arXiv:2303.00242v1 [cs.CL] 1 Mar 2023

DIFFQG：生成问题以总结事实变化

JeremyR.Cole[1*] **Palak Jain**[1*] **Julian Martin Eisenschlos**[1]**Michael J.Q. Zhang**[3] **Eunsol Choi**[3]
BhuwanDhingra[1,2†] GoogleResearch [2] 杜克大学 [3] 得克萨斯大学奥斯汀分校
{
j
rcole,palakj,eisenjulian,bdhingra}@google.com{mjqzhang,eunsol}@utexas.edu

摘要

识别同一篇文章两个版本之间的差异，有助于更新知识库并了解文章如何演变。在各种情境中，自然会出现成对文本：记者会撰写相似的新闻报道，而权威网站的维护者则必须确保信息始终保持最新。我们提出将成对文档之间的事实变化表示为问答对，即同一个问题在两个版本中的答案有所不同。我们发现，问答对能够灵活而简洁地捕捉更新后的内容。给定成对文档后，标注员会找出这样的问题：一个段落能够回答该问题，而另一个段落给出的答案不同或无法回答。我们发布了DIFFQG，其中包含 759个问答对，以及1153个没有事实变化的成对段落示例。这些问题旨在做到明确且能获取信息，并涉及复杂的编辑，超越了当前问题生成和事实变化检测系统的能力。我们的数据集将文档两个版本之间的变化概括为问题和答案，以一种新颖的方式研究自动更新摘要。

1 引言

给定一对陈述，我们如何识别它们信息内容上的差异？这一问题以不同形式存在于自然语言处理研究中，例如文本蕴含识别 (Dagan等人，2010年) 和 自然语言推理(Bowman等人，2015年）。这类研究最初关注的是寻找句子之间的逻辑蕴含关系。

近年来，专门用于蕴含判断的资源和模型已被应用于事实核查（Thorne等人，2018b），应用领域包括科学、教育和新闻业。这一趋势

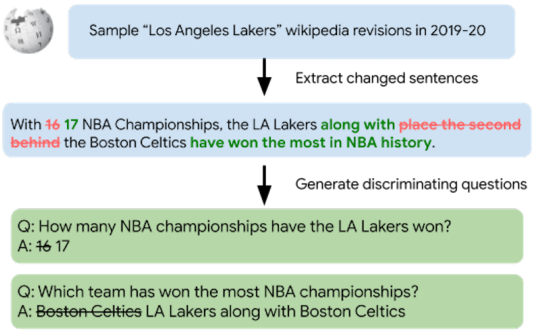

图 1：DIFFQG 由相互配对的维基百科段落组成，这些段落对应于事实编辑。其目标是给定一个答案片段，生成一个区分性问题，使得其中一个段落能够回答该问题，而另一个不能，或给出不同的答案。

凸显了逻辑蕴含与一般事实变化检测之间有限的迁移能力（Thorne等人，2018a），也凸显了这项任务对可解释模型的需求（Kumar和Talukdar，2020年）。

不同时期的维基百科修订提供了大量且易于获取的句子对，催生了 WIKIATOMICED-ITS（Faruqui 等人，2018年）和 VITAMINC（Schuster 等人，2021年）等新资源。然而，先前的工作仅限于涉及单个事实添加或变化的最小改动。我们推出了 DIFFQG，这是一个涵盖多年变化的人工标注数据集。DIFFQG 由配对段落构成，其中包含复杂的事实变化，包括同一示例中的多处添加和删除。此外，它还提供了一种以判别式问答对的形式解读预测结果的方法，用以识别这些变化。

问答对能以半结构化摘要的形式概述变化：它比知识图谱三元组更灵活，也比自由文本更实用。例如，问答对可以表示不同类型的更新：新任首相可能会更新答案，而一种新型的

*共同贡献。

部长则会新增一个全新问题。

问题生成（QG）是一项新的自然语言处理任务，其目标是生成一个由给定文档作答的问题。这种方法已有多种成功应用，包括扩充用于训练问答系统的数据集（Duan 等人，2017年；Lewis 等人，2021年）、捕捉文本中关于某主题的隐含信息（Pyatkin 等人，2021年），以及构建软知识库（Chen 等人，2022年）。以往的问题生成工作将底层段落视为静态内容（Lewis 等人，2021年），但现实生活中的文档会不断更新（Dhin- gra 等人，2022年）。随着源语料库的更新，必须添加新的问答对，并更新已有的问答对。

因此，DIFFQG 同时应对两项挑战：提供对事实变化的可解释性总结，并更新由问答对组成的软知识库。我们希望，该数据集还能够帮助评估问题生成模型生成自然、语义正确、无歧义且寻求信息的问题的质量。用于我们实验的数据集和代码将开源。[1]

我们的贡献如下：**(a)** 我们推出了 DIFFQG，这是一个由专家标注的评估数据集，包含用于概括两个段落之间差异的问题。据我们所知，此前尚无数据集涵盖如此冗长而复杂的编辑。
(b) 我们提出了一组评估指标，可用于衡量问题生成或事实变化检测方面的改进。

(c) 我们评估了一组全面的基线，以揭示当前系统的不足。

2 DIFFQG任务

DIFFQG 的目标是利用问题—答案对来捕捉两个相似段落之间的差异。具体而言，给定基础段落x_b 和目标段落 x_t，其中 x_t 和 x_b 是同一篇文章的不同版本，我们旨在生成区分性问题 Q_t。对于每个$q_t \in Q_t$，推导相应答案片段 $a_t \in A_t$ 所需的信息必须在 x_b中缺失。为限制可能问题的范围，每个答案片段 $a_t \in A_t$ 都必须是 x_t的子字符串。虽然 a_t也可能是 x_b的子字符串，但 x_b 必须缺少推导出 a_t 是正确答案所需的信息。或者，也可能存在相应的答案片段 a_b，即用 x_b回答 q_t 后得到的答案。请注意，我们认为 a_t的释义（例如将数字词汇化以及使用其他实体名称）属于等价答案。

这类区分性问题还需满足一些额外要求：问题应询问事实信息，并且能够独立成立（Choi 等人，2021年）（即，即使脱离段落单独呈现，也能被理解）。有时可能无法写出这样的区分性问题。只有在标注员相当确信目标段落中没有关于答案片段的新信息时，他们才会标记为不存在事实变化。

