结合深度神经重排序与无监督抽取的多查询聚焦摘要

Philipp Seeberger 和 **KorbinianRiedhammer**纽伦堡乔治·西蒙·欧姆应用技术大学 {philipp.seeberger,korbinian.riedhammer}@th-nuernberg.de

arXiv:2302.01148v1 [cs.CL] 2 Feb 2023

摘要

CrisisFACTS 赛道旨在应对事件追踪领域中的多流事实查找等挑战；参赛系统从多个灾害相关事件中提取重要事实，同时纳入时间顺序。我们提出一种结合检索、重排序以及广为人知的整数线性规划（ILP）和最大边际相关性（MMR）框架的方法。在前两个模块中，我们探索了多种方法，包括基于实体的基线、预训练和微调的问答系统，以及 ColBERT。随后，我们将后一个模块用作抽取式摘要组件，并将多样性和新颖性标准纳入考量。自动评分结果显示，系统在各项评估设置中表现出色，但也揭示了不足之处和挑战。

1 引言

如果在紧急事件持续期间，关键信息匮乏导致态势感知不足，那么自然灾害和人为灾害都可能造成重大人员伤亡、财产损失和环境破坏。当今的信息生态系统通过整合各种在线信息源，为应急响应带来了新的机遇和方向（Buntain 等人，2021年；Kruspe 等人，2021）。社交媒体和微博平台等其他信息源能够即时提供有关当前事态发展的细节（Sakaki 等人，2010；Reuter 等人，2018）。由此形成多流信息环境，传统信息源由各种新近兴起的在线信息源加以补充。过去数十年间不断演进的任务表明，先前的研究已认识到这一环境具有广阔前景（Allan 等人，1998；Aslam 等人，2015；Sequiera 等人，2018；Buntain 等人，2021）。

然而，内容生成的高速特性以及不同信息源（Kaufhold，2021年）和事件（Seeberger 和 Riedhammer，2022年）固有的属性，给现有模型带来了新的挑战。这些模型能够向用户提供相关的事件相关结果，但仍不适用于多流事实查找和摘要需求。新设立的 CrisisFACTS 赛道旨在解决这些问题，并促使学界开发更适合随时间进行事实片段抽取的系统。总体而言，该任务要求参与者的系统从危机相关数据集中抽取查询聚焦的事实列表，数据集包括*Twitter*、*Reddit*、网络新闻和*Facebook* 等数据源。每份抽取的事实列表均基于一个事件-日期对，并附有重要性分数，作为后续摘要的依据。事实上，这可以视为信息检索（IR）和摘要领域既有任务的延伸。

近年来，预训练语言模型（如 BERT（Devlinet al.，2019）的引入显著提升了即席排序（Linet al.，2022年）和摘要（Ma et al.，2022年）的效果。具体而言，基于 BERT 的交叉编码器相比经典检索和双编码器方法取得了显著提升，但计算成本高得难以实际应用（KhattabandZaharia，2020年）。为缓解这一问题，深度神经排序模型通常被用作第二阶段重排序器，而第一阶段通常由高效的检索器承担，用于生成一组候选文档（MacAvaney etal.，2022年）。类似地，现代摘要方法依赖交叉编码器检索相关文档、段落或句子，用于抽取式摘要（Xu 和 Lapata，2020年；Ahuja et al.，2022年），或作为生成式摘要的初步筛选步骤（Xu 和 Lapata，2021）。最后，还可以通过最大边际相关性等方法进一步优化排序后的文档集合（Carbonell andGoldstein，1998），

