg1、Arg2。
- 信心阈值。0.99;
- 假设格式。专题。

C.3 论据提取模块

如附录B所述，我们用三种表征（BERT、RoBERTa和BART）及其基础和大型版本进行了实验。

预训练数据 我们有两个用于预训练的抽取式QA数据集，即SQuAD2.0和QAMR（也包括它们的组合）。

问题格式 我们试验了两种问题格式。

- *静态的*。每个事件类型的问题都是固定的。例如，攻击事件中地点参数的问题始终是 "*攻击在哪里？*".

[19]所有假说的清单见补充材料。

- *语境化*。在可能的情况下，这些问题与事件实例的触发器一起被实例化了。例如，攻击事件中的地点参数的问题是 "*触发器在哪里？*"，其中 "*触发器*"是当前事件中的特定触发标记[20].

信心阈值 对于抽取式质量保证模型预测一个非空的答案，我们要求其信心分数应高于一个阈值。我们在[0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99]范围内搜索。

萃取参数的最佳配置是。
- 训练前的表示。RoBERTa-large。
- 训练前的数据。QAMR。
- 问题格式。语境化。
- 信心阈值。0.0（阈值几乎没有区别，因为大多数模型预测的置信度都超过0.99）。

D 全部结果

作为对第4节的补充，表5和表6分别显示了ACE和ERE的全部结果，包括精确度、召回率和F1得分。

E 分析（续）

本节详细介绍了第5节未涉及的其余错误类型和消融研究实验。

E.1 触发器提取

E.1.1 误差分析

忽视背景（M-错误）。这是另一个普遍存在的错误类型（11%），也可以归因于TE模型。该模型过于关注候选触发器本身而忽视了上下文。*考虑一下这句话："他在创造'已婚有子女'这样的节目中起了重要作用....."。*".词 "*已婚*"被错误地预设为一个MARRY事件的触发器。TE模型将其识别为一个实际的事件，而不是一个节目的名称。

SRL覆盖率（U-Error）。在所有的错误中，有3%源于目标触发器首先没有被SRL覆盖的事实。这是一个问题

[20]所有问题的清单见补充材料。{}{ }

[21]对10种事件类型进行了训练；对未见过的事件进行了测试。

设置	系统	钛合金			TI+TC			AI			AI+AC		
		P	R	F	P	R	F	P	R	F	P	R	F
划伤(监制)	(Lin et al. 2020)	-	-	78.2	-	-	74.7	-	-	59.2	-	-	56.8
擦伤（零照）	(Huang et al. 2018)[21]	85.7	41.2	55.6	75.5	36.3	49.1	28.2	27.3	**27.8**	16.1	15.6	15.8
	(Zhang et al. 2020)	58.9	57.8	**58.3**	54.6	53.5	**54.0**	19.8	38.9	26.3	9.4	18.5	12.5
	我们的	34.7	66.3	45.5	31.7	60.6	41.7	20.2	40.4	27.0	12.6	25.2	**16.8**
黄金TI(0-shot)	(Huang et al. 2018)	-	-	-	-	-	33.5	-	-	-	-	-	14.7
	(Zhang et al. 2020)	-	-	-	-	-	82.9	-	-	-	-	-	-
	我们的	-	-	-	-	-	**83.7**	35.1	43.7	**38.9**	21.8	27.2	**24.2**
黄金TI+TC(零距离拍摄)	(Liu et al. 2020)	-	-	-	-	-	-	-	-	-	25.5	26.0	25.8
	我们的	-	-	-	-	-	-	39.4	50.7	**44.3**	24.4	31.4	**27.4**

表5：ACE-2005的全部性能。

设置	系统	钛合金			TI+TC			AI			AI+AC		
		P	R	F	P	R	F	P	R	F	P	R	F
划伤(监制)	(Lin et al. 2020)	-	-	68.4	-	-	57.0	-	-	50.1	-	-	46.5
划伤	我们的	34.5	68.2	45.8	30.2	59.7	40.1	18.2	37.9	25.1	12.1	24.3	16.1
金色的TI金色的TI+TC（0-shot）。		-	-	-	-	-	80.0	33.6	41.1	37.0	21.0	25.7	23.1
		-	-	-	-	-	-	39.4	50.6	44.3	24.4	31.3	27.4

表6：ERE的全部性能。

我们对TE模型的使用。具体来说，目前的SRL系统不能完全处理名词性触发器，也不能检测像 "*靠边站*"这样的多词性触发器或像 "*死亡*"这样的形容词性触发器。**其他。**除了正文中提到的那些，其他不太常见的错误类型与核心推理有关（例如，当代词如 "*这*"成为触发器时，），专有名词（例如，历史事件如 "*起义*"），信心分数太低（因此无法识别黄金触发器），假设的模糊性（例如，"*核试验*"被预测为一个审判-听证事件，因为有 "*试验*"这个词和假设 "*有一个审判或听证*"。）

E.1.2 消融研究

SRL模型。为了研究**SRL覆盖率的**影响，我们又用两个SRL模型进行实验。伊利诺伊州的SRL（Punyakanok等人，2008年）。[22]和一个能识别几乎所有动词和名词的模型。[23].这三个模型都不能识别形容词/多词谓语。相比之下，每个模型都能覆盖90%以上的动词触发器，而名义触发器的覆盖率从60%到95%不等。在T-Ext上，覆盖率最高的模型表现最好（在TI上+4.0 F1，在TC上比覆盖率最低的模型+6.8），证明了贪婪识别的收益确实弥补了预审的成本。