请看下面的示例：

- x_b = John Doe 在2012年奥运会上赢得了两枚金牌。
- x_t = John Doe 在2012年奥运会上赢得了一枚金牌。

标注员会获知，该流程的目标是收集能够通过一个段落回答、但无法通过另一个段落回答的已消歧和信息寻求型查询。我们所说的已消歧查询，是指大致指向单一答案、无需任何上下文的查询。例如，“谁在2012年奥运会上赢得了两枚金牌？”可能指向不同的人，而“他在2012年奥运会上赢得了多少枚金牌？”这类问题，如果没有 John Doe 段落，就完全无法回答。

信息寻求型查询是指，提问者无需事先知道答案，问题本身才有意义。这与 Natural Questions 的最初目标相关（Kwiatkowski 等人，2019年），也对应于*Cranfield*-style 问题，相关描述见 Rodriguez 和 Boyd-Graber（2021年）。例如，“阿尔·卡彭的母亲以什么为生？”看起来就是一个信息寻求型查询。另一方面，“哪位意大利裔美国黑帮分子的母亲是女裁缝？”则不是：除非提问者已经知道答案，否则他们为什么会假定这样的人确实存在？我们将在下一节介绍用于获取这类区分性问题集的标注流程。

3 数据收集

收集此类区分性问题并非易事。因此，我们引入了由专家标注员（本文作者）参与的分阶段标注流程，

[1] https://github.com/google-research/ 语言/tree/master/language/diffqg

并使用问题生成模型辅助标注。下面将介绍这一流程（如图所示：图 2）。

3.1 输入段落对选择

首先，我们从 Natural Questions（NQ）训练集中提取包含实体的维基百科页面（Kwiatkowski 等人，2019年）。具体而言，我们查找 2008 年至 2020 年间的维基百科快照中的页面。抽取一篇基础文档后，我们找到该文档一年后的版本，并将其用作目标文档。将这两篇文档作为语料库，我们计算各句子对 TF-IDF 向量之间的余弦相似度，并选取相似度最高的一对。相似度大于 0。8 或小于 0。25 的句子对将被舍弃。为保留有意义的变化，我们确保至少编辑一个名词或数字，并对编辑内容涉及命名实体或至少五个词元的样本进行过采样。

因此，该过程侧重于一年间累积的编辑，所涉及的变化范围从五个到二十个词元不等，因此这些变化在语义上比现有的事实变化检测数据集更加丰富，适用范围也更广。

3.2 种子问题-答案对生成

每个目标段落都有数量极其庞大的可能答案片段；为方便起见，我们将其限制为仅使用伯克利神经句法分析器（Kitaev 和 Klein，2018年）识别出的名词短语。为提高标注速度，每个示例都以一个种子问题开始，该问题由问题生成模型根据目标段落和答案片段生成。具体而言，我们使用了一个在 SQuAD 数据集（Raffel 等人，2020年）上进行微调的 T5-XXL 模型（Rajpurkar 等人，2016年）。

3.3 标注流程

DIFFQG 的标注由六名专家志愿者分三个阶段完成。首先，标注员会获得上文所述的成对段落，以及答案片段和种子问题，这些内容共同构成一个示例候选项。然后，他们为每个示例候选项标注以下五个选项之一：

接受 种子问题满足区分性问题的所有要求，无需修改。

上下文 种子问题询问的是恰当的话题，但脱离段落的上下文便无法作答。例如，像“他赢得了什么？”或“奥运会在哪里举行？”这样的问题，都缺少成功作答所需的上下文。

编辑 这个候选示例答案有一个区分性问题，但该问题与种子问题不同。有时，这是因为种子问题没有涵盖段落中包含的新信息；另一些时候，种子问题本身就毫无意义。

拒绝 这个候选示例没有有效的区分性问题。换句话说，这些都是负例。有时，目标段落根本不包含任何新信息；不过，它可能包含关于其他答案片段的新信息，而非候选示例中的答案片段。在我们之前的 John Doe 示例中，除了间接提及外，并没有关于“奥运会”的新信息。

跳过 目前无法确定这个候选示例是否存在有效的区分性问题。这可能是因为答案片段表达生硬或冗长，例如“在 2012年奥运会上获得两枚金牌”。也可能是因为答案片段与其他实体存在间接关系，导致难以判断是否有关于该答案片段的新信息。最后，即使显然存在新信息，编写信息寻求问题也可能很困难，例如在前一个示例中以“John Doe”作为答案片段来编写问题。

每个示例候选项均由两名标注员进行评审。除非两名标注员都同意添加、拒绝或跳过，否则将由第三位标注人员作出决定。对于一名标注员选择了上下文或编辑的示例，第三位标注人员负责根据指南编写正确的问题。如果一名标注员选择添加或拒绝，而另一名标注员选择跳过，则第三位标注人员可以确认添加或拒绝；如果无法作出决定，也可以选择跳过。请参阅附录 A.2 以了解标注界面。

3.4 问题编写指南

请注意，撰写一个单一、不依赖上下文且具有信息寻求性质的问题来概括两个段落之间的差异，可能颇具挑战。遇到似乎无法完成的情况时，鼓励标注员跳过该示例。如果需要补充上下文，鼓励标注员在不牺牲流畅性的前提下尽可能添加上下文，以便读者无需了解源段落也能回答问题。当标注员从头开始撰写问题时