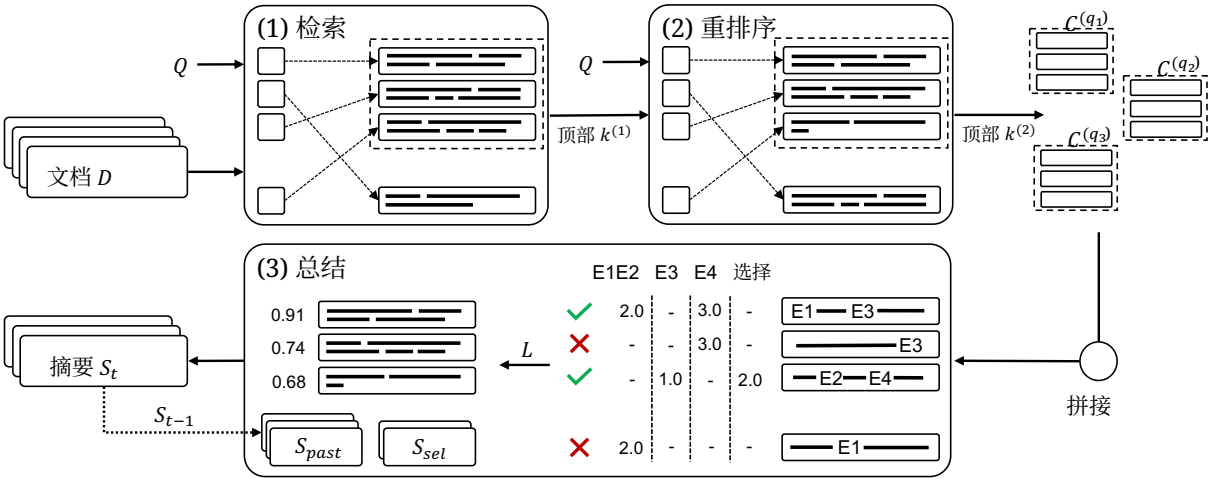

图1：我们提出的框架概览。每个事件和时间段的所有查询和文档均分别经过以下三个主要组件处理：(1) 检索，(2) 重排序，以及 (3) 总结。符号 E1 至 E4 表示整数线性规划公式中的相关概念。在最终评分时，使用选定的 $\mathcal{S}_{sel}$ 和过去的摘要 $\mathcal{S}_{past}$ 文档来惩罚冗余。

整数线性规划（McDonald，2007），以及 TextRank（Mihalcea 和 Tarau，2004）。

在这项工作中，我们探索了多种信息检索和重排序流程，涵盖从预训练到经过微调的最先进模型。此外，我们提出在抽取式摘要设置中，利用广为人知的整数线性规划和最大边际相关性框架相结合的方法，对事实列表进行后续处理。借此，我们旨在解决多样性和冗余相关的问题。

2 方法

如图1所示，我们提出的框架首先根据一项信息请求 $\mathcal{Q} = \{q_1, \ldots, q_M\}$ 检索并重排序一组文档 $\mathcal{D} = \{d_1, \ldots, d_N\}$，其中每个查询 q_i 通常由一段简短文本或一组指示性词语构成。文档数量[1]和查询数量分别记为 N 和 M。令 $\mathcal{C} = \{\mathcal{C}^{(q_1)} \ldots, \mathcal{C}^{(q_M)}\}$ 表示由此得到的查询相关簇集合，其中 $\mathcal{C}^{(q_i)} = \{d_1^{(q_i)}, \ldots, d^{(q_i)}{}_k\}$ 由按相关性排序的前 k 个候选项组成。随后，摘要组件会从簇池 $\bigcup_{i=1}^{M} \mathcal{C}^{(q_i)}$ 中进一步选择 L 个候选项，以生成摘要 $\mathcal{S}$。按照赛道设计，每个摘要 $\mathcal{S}_t$ 均针对一个时间段 $p_t \in \{p_1, \ldots, p_T\}$ 生成，其中 T 表示时间段的数量。在 CrisisFACTS 的范围内，时间段 p_i 对应于一天。下文将详细介绍我们提交方案中使用的各个组件。