预培训任务。我们的结果表明，TE-

[22] https://cogcomp.seas.upenn.edu/page/软件_视图/SRL

[23] 也可从 https://github.com/CogComp/SRL-English。

基于TC的数据远远超过其Yes/No QA的对应部分（52.6%）。一个假设是，TE模型的预训练数据（MNLI；约40万个例子）比QA模型（BoolQ；约9万个）大得多。为了验证这一点，我们在MNLI中与BoolQ相同大小的部分上重新训练了一个TE模型。结果，差距缩小到31.4%，尽管仍然相当大。这证明了训练数据大小的重要性。这也意味着，为了进一步改进目前基于TE的方法，使用更大范围的训练数据可能是有希望的。**假设设计。**我们观察到，Hy-
术语格式也起着非同小可的作用。正如附录C.2中所说，我们用两种hy- pothesis设计进行了实验，即*topical*和*natural*。实验表明，"topical "比"natural "好1.9%，表明目前的TE系统对文本措辞的敏感性。

E.2 论据提取

E.2.1 误差分析

参数类型太宽泛（M-Error/U-Error）。对于这种错误类型（9%），模型和我们的用法都是有责任的。尽管ACE对论据有严格的定义，但QA模型有时对它们的解释过于广泛。例如，在"*一名被蒙住眼睛的妇女被一名戴着头罩的武装分子射中头部*"的情况下，考虑到"*射在哪里*"的问题，模型回答"*在头部*"。这在技术上并不是错误的，但肯定也不是去掉了Place 的论点。我们不能完全追究QA模型的责任，因为这些问题

也确实太泛滥了。

不精确的跨度（U-Error）：15%的错误是由于黄金和预测的语料跨度不精确的匹配。例如，金句是 "*星期六上午*"，而预测的是 "*上午*"。尽管在我们的评估中，我们尽可能只比较短语的头部，但并不是所有的ACE论据（即那些 "价值 "类型而不是 "实体 "类型）都有头部注释。在这种情况下，目前的评估框架并没有对部分匹配给予奖励，这可能是一个潜在的改进的不完善之处。

上下文不充分（U-Error）。与触发器提取一样，在预测参数时，模型有时会被赋予不充分的语境（11%）。目标参数可能完全在谓词的SRL成分之外，因此无法提取。

核心参考和注释模糊（T-错误）。归于该任务的错误类型包括 "核心参考"（5%）和 "注释模糊"（17%）。前者指的是模型预测了黄金论据的核心参考的情况。然而，目前的评估框架仍然将其作为一个错误。后者发生在模型做出了合理的预测，但与注释不一致的情况下。例如，在 "*伊拉克军队用炮火回应* "这句话中，模型将 "*炮火* "识别为由 "炮火 "引发的攻击事件的工具。然而，没有注释任何工具。未来的评估框架应该考虑在这种人类分歧的情况下允许多个正确答案。

其他。其他错误与多个论点有关（即模型只预测了其中一个论点），缺乏文件层面的知识（即感性本身没有足够的信息），以及没有明显理由的任意预测。

E.2.2 消融研究

预训练数据。从第5.2.2节中的 "预训练数据 "段落开始，我们对QAMR和SQuAD训练之间的差距进行了三个假设的测试。

假设（1）。QAMR和ACE都有单句语境，而SQuAD有段落。

我们试图通过在具有较长上下文的新版QAMR上重新训练QA模型来验证这一点，该模型与SQuAD的长度分布相同。具体做法是：a）增加随机句子，或者b）重复原句。我们注意到

a）几乎完全不影响人工智能，但对交流有一点影响（3%），b）使人工智能降低4%，交流降低3%。因此，尽管较长的语境确实会略微削弱性能，但这并不是QAMR和SQUAD之间差距的主要原因。

假设（2）。SQuAD中的NA问题 "混淆 "了模型，即SQuAD和ACE的HA问题类型相似，而NA问题的类型不同。

为了检验这一假设，我们保留了SQuAD中所有的HA问题，组成一个新的数据集。我们还构建了一个相同大小的控制集，但从原来的SQuAD中随机抽取了NA和HA问题。我们在每个数据集上重新训练QA模型，发现与控制集相比，仅有HA的数据集在AI上带来了7%的增长，但在AC上则下降了2%。这表明，在SQuAD中增加NA问题对事件提取有不同的影响。未来的研究应该关注如何更好地将模型识别NA问题的能力转移到不同的领域。

假设（3）。在QAMR和ACE中，每个参试者的答案*密度都*很高，而在SQuAD中则很低。

为了了解这是否是原因，我们构建了一个新版本的QAMR，为每个句子只保留一个QA对。通过随机删除原始QAMR中的句子（以及它们的所有QA对），我们还构建了一个同样大小的控制集，但每个句子有多个QA对。结果显示，低密度集在人工智能方面只比控制集差0.5%，在交流方面差0.2%，表明答案的密度不是一个关键的方面。

问题中的类型约束。由于通用的问题可能是导致**论据类型过于宽泛的**原因，我们尝试使用一套新的问题模板，尽可能包含具体的实体类型要求。例如，我们不问 "*镜头在哪里*"，而是问 "*镜头的位置是什么*"，这样可以防止模型回答 "*在头上*"。然而，在重新预测后，只有11%的错误被修复，这表明编码类型约束是不肤浅的。

问题设计。 和触发器提取中的假设格式一样，问题的设计也会对论证产生影响。正如附录C.3中提到的，我们探索了两种格式，*静态的*和*上下文的*。实验表明，从 "静态 "切换到 "语境化 "可以将人工智能提高7%。

而对AC的损害为3%，这表明contex- tualized问题总体上有助于模型更好地定位事件。

语境设计。为了衡量*充分的上下文的*影响，我们现在使用整个句子作为这些实例的上下文，类似于三角函数的提取。结果显示，现在有27%的人是正确的，另外27%的人是部分正确的（不精确的跨度）。