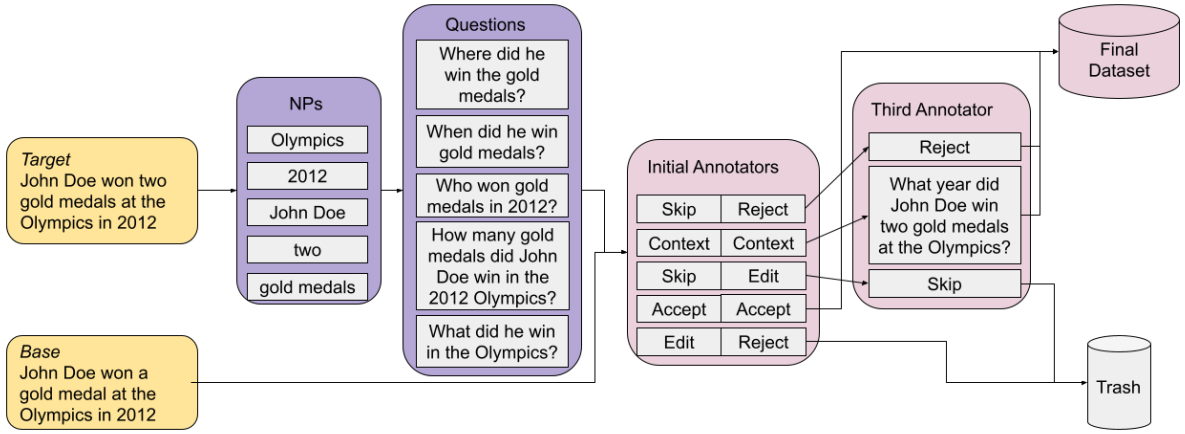

图 2：DIFFQG 标注流程。从目标段落中提取名词短语，并由问题生成模型生成初始问题。标注员判定生成的问题是否可作为令人满意的区分性问题（接受）、是否必须编辑（上下文、编辑），或是否不包含新的事实信息（拒绝）。第一阶段结束后，第三位标注人员会解决无法判定的情况（跳过），最终得到一组问题和负例。如果最初两名标注员意见不一致，或第三位标注人员无法解决无法判定的情况，则丢弃该示例。

对于编辑情形，鼓励他们设想一个问题：如果针对基础段落提出，该问题的答案会不同、无法回答，或其前提不成立。虽然将问题限定为一个答案片段可以减少歧义，但这项任务仍存在歧义；在处理大型且复杂的编辑时，这种情况不可避免。

	段落	答案	平均编辑后的词元数
有变更	391	759	12.9
无变更	478	1153	14.1
总计	672	1912	13.7

表1：DIFFQG 的数据集统计信息。编辑后的词元表示给定段落对中添加或删除的词元平均数。

3.5 数据统计

我们的初始标注流程从 8,530 个示例候选项开始，这些候选项取自999 个段落对。标注员跳过了近75% 的示例候选项，最终留下 1,912 个示例。其中，约40%，即 759 个示例存在事实变化，因此为其撰写了用于区分的問題，留下1153 个负例。在检测到事实变化的文本片段中，标注员修改了65%，即 494 个示例中的问题： 45% 标注为上下文，20% 标注为编辑。详细的数据集统计信息见表1。

请注意，在所有问题被原样接受或被视为负面示例的情况下，至少有两名标注员同意这一评定。然而，人工编写的问题未经核验；两名名标注员都同意存在一个区分性问题，但不一定知道具体是什么问题。为此，我们评估了一组包含五十个问题的小型样本，发现第二名标注员约有 85% 的概率会写出一个等价的问题。

4 动机

在上一节中，我们介绍了 DIFFQG 及其标注流程。如前所述，DIFFQG 的目的是检测并描述事实变化。具体而言，DIFFQG 粗略衡量了模型自动构建问答对数据库、以概括这些变化的能力。还可以采用许多其他格式来概括事实变化，例如段落、知识库三元组或单独的主张。

段落虽然能够表达细微差别，却缺乏原子性。因此，很难确切判断发生了什么变化，也难以比较两项变化。这使得段落不太适合作为数据库。

另一方面，知识库三元组能够描述的事实变化类型有限：无论具体设置如何，节点和关系都来自某种固定词汇，这可能导致一些有趣的变化无法保留。例如，涉及一组实体、日期范围、各种数字或抽象信息的变化，都可能难以描述。

另一种替代方法是列出主张，类似于 Vita min-C（Schuster 等人，2021年）。这种方法同样是原子化的，而且比

知识库三元组更灵活。不过，问答对有一些优势。首先，问答对是半结构化信息，构成松散的键值对。事实编辑可能会改变现有问题的答案，或添加与全新信息相对应的信息，因此需要提出一个新问题。相比之下，主张之间更难建立关联。

最后，问答对很有意思，因为问答本身也很有意思。以往研究曾将问答对数据库用作提升问答性能的方法（Lewis 等人，2021年）。因此，自动创建和更新此类数据库的方法似乎非常有用。随着事实语料库不断变化，我们设想，此类数据库需要通过迭代方式进行更新。

5 评估指标

DIFFQG 可用于衡量三项相关任务的表现。

事实变化检测 给定一个由基础段落、目标段落和答案片段组成的示例，目标是判断是否存在有效的区分性问题。换言之，就是判断与基础段落相比，目标段落中是否包含有关该答案片段的新信息。为衡量这一点，我们报告标注中是否存在区分性问题的准确率、精确率、召回率和 F1 分数。请注意，始终预测没有变化可达到 60.3% 的准确率，但 F1 分数为 0%；而随机猜测的 F1 分数为 44.1%。

判别性问题生成 给定一个目标段落和答案片段，撰写一个具体、明确且以获取信息为目的的查询，使其能够通过目标段落得到回答。为衡量这一点，我们将机器生成的问题与经过人工核验、编辑或手写的问题进行比较。我们使用两个不依赖模型的评估指标 Rouge-1 和 ROUGE-L（Lin，2004年），分别衡量问题之间的词元级重叠和最长子序列重叠。我们还考虑了两个基于模型的评估指标：BLEURT（Sellamet al.，2020年），这是一种基于 BERT（Devlin et al.，2019年）学习得到的文本相似度评估方法；以及一个查询相似度模型（Reimersand Gurevych，2019年），该模型在 Quora 问题对上进行训练 [2]。

[2]huggingface.co/cross-encoder/quora-roberta-large

请注意，尽管我们在标注过程中使用了问题生成模型，但我们仍然对判别性问题生成进行评估。请注意，所有这些问题都经过人工审核，只有非常流畅的问题才会被保留。由于问题生成模型生成的内容中流畅程度很高的问题各不相同，这组问题并不像初看起来那样简单。不过，我们还将人工撰写或编辑的问题分离出来，并对该集合进行独立评估。

完整系统 这是对 DIFFQG 整体性能的衡量。我们沿用判别性问题生成的评估指标；如果事实变化检测有误，则将 BLEURT、ROUGE-1、ROUGE-L 和 QuerySimilarity 的值设为 0.0。

6 方法

如前所述，DIFFQG 可以视为两项任务的组合：事实变化检测和判别性问题生成。因此，我们简单的基线系统将其作为一条流水线处理：首先预测是否存在事实变化，若存在，再生成一个区分性问题。我们还提出了基线模型，通过一次预测联合解决这两项任务。我们的方法如图3所示。