[1]在本文中，我们交替使用“文档”一词和 CrisisFACTS 流式条目。这些流式条目更接近句子或简短帖子，而非长篇文档。

2.1 阶段1：检索

在第一阶段，我们采用词汇检索方法，以避免深度神经模型（如基于 BERT 的交叉编码器）带来的高昂计算成本。因此，我们首先为每个查询检索排名靠前的$k^{(1)}$ 候选项 $q \in \mathcal{Q}$，这些候选项来自一组文档 $\mathcal{D}$，检索依据是一系列指示性词语。对于每个查询 q，缩减后的 $k^{(1)}$ 候选项集合随后会交由指定的重排序阶段处理。在检索过程中，我们采用广为人知的 BM25 模型（Robertson 和 Zaragoza，2009），但也可以轻松地用更先进的方法替代。我们在初步实验中发现，候选项数量相对较少，限制了后续重排序组件的表现。为解决这一问题，可以采用查询扩展（Amati 和 Van Rijsbergen，2002）、文档扩展（Nogueira 等人，2019）或自适应重排序（MacAvaney 等人，2022年）方法。本文中，我们整合了查询扩展，以扩大候选项池，克服召回率限制。

2.2 阶段 2：重排序

经典检索方法可能不足以捕捉查询中表达的语义及其与文本内容的关系。因此，对前一检索阶段的每个查询相关簇 $\mathcal{C}^{(q)}$ 进行重排序，并选出排名靠前的 $k^{(2)}$ 个候选项，

从而得到一个新的集合 $\mathcal{C}'^{(q)}$。在此，我们利用预训练深度神经模型，从其他现有数据集中获取监督信号。此外，我们还纳入一个基于实体和关键词的基线，以评估在大型文本语料库上预训练的模型所带来的性能增益。下面将详细介绍所考虑的重排序模型：

BoE 作为基线，我们采用 Xiong 等人 (2016年) 提出的用于文档排序的实体袋表示，该方法依赖于实体链接系统所获取的语义信息。我们通过根据实体类型为每篇文档构建实体袋向量，并使用关键词袋向量对其进行扩展，实现了一个简化版本。最终模型通过累加文档中出现的预期查询实体类型和关键词的频率来为文档评分。

问答 与 Xu 和 Lapata (2020年) 类似，我们采用问答（QA）系统，以利用与最佳答案选择相关的远程监督信号。问答方法支持句子和片段选择，但我们仅采用句子级选择，这更适合组织者提供的查询和流式条目。具体而言，我们将查询 q 与候选文档 d 拼接成一个序列 [CLS] q [SEP] d，并使用基于 BERT 的交叉编码器预测相关性得分。在本研究中，我们采用预训练并经过微调的问答版本，其中微调后的系统适配于危机领域。

ColBERT 该模型采用上下文化后期交互方法，同时利用第一阶段的近似最近邻搜索和重排序阶段来计算排序分数（Khat- tab 和 Zaharia，2020年）。具体而言，ColBERT 支持重排序机制，可基于候选项池生成更精确的分数，也可用于端到端检索。我们采用端到端检索方法。这样，我们考虑了一种不受 BM25 等经典检索模型相关限制的方法。请注意，这种替代方案跳过了第一个检索阶段，直接检索文档集合 $\mathcal{D}$，以获取每个查询的前 $k^{(2)}$ 个候选项 q。

2.3 阶段3：总结

选择 在摘要任务中，找到一组多样且冗余较少的事实至关重要。然而，如果不进行任何后处理，重排序后的候选项在多样性和冗余方面仍然存在问题——冗余。为解决这一问题，我们使用一个额外的选择步骤，并将其形式化为整数线性规划。我们采用基于概念的模型（Gillick 和 Favre，2009；Riedhammer 等人，2010），其中概念可以是事实、事件或信息单元。在这一问题设定中，目标函数是在长度约束下，最大化所选内容中各概念的加权和。最后，我们得到一个抽取式摘要 $\mathcal{S}_t = \{d_1, \ldots, d_l\}$其中 $|\mathcal{S}_t|$ 受限于 $l \leq L$ 且 L 为文档的最大数量。

评分 广为人知的最大边际相关性算法会贪心地选择文档，在基于查询的相关性与相对于此前所选文档的冗余之间进行权衡，直至满足摘要长度限制。然而，为了提高排名靠前文档的多样性，可以放宽这一限制，对摘要 $\mathcal{S}_t$ 进行重排序。形式上，我们将文档 i 的最终分数定义为

$$\lambda \cdot \text{Rel}_i - (1-\lambda) \cdot \max_{j \in \mathcal{S}_{sel} \cup \mathcal{S}_{past}} \text{Red}_{ij} \quad (1)$$