请注意，我们的方法均未使用 DIFFQG 中的任何部分作为训练数据，因为该数据集仅用于评估。相反，模型是在规模更大的现有问题生成和事实变化检测数据集上训练的。

6.1 事实变化检测

我们提出了五种基线，它们基于答案等价性，或同时基于问题和答案等价性。

答案等价性基线

我们的简单基线（重叠答案）仅当答案片段不在基础段落中时，才将示例归类为存在事实变化。在查找与段落的词元重叠之前，会先对该片段进行规范化。

我们基于模型构建的简单基线与之类似，但使用答案等价性模型（Bulian 等人，2022年）。该模型将目标答案片段与所有有效的基础答案片段进行比较，如果与其中任何一个都不匹配，就判定发生了事实变化。答案等价性模型还会将每个答案片段对应的候选问题作为输入。

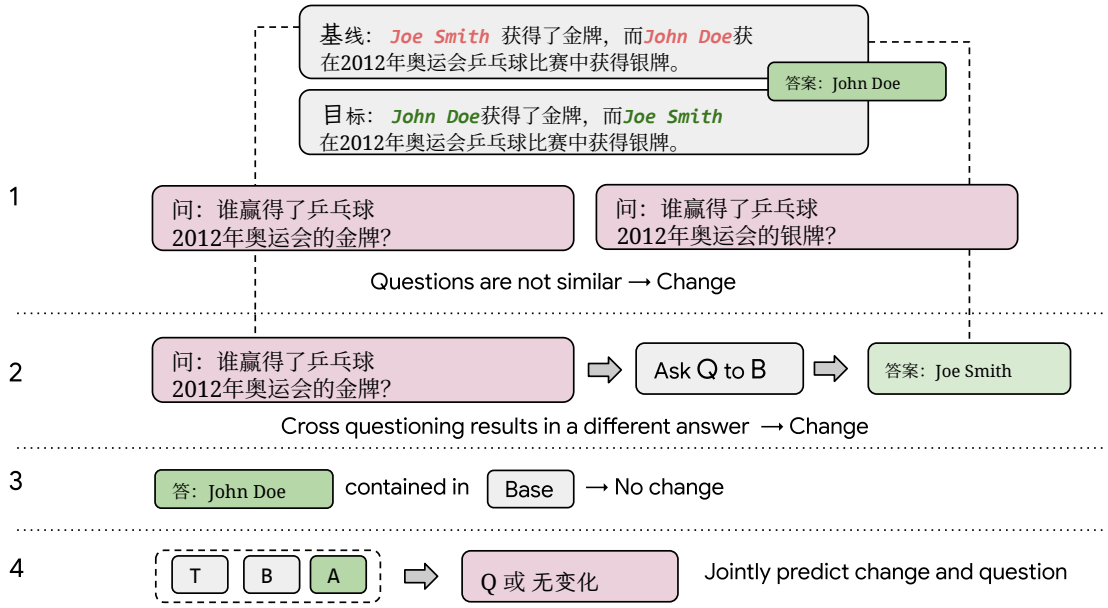

图3：DIFFQG 上事实变化检测的方法：(1) 问题相似度 (2) 交叉提问。问答等价方法结合了 (1) 和 (2)，只有当两个系统都判断发生了变化时，才判定为变化。(3) 答案重叠 (4) 一种语言模型，它联合学习事实变化检测和区分性问题生成任务，并解码出区分性问题或表示无变化的特殊标记。

问答等价基线

先前的方法只考虑答案片段，而忽略上下文。在这里，我们考虑使用问题生成模型处理段落和答案片段，以确定是否存在新信息的方法。

首先，我们使用 答案重叠方法，找出与目标答案片段等价的基础答案片段。接着，我们要考察根据这些答案生成的问题在两个段落中是否也等价。为此，我们使用一个在 Quora 问题对上训练的 T5-XXL（Raffel 等人，2020年）模型 [3]，预测问题对是“重复”还是“不重复”。如果问题并非重复，我们就认为这个示例存在事实变化。因此，现在可以识别出在两个段落中都出现、但上下文不同的答案片段，并判定其存在事实变化。这将提高答案重叠方法的召回率。

第二种方法添加了交叉提问过滤器（*Cross-Q*）。给定根据目标段落生成的候选问题，我们使用阅读理解模型尝试根据基础段落回答该问题。我们在 SQuaD v2（Rajpurkar 等人，2018年）问答数据集上训练 T5-XXL 模型，使其以段落和问题为输入，并输出答案。

[3]https://www.kaggle.com/c/quora-question-pairs

如果模型预测没有答案，或预测的答案与目标片段不同，我们就将该示例归类为存在事实变化。最后，问答等价性方法结合查询相似度模型和交叉提问模型，以提高精确率。此时，只有当两种方法都判定存在事实变化时，我们才认为示例存在事实变化。

6.2 问题生成

随后，我们将每个事实变化检测基线与一个问题生成模型相结合。我们使用一个类似的、在 SQuAD 上微调的 T5-XXL 模型，如第3.2节所述。不出所料，用来生成问题的模型在它最初编写的问题上表现良好；然而，这并不是公平的基线。因此，我们还测试了该模型的一个采样版本，以及使用不同随机种子重新训练的版本。我们还测试了一个在 Natural Questions 上训练的类似模型（Kwiatkowski 等人，2019年）。

6.3 联合系统

文中介绍的许多技术第 6.1 节效率低下，需要多次运行各种模型。例如，查询相似度方法需要针对基础段落中每个答案片段、每个示例运行一次模型，这意味着每对段落都需要进行平方级次数的运行。我们

还探索了能够直接比较基础段落和目标段落、无需任何中间步骤的方法。这些方法会同时检测是否存在事实变化，并生成区分性问题。

在银标数据上微调

我们使用与第3.1节相同的流程，挖掘更多维基百科段落对。然后，我们从目标段落中找出所有可能的答案片段。为了给任务中的事实变化检测组件创建银标训练样例，我们使用效果最佳的启发式方法问答等价性为每个目标答案片段标注。

然后，我们将这些标签转换为文本到文本任务。对于每个存在事实变化的示例，我们将 SQuAD 模型生成的问题（第6.2节）作为目标。对于不存在事实变化的问题，我们将“None”作为目标。模型的输入是基础段落与目标段落的拼接，并用一个特殊标记标出目标答案片段。

该模型是一个 T5-XXL，其初始化所用的问答生成模型与生成原始问题的模型相同。

在 VITAMINC 上微调

VITAMINC 任务（Schuster 等人，2021年）也包含事实变化检测组件。我们从 VITAMINC 修订标记数据集中抽取负例，使用随机选择的名词短语作为答案片段。

对于正例，我们需要找出一个包含事实变化的特定答案片段，以及与之对应的区分性问题。虽然在 VitaminC 中没有与此任务直接对应的任务，但事实核查任务与之有些相似。该数据集由一条证据*e*、一条支持该证据*c* 的简化声明 *e*、与之对应的编辑后句子 *e’*，以及一条与该证据相矛盾、却支持*c’* 的编辑后声明 *e*，并支持*e’*。根据 *(c, c’)* 之间的词元级差异确定答案片段*a*，并使用第*c* 6.2节中的问题生成模型，根据*a* 生成一个以c 为条件的问题。由于*c* 是一个简单句，我们根据经验发现生成的问题质量很高。

数据集（*e*、*e’*、*a*）被转换为文本到文本任务，并按照上述相同步骤微调 T5-XXL 模型。值得注意的是，使用等量正例和负例训练的模型在 DIFFQG 上表现不佳。在我们最终的 VITAMINC 银标数据集中，我们仅使用了 10% 负例，以取得合理的性能。

模型	Acc	P	R	F1
随机基线	50.0	39.5	50.0	44.1
答案重叠	82.1	79.1	74.6	76.7
答案等价性	81.1	84.6	64.2	73.0
查询相似度	77.7	65.4	93.3	76.9
Cross-Q	76.9	65.9	86.8	74.9
问答等价性	**83.9**	76.7	85.5	**80.9**
基于问答等价性的微调	82.5	78.4	77.1	77.7
基于 VITAMINC 的微调	81.5	79.9	71.4	75.4

表 2：事实变化检测的评估指标。请注意，这些模型都没有事实变化检测训练数据，而是通过其他任务或启发式方法进行验证。随机基线假设猜测“变化”或“未变化”的概率相同。准确率=准确率，P=精确率，R=召回率，F1=F1 分数。**粗体**表示最佳模型，第二佳模型带下划线。

7 结果与讨论

我们分别报告事实变化检测、问题生成和完整系统的结果。此外，我们还分别报告整体性能，以及仅针对人工编写问题子集的性能；这些问题对应于标注阶段标记为“编辑”或“上下文”的句子。附录中提供了带有模型输出的部分示例，以展示基线方法的能力和典型错误。附录 A.1 。

7.1 事实变化检测

表 2比较了不同系统在事实变化检测任务上的表现。我们发现，问答等价性优于启发式基线方法。尤其是，当答案片段文本保持不变、但周围上下文发生变化时，它能更好地处理这类情况。例如，在段落 “在新罕布什尔州行政委员会中，*Laconia* 属于第一区，其代表是 <添加：共和党人 *Joe Kenney*> <删除：民主党人 *Michael J. Cryans*>。”中，问答等价性正确捕捉了与答案片段 “新罕布什尔州行政委员会”相关的新信息，并将其转化为问题：*“Joe Kenney* 代表 *Laconia* 所属的哪个州议会？”然而，如附录 A.1所示，即使段落之间没有语义编辑，该方法也容易检测出虚假变化。

在问答等价性银标数据上微调的联合系统似乎并未优于

问答等价性。尽管它似乎受益于额外的上下文，但仍难以处理冗长且复杂的编辑。不过，该模型只需进行一次推理即可完成这两项任务。

尽管能够使用额外数据，经过 VITAMINC 训练的模型仍无法超越我们的基线。VITAMINC 风格的编辑与 DIFFQG 编辑有很大不同，通常只涉及小幅改动。因此，在 VITAMINC 上微调的模型难以处理大幅短语改动或句子重构。

7.2 问题生成

模型	R-1	R-L	QSim	BLRT
SQuAD-seed	71.7	74.9	66.3	74.1
SQuAD-sampled	59.6	63.2	57.8	65.6
SQuAD-retrained	58.0	61.7	58.9	65.6
NQ	20.4	39.6	22.9	41.5

表3：DIFFQG 正例子集上问题生成模型的性能变化。我们报告了 SQuAD 模型的三个不同版本，其中第一个模型与用于生成标注种子的模型相同。R-1=ROUGE-1，R-L=ROUGE-L，QSim=基于查询相似度模型的准确率，BLRT=BLEURT。

在表3中，我们比较了正例子集上的问题生成基线模型。在表4中，我们考察了相同模型在人工编写样例子集上的表现：这些样例的变化来自表1。

模型	R-1	R-L	QSim	BLRT
SQuAD-seed	56.5	61.4	48.2	61.3
SQuAD-sampled	50.9	55.2	47.0	58.3
SQuAD-retrained	50.3	54.5	50.4	59.4
NQ	20.4	39.1	20.0	40.2

表4：DIFFQG 人工编写问题上的问题生成模型性能变化。第一个模型与我们用于生成种子问题的模型相同。R-1=ROUGE-1，R-L=ROUGE-L，QSim=基于查询相似度模型的准确率，BLRT=BLEURT。

本次评估的主要目标是检验：第3.2节中所述的种子模型直接生成的问题，是否仍然适用于评估 DIFFQG 系统。我们发现，从同一模型中采样以及按照相同流程重新训练（如第6.2节所述）整体数据集上的性能会显著下降。这表明，除非能够使用同一个模型，否则这些经过人工验证但并非人工编写的问题仍可用于评估。尽管如此，种子模型仍可视为当前 DIFFQG 问题生成性能的大致上限。

人工编写的问题（见表4）对问题生成模型而言，似乎更难复现。模型性能大幅下降：最明显的是，曾为整个数据集编写部分问题的种子模型表现下降最多，而它本应能准确复现这些问题。

我们还注意到，在 Natural Questions（Kwiatkowski 等人，2019年）上微调的问题生成模型，其问题风格与 SQuAD 有显著差异。这可能是因为 SQuAD 问题最初是根据段落生成的，而 Natural Questions 的问题形式则更加自由。此外，我们发现 Natural Questions 模型在许多场景中都会产生幻觉。这也反映出在 DIFFQG 上，经过 Natural Questions 训练的问题生成模型表现不佳。