其中，Rel_i 是文档 i 的相关性得分，Red_{ij} 是同时将文档 i 和 j 纳入摘要 $\mathcal{S}_{sel}$ 以及以往摘要 $\mathcal{S}_{past} = \bigcup_{i=1}^{t-1} \mathcal{S}_i$所产生的冗余惩罚。然而，由于存在多个匹配的查询，单个检索到的文档可能对应多个得分。我们认为，涵盖多个查询的文档能为摘要提供更多相关信息。具体而言，我们将相关性得分表示为 $\text{Rel}_i = |\mathcal{Q}^{(i)}| \cdot score_i$，其中 $score_i$ 是文档 i 的平均得分，并根据匹配的查询数量 $\mathcal{Q}^{(i)} \subseteq \mathcal{Q}$进行加权。

3 实验

本节详细介绍实验设置，并讨论我们提交的各组实验结果。在所有实验中，我们主要考虑来源*Twitter*、*Reddit*和网络新闻，由于难以获取帖子内容，因此不考虑*Facebook*。

3.1 预处理

我们对所有推文进行规范化，使其文本内容的表示方式与其他在线信息源相似。具体而言，我们会移除所有转推标识前缀、用户提及、颜文字、表情符号和网址。此外，我们还会移除所有井号符号，并使用*WordSegment*将文本拆分为相应的词语。[2]

[2] https://grantjenks.com/docs/WordSegment

	摘要			匹配	
	ICS	NIST	Wiki	全面性	冗余
ColBERT	.050/**.450**	.139/.546	.031/.542	.189	.201
BM25 →BoE	.047/.436	.142/.560	.030/.533	.185	**.176**
BM25 →问答$_{\text{ASNQ}}$	**.051**/.448	**.147**/.563	**.036/.565**	**.213**	.226
BM25 →问答$_{\text{Crisis}}$	.046/.443	**.147/.564**	.034/.545	.210	.226
TREC 最佳	.058/.459	.147/.564	.036/.565	.217	.125

表 1：我们自动提交运行的总体结果。我们报告摘要的 Rouge-2/BERTScore，以及匹配的全面性和冗余率，定义见附录 A。我们提出的系统中表现最好的结果以粗体标出。

事件	ICS	NIST	Wiki
001	.116*/.522	**.273**/.560*	.013*/.540
002	**.066**/.561	.050*/.563	.043/**.579**
003	.053/.516	.238/.611*	.021/.593*
004	.061/.480	.171*/ **.585**	**.060/.582**
005	-	.136/.544*	.032/.526*
006	.057/.506	.048/.533	.019*/.580
007	.040/.494*	.104*/.524*	**.057/.554**
008	.012/.501	**.154**/.583	**.044/.562**

表 2：针对问答$_{\text{ASNQ}}$ 系统的各个事件的 Rouge-2/BERTScore 结果。所有提交中 TREC 的最佳结果以粗体显示。低于 TREC 中位数的结果以灰色显示，而低于我们的基线 BoE 的结果则以 *标记。

3.2 危机问答

由于首届 CrisisFACTS 赛道未提供与任务相关的任何标注，我们决定创建一个反映查询聚焦句子选择的合成版本。我们利用 DocEE 数据集（Tong 等人，2022年），这是一个近期发布的文档级事件抽取基准数据集。我们抽取了其中一个包含 6818 篇文档的子集，该子集仅涵盖与危机相关的事件及其对应的事件论元。[3] 首先，我们为每个事件论元手动创建粗粒度问题。其次，使用一个 T5$_{\text{BASE}}$ 问题生成模型[4] 对数据集进行增强，以获得细粒度问题。最后，我们根据论元的位置合成问题-句子对，并将这些配对标记为二元相关性分类任务。模型验证时，我们使用已发布的数据集划分。