需要注意的是，用于概括事实变化的问题可能有极其广泛的形式。虽然将答案限制为单个答案片段缩小了这一范围，但我们仍发现存在多个有效问题的情况。因此，模型生成的问题可能完全有效，只是根据标注员的判断，并非最切题的问题，因而双方可能存在一些分歧。

7.3 完整系统

完整的 DIFFQG 评估指标见表5，其中包括两个分别在 VITAMINC 和问答等价性数据上训练的微调系统，以及两个流水线系统；后者将事实变化检测模型连接到问题生成模型。在流水线实验中，我们使用第6.2节所述的重新训练版 SQuaD 模型。我们在完整的 DIFFQG 数据集和人工撰写子集上评估这些模型。

总体而言，所有系统的表现都相差不大。问答等价性表现最佳，微调版本和简单的启发式模型紧随其后，这表明未来仍有很大的创新空间。在人工撰写子集上，性能进一步显著下降，突显了人工编写的问题所带来的挑战。

模型		事实变化检测			完整系统					
					All			人工编写		
变化	QG	P	R	F1	R-L	Qsim	BLEURT ROUGE-L		Qsim	BLRT
流水线系统										
重叠答案	SQuAD	79.1	74.6	76.7	71.3	70.8	72.3	40.6	38.1	43.6
问答等价性	SQuAD	76.7	85.5	**80.9**	71.4	70.6	**72.7**	46.8	43.7	**50.8**
联合系统										
基于问答等价性的微调		78.4	77.1	77.7	71.2	70.0	72.3	42.7	36.6	45.9
基于 VITAMINC 的微调		79.9	71.4	75.4	69.9	68.1	71.1	36.3	32.2	39.9

表5：流水线系统和联合系统在 DIFFQG 上的完整系统性能。请注意，完整系统指标中的“All”部分包含 DIFFQG 的全部内容，而“人工编写”部分仅包含由标注员编辑的问题。流水线系统的问题生成组件使用重新训练的 SQuaD 模型。**粗体** 表示最佳系统，次优结果以下划线标出。R-L=ROUGE-L，Qsim=基于查询相似度模型的准确率，BLRT=BLEURT，FT=微调

8 相关工作

事实编辑 事实变化检测近期受到学界关注。例如，WIKIATOMICEDITS (Faruqui 等人, 2018年)利用维基百科修订来学习区分事实编辑。与我们的工作最接近的是 VITAMINC（Schuster 等人,2021年），其目标是根据一对编辑后的句子生成具有区分性的主张。不过，这两个数据集都主要依赖较小幅度的编辑，通常只是替换单个实体或数字。例如，VITAMINC 示例的词元变更数量中位数为 4，而 WIKIATOMICEDITS 示例的中位数为 2。此外，利用名词或实体重叠等启发式方法，更容易检测出这些编辑。另一方面，DIFFQG 示例的词元变更数量中位数为 13，且可能涉及多个实体更新。此外，即使某个实体同时出现在两个段落中，其周围的上下文信息也可能发生变化。这使得 DIFFQG 中的编辑更难概括，也与以往工作有很大不同；在 第 7.1 节 中也观察到了这一点：使用 VITAMINC 训练数据解决 DIFFQG 时，表现不佳。

近期的工作，例如 Fruit（Iv 等人，2022年）和 PEER（Schick 等人，2023年）也处理更复杂的编辑。Fruit 根据维基百科文章中的新证据，从基础段落生成更新后的句子。PEER 则尝试通过一系列规划步骤来模拟编辑过程。不过，这两种方法主要关注生成目标更新，而我们关注的是简洁地捕捉编辑后的信息。进一步而言，据我们所知，将问题生成用作一种判别手段尚属首创。

问题生成 问题生成已成功应用于多种目的，包括增强问答系统（Duan 等人，2017年；Lewis 等人，2021年）、捕捉关于文本的隐含信息（Pyatkin 等人，2021年），以及构建软知识库（Chen 等人，2022年）。在本研究中，我们将问题生成应用于辨别编辑后的句子这一任务。据我们所知，此前尚无关于评估问题生成系统的研究。

9 结论

在这项工作中，我们提出了 DIFFQG 任务和数据集，用于评估自然语言处理系统通过问题生成来概括两个相关段落之间变化的能力。我们介绍了若干启发式基线和模型基线，以及一组用于衡量数据集表现的评估指标。DIFFQG任务要求模型识别事实关系的变化，并忽略其他风格上的编辑。我们发现，现有方法在这种条件下表现不佳。经过训练、同时执行事实变化检测和问题生成的模型，有时甚至无法理解简单的编辑。我们希望这项工作能为未来对这一重要问题的研究带来价值。

局限性

DIFFQG 的规模相对较小，包含不到一千个问题和不到两千个

示例总数。这使我们无法提供训练集，也限制了我们对任务难度所能做出的主张。此外，总结复杂编辑可能存在大量有效解。虽然根据答案生成问题可以缩小解空间，但仍存在歧义。

为简化标注，我们使用问题生成模型；然而，我们的目标也包括评估问题生成模型，这让我们的叙述变得复杂。最后，我们评估的大多数基线都属于某种形式的 T5（Raffel 等人，2020年）模型。其他模型架构或许能更有效地解决这项任务。

致谢

我们感谢 Jannis Bulian、Tal Schuster 和 William Cohen，以及匿名审稿人提出的深刻意见和宝贵反馈。

参考文献

Samuel R. Bowman、Gabor Angeli、Christopher Potts 和 Christopher D. Manning。2015年。用于学习自然语言推理的大型标注语料库。载于《*2015*年自然语言处理经验方法会议论文集》，第632–642页，葡萄牙里斯本。计算语言学协会。

Jannis Bulian、Christian Buck、Wojciech Gajewski、Benjamin Boerschinger 和 Tal Schuster。2022年。Tomayto, tomahto：超越问答评估中的词元级答案等价性。载于《*2022*年自然语言处理经验方法会议论文集》。

Wenhu Chen、Pat Verga、Michiel de Jong、John Wi- eting 和 William Cohen。2022年。为开放域问答中的预训练语言模型增添问答记忆。

Eunsol Choi、Jennimaria Palomaki、Matthew Lamm、Tom Kwiatkowski、Dipanjan Das 和 Michael Collins。2021年。去语境化：使句子能够独立成篇。计算语言学协会会刊，9:447–461。

Ido Dagan、Bill Dolan、Bernardo Magnini 和 Dan Roth。2010年。识别文本蕴含：基本原理、评估与方法。《自然语言工程期刊》，4。

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

Bhuwan Dhingra、Jeremy R. Cole、Julian Martin Eisenschlos、Daniel Gillick、Jacob Eisenstein 和 William W. Cohen。2022年。时间感知语言模型作为时序知识库。计算语言学协会会刊，10:257–273。

Nan Duan、Duyu Tang、Peng Chen 和 Ming Zhou。2017年。面向问答的问题生成。载于 *2017*年自然语言处理实证方法会议论文集，第866–874 页，丹麦哥本哈根。计算语言学协会。

Manaal Faruqui、Ellie Pavlick、Ian Tenney 和 Dipanjan Das。2018年。WikiAtomicEdits：用于语言和话语建模的维基百科编辑多语言语料库。载于 *2018*年自然语言处理实证方法会议论文集，第 305–315 页，比利时布鲁塞尔。计算语言学协会。

Robert Iv、AlexandrePassos、Sameer Singh 和 Ming-Wei Chang。2022年。FRUIT：忠实反映文本中更新信息。载于计算语言学协会北美分会人类语言技术：*2022*年会议论文集，第 3670–3686 页，美国西雅图。计算语言学协会。

Nikita Kitaev 和 Dan Klein。2018年。使用自注意力编码器进行成分句法分析。载于计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），第2676–2686页，澳大利亚墨尔本。计算语言学协会。

SawanKumar 和 Partha Talukdar。2020年。NILE：借助忠实的自然语言解释进行自然语言推断。载于计算语言学协会第*58*届年会论文集，第8730–8742页，线上。计算语言学协会。

Tom Kwiatkowski、Jennimaria Palomaki、Olivia Redfield、Michael Collins、Ankur Parikh、Chris Alberti、Danielle Epstein、Illia Polosukhin、Jacob Devlin、Kenton Lee、Kristina Toutanova、Llion Jones、Matthew Kelcey、Ming-Wei Chang、Andrew M. Dai、Jakob Uszkoreit、Quoc Le 和 Slav Petrov。2019年。Natural Questions：用于问题回答研究的基准。计算语言学协会会刊，7：452–466。

Patrick Lewis、Yuxiang Wu、Linqing Liu、Pasquale Minervini、Heinrich Küttler、AleksandraPiktus、Pontus Stenetorp 和 Sebastian Riedel。2021年。PAQ：6500万个可能被问到的问题，以及你可以如何利用它们。计算语言学协会会刊，9：1098–1115。

Chin-Yew Lin。2004年。Rouge：一个用于摘要自动评估的软件包。载于文本摘要分支拓展，第74–81页。

Valentina Pyatkin、Paul Roit、Julian Michael、Yoav Goldberg、Reut Tsarfaty 和 Ido Dagan。2021年。事事皆问：为任何语义角色生成语境化问题。载于*2021*年自然语言处理实证方法会议论文集，第1429–1441页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Colin Raffel、Noam Shazeer、AdamRoberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020年。探索统一文本到文本 Transformer 中迁移学习的极限。《机器学习研究杂志》，21(140):1–67。

Pranav Rajpurkar、Robin Jia 和 Percy Liang。2018年。知道自己不知道什么：SQuAD 中无法回答的问题。载于第*56*届计算语言学协会年会论文集（第*2*卷：短文），第784–789页，澳大利亚墨尔本。计算语言学协会。

Pranav Rajpurkar、Jian Zhang、KonstantinLopyrev 和 Percy Liang。2016年。SQuAD：100 个000+ 用于文本机器阅读理解的问题。载于*2016*年自然语言处理实证方法会议论文集，第 2383–2392 页，得克萨斯州奥斯汀。计算语言学协会。

Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。载于*2019*年自然语言处理实证方法会议论文集。计算语言学协会。

Pedro Rodriguez 和 Jordan Boyd-Graber。2021年。问答中的评估范式。载于*2021*年自然语言处理实证方法会议论文集，第 9630–9642 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Timo Schick、Jane A. Yu、Zhengbao Jiang、Fabio Petroni、Patrick Lewis、Gautier Izacard、Qingfei You、Christoforos Nalmpantis、Edouard Grave 和 Sebastian Riedel。2023年。PEER：协作式语言模型。载于国际学习表征会议。

Tal Schuster、Adam Fisch 和 Regina Barzilay。2021年。获取维生素 C！使用对比证据进行稳健的事实核查。载于 计算语言学协会北美分会*2021*年会议：人类语言技术论文集，第 624–643 页，线上。计算语言学协会。

Thibault Sellam、Amy Pu、Hyung Won Chung、Sebastian Gehrmann、Qijun Tan、Markus Freitag、Dipanjan Das 和 AnkurParikh。2020年。学习如何评估——超越英语的翻译评估：BLEURT 在 WMT 2020 指标共享任务中的提交。载于第五届机器翻译会议论文集，第 921–927 页，线上。计算语言学协会。

James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018a。FEVER：用于事实提取和验证的大规模数据集。载于 计算语言学协会北美分会*2018*年会议：人类语言技术论文集，第 *1* 卷（长篇论文），第 809–819 页，美国路易斯安那州新奥尔良。计算语言学协会。

编者：James Thorne、Andreas Vlachos、Oana Cocarascu、Christos Christodoulopoulos 和 Arpit Mittal。2018b。第一届事实提取与验证研讨会（*FEVER*）论文集。计算语言学协会，布鲁塞尔，比利时。

A 附录

A.1 定性示例

DiffQG 数据集中的示例如图 4所示。图中还同时列出了各个系统的模型输出（成功或失败）。

A.2 标注界面

请参阅图 5和图 6，分别查看阶段 1 和阶段 2 的标注界面。

图4：不同系统的 DIFFQG 示例及其预测结果。编辑后的句子采用颜色编码：新增词元标为绿色，删除的词元标为红色；答案片段带下划线。除非为说明所必需，否则省略额外上下文（以灰色显示）。无变化表示该示例没有事实变化。