[3]我们检查了 DocEE 与 Cri- sisFACTS 事件之间是否存在重叠。事实上，其中一些事件属于 DocEE 数据集，因此我们在实验开始前移除了相应的文档。[4]

https://huggingface.co/mrm8488/t5-base-finetuned-question-generation-ap

3.3 实验设置

检索 在第一阶段，我们使用 PyTerrier 库的默认设置运行 BM25 模型（Macdonald 和 Tonellotto，2020年），并将其与 Bo1 查询扩展（Amati 和 Van Ri- jsbergen，2002）组件结合。对于每个查询，我们将查询文本与指示性词语拼接起来，检索排名靠前的 $k^{(1)} = 100$ 候选项，并删除完全重复的项。大多数重复项出现在推文文档中，且主要与转发推文有关。

重排序 BoE 模型基于一组经过人工整理的实体类型，这些类型大多符合每个查询的预期信息需求。例如，关于失踪人员的查询通常分别涉及人数和地点。组织者提供的指示性词语被用作关键词。问答$_{\text{ASNQ}}$ 系统基于 RoBERTa$_{\text{BASE}}$，并在 ASNQ 数据集上进行预训练（Garg 等人，2020年），且未做任何进一步调整。类似地，我们采用 ColBERT$_{\text{v2}}$版本（Santhanam 等人，2022年），该版本在 MS MARCO 段落排序任务上训练。在问答$_{\text{Crisis}}$方面，我们遵循Garg 等人（2020年）所采用的适配步骤，使用领域专属的危机问答数据集对问答$_{\text{ASNQ}}$模型进行微调，从而得到适配后的问答系统。尽管合成数据集涵盖了范围广泛的已标注事件论元，我们仍发现问句—句子对中存在相当比例的假阴性。因此，我们首先使用问答模型$_{\text{Crisis-0}}$ 对数据集进行去噪，阈值上限设为 0。1，然后在第二步中训练新的问答模型$_{\text{Crisis-1}}$，这一做法与 RocketQA（Qu 等人，2021）等先前工作一致。我们使用 Transformers 库（Wolf 等人，2020年）来构建问答模型，使用 ColBERT 官方实现，并

选择每个查询的前 $k^{(2)} = 25$ 个候选项。

总结 为了公平比较不同的检索和重排序组件，我们在每次运行中重复使用相同的选择和评分流程。具体而言，受信息抽取（Martinez-Rodriguez 等人，2020年）的启发，我们提取实体[5] 作为概念，将实体频率作为权重，并将 $L = 150$ 用于整数线性规划公式。对于最大边际相关性，我们选择 $\lambda = 0.8$ 并根据 TF-IDF 特征和余弦相似度计算冗余度 Red_{ij}。

3.4 结果

在表1中，我们展示了各流程设置的整体表现。由于这是 CrisisFACTS 赛道的首届赛事，我们主要将分析范围限定在各提交运行结果之间。不过，我们也将模型与摘要任务的中位数和最佳结果进行比较，供读者参考（表2）。

总体而言， 问答模型在几乎所有评估设置中都优于基线 BoE 和 ColBERT。这些结果印证了以往文本检索研究的发现：交叉编码器架构的表现更佳。有趣的是，经过微调的问答模型在两种摘要设置下以及全面性方面的表现有所下降。我们推测，经过适配的问答模型对危机问答数据集中的实体存在偏向。这可能导致它仅对部分事实给出较高分数。此外，我们也注意到，由于 CrisisFACTS 与危机问答事件的时间重合，可能存在数据重叠问题。不过，性能提升仅出现在 NIST 参考摘要上，因此我们将相关分析留待未来研究。

摘要 表2中的深入分析发现，预训练问答模型在各类事件和参考摘要上都取得了最佳结果。与 BoE 基线相比，性能提升在不同事件、指标和参考摘要之间有所差异。不过，只有三项性能指标低于 TREC 中位数，这表明整个流程表现出色。尽管如此，与自动摘要评估不同，人工匹配显

[5] https://stanfordnlp.github.io/stanza/ ner.html

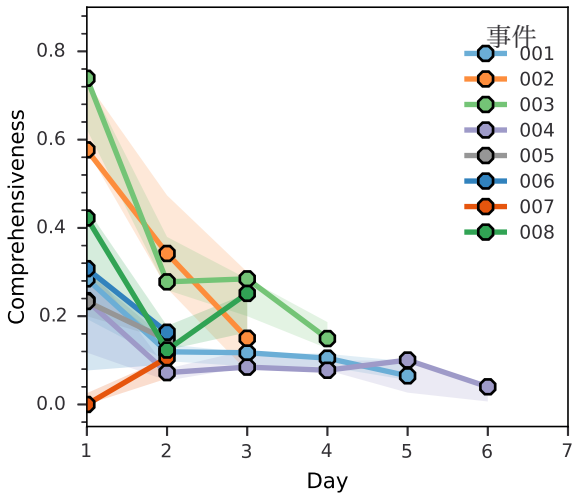

图 2：所有事件的全面性趋势。问答ASNQ 系统以粗体显示，而所有模型的最小值与最大值范围则以高亮显示。

示，同一事件在不同日期的结果存在较大差异。

匹配 如果我们绘制全面性随天数变化的趋势（图 2），可以看到，各类事件的性能都大幅下降。由于所有模型都呈现这一趋势，我们推测这至少由两个因素造成。首先，该流程中的检索和重排序阶段没有考虑每个查询的多样性，因此可能会舍弃罕见事实，转而保留相关性更高、分布在事件时间线各处的事实。其次，在选择阶段，针对以往摘要进行多样化仍然是一项具有挑战性的任务。例如，一些具体句子可能仅相差一个数字（如烧毁面积），而不够成熟的相似度度量可能会无意中惩罚新事实。

4 结论

在这项工作中，我们研究了深度神经重排序与全局无监督抽取相结合的方法，用于多查询聚焦摘要任务。实验表明，基于远程监督的问答的交叉编码器表现出色。然而，我们发现，在处理时间方面时仍存在不足和挑战，这凸显了下游摘要作为关键环节的重要性。我们认为仍有很大的改进空间，尤其是通过整合更复杂的抽取式方法、抽象式摘要技术，甚至采用联合优化。

致谢

作者感谢德国联邦教育与研究部通过 ISAKI 项目（项目编号 13N15572）提供的资金支持。

参考文献

Ojas Ahuja、Jiacheng Xu、Akshay Gupta、Kevin Horecka 和 Greg Durrett。2022年。ASPECTNEWS：新闻文档的面向方面的摘要。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6494–6506页，爱尔兰都柏林。ACL。

James Allan、Jaime G. Carbonell、George Doddington、Jonathan Yamron 和 Yiming Yang。1998年。主题检测与跟踪试点研究最终报告。出版方：卡内基梅隆大学。

Gianni Amati 和 Cornelis Joost Van Rijsbergen。2002年。基于测量与随机性的偏差的信息检索概率模型。*ACM*信息系统汇刊，20(4):357–389。

Javed A. Aslam、Fernando Diaz、Matthew Ekstrand-Abueg、Richard McCreadie、Virgil Pavlu 和 Tetsuya Sakai。2015年。TREC 2015时间摘要赛道概述。载于第*24*届文本检索会议论文集，*TREC 2015*，美国马里兰州盖瑟斯堡，*2015*年*11*月*17*日至*20*日，第500-319卷，*NIST*特别出版物。美国国家标准与技术研究院（NIST）。

Cody L. Buntain、Richard McCreadie 和 Ian Soboroff。2021年。事件流 2020年：COVID-19 时期的 TREC-IS。载于*ISCRAM 2021*：第 *18* 届国际危机响应与管理信息系统会议。

Jaime Carbonell 和 Jade Goldstein。1998年。MMR 的应用：基于多样性的重排序，用于重新排序文档并生成摘要。载于第 *21* 届 *ACM SIGIR* 信息检索研究与开发年度国际会议论文集——*SIGIR '98*，第 335–336 页，澳大利亚墨尔本。ACM。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向 Transformer 预训练。载于 *2019* 年计算语言学协会北美分会人类语言技术会议论文集，第 4171–4186 页，美国明尼