On the *New Hampshire Executive Council*, Laconia is in the 1st District, represented by **Republican Joe Kenney.** **Democrat Michael J. Cryans.**
Ground Truth: What state council does Joe Kenney represent Laconia in
Overlapping Answer: No change **QA-equivalence**: What state council does Joe Kenney represent Laconia in **Ft on QA-equivalence**: No change
..triggering the Spanish–American War of 1898.. **In the Pacific, Cuba became independent while** the *United States* took possession of **Puerto Rico and Spain's Pacific colonies of** the **Spanish** Philippines and Guam.
Ground Truth: Which country took possession of Puerto Rico after the Spanish-American War of 1898
Overlapping Answer: No change **QA-equivalence**: Which country took possession of Puerto Rico after the Spanish-American War **Ft on QA-equivalence:** What country took possession of Puerto Rico after the Spanish-American
Dances with Wolves is a 1990 epic film which tells the story of a **United States cavalry officer from the** Civil War **era United States Lieutenant** who travels **to alone into** the **American** frontier **near to find** a **military post.** ***Sioux tribe*.**
Ground Truth: In Dances with Wolves, the protagonist travels close to which tribe
Overlapping Answer: Who did the cavalry officer come close to in the story **QA-equivalence**: Who did the cavalry officer come close to in the story **Ft on QA-equivalence**: No change
Additional music, So far Ryu and Chun Li each have one new costume each for $.99 and '*Street Fighter II Arranged BGM*' can also be purchased to provide alternate in game audio.
Ground Truth: No change
Overlapping Answer: No change **QA-equivalence**: What can be purchased to change the in-game audio **Ft on QA-equivalence**: No change

DIFF 100: Susan Boyle (Id: 22378444_902516498_0)

Base	Target
Context: 2018. [SEP] Susan Magdalane Boyle (born 1 April 1961) is a Scottish singer who came to international attention when she appeared as a contestant on the TV programme Britain's Got Talent on 11 April 2009, singing "I Dreamed a Dream" from . [SEP] Her first album, I Dreamed a Dream, was released in November 2009 and became the UK's best-selling debut album of all time, beating the previous record held by Spirit by Leona Lewis.	**Context**: 2019. [SEP] Susan Magdalane Boyle (born 1 April 1961) is a Scottish singer, who rose to fame after appearing as a contestant on the third series of Britain's Got Talent, singing "I Dreamed a Dream" from . [SEP] Her debut studio album, I Dreamed a Dream, was released in November 2009 and became the UK's best-selling debut album of all time, beating the previous record held by Spirit by Leona Lewis.
Sentence: Susan Magdalane Boyle (born 1 April 1961) is a Scottish singer who came to international attention when she appeared as a contestant on the TV programme Britain's Got Talent on 11 April 2009, singing "I Dreamed a Dream" from .	**Sentence**: Susan Magdalane Boyle (born 1 April 1961) is a Scottish singer, who rose to fame after appearing as a contestant on the third series of Britain's Got Talent, singing "I Dreamed a Dream" from .

Susan Magdalane Boyle (born 1 April 1961) is a Scottish singer, who **came** **rose** to **fame after appearing** **international attention when she appeared** as a contestant on the **TV programme** **third series of** Britain's Got Talent, **on 11 April 2009,** singing "I Dreamed a Dream" from .

Diff has new info?			Yes
Question	**Answer (NP)**	**Mark QA**	**Edited Question (if any)**
What is Susan Boyle's real name?	Susan Magdalane Boyle	Edit	
When was Susan Boyle born?	1 April 1961	Reject	
Who is Susan Boyle?	a Scottish singer	Skip	
In which series of Britain's Got Talent did Susan Boyle compete?	the third series	Accept	
What show did Susan Boyle appear on?	Britain's Got Talent	Reject	
What is the name of the song that Susan Boyle sang on Britain's Got Talent?	a Dream	Skip	

图5：标注第一阶段的界面，标注员可从五个选项中选择一个：接受/拒绝/编辑/上下文/跳过。每个示例由两位标注员进行标注。如果两人意见一致，则直接接受该示例，或交由第三位标注人员编辑。如果其中一位标注员选择跳过，则由第三位标注人员做出最终决定。

DIFF 78: Susan Boyle (Id: 22378444_902516498_0)

Base	Target
Context: 2018. [SEP] Susan Magdalane Boyle (born 1 April 1961) is a Scottish singer who came to international attention when she appeared as a contestant on the TV programme Britain's Got Talent on 11 April 2009, singing "I Dreamed a Dream" from . [SEP] Her first album, I Dreamed a Dream, was released in November 2009 and became the UK's best-selling debut album of all time, beating the previous record held by Spirit by Leona Lewis.	**Context:** 2019. [SEP] Susan Magdalane Boyle (born 1 April 1961) is a Scottish singer, who rose to fame after appearing as a contestant on the third series of Britain's Got Talent, singing "I Dreamed a Dream" from . [SEP] Her debut studio album, I Dreamed a Dream, was released in November 2009 and became the UK's best-selling debut album of all time, beating the previous record held by Spirit by Leona Lewis.
Sentence: Susan Magdalane Boyle (born 1 April 1961) is a Scottish singer who came to international attention when she appeared as a contestant on the TV programme Britain's Got Talent on 11 April 2009, singing "I Dreamed a Dream" from .	**Sentence:** Susan Magdalane Boyle (born 1 April 1961) is a Scottish singer, who rose to fame after appearing as a contestant on the third series of Britain's Got Talent, singing "I Dreamed a Dream" from .

Susan Magdalane Boyle (born 1 April 1961) is a Scottish singer, who **came** **rose** to **fame after appearing** **international attention when she appeared** as a contestant on the **TV programme** **third series of** Britain's Got Talent, **on 11 April 2009,** singing "I Dreamed a Dream" from .

Question	**Answer (NP)**	**Phase 1**	**Mark QA**	**Edited Question (if any)**
x	Susan Magdalane Boyle	Edit/Skip	Edit	Which Scottish singer appeared on the third season of Britain's Got Talent?
x	a Scottish singer	Reject/Skip	Reject	
On which television show did Susan Boyle appear?	the third series of Britain's Got Talent	Accept/Skip	Accept	

图6：标注第二阶段的界面，第三位标注人员将改写问题，和/或处理存在分歧的标注。在此，标注员根据编辑 标注，为给定的答案片段编写一个新问题，并决定是否确认另外两个示例的拒绝 和接受 标注。请注意，对于编辑 或拒绝 标注，为避免偏差，我们不会向标注员显示种子问题，而是显示一个x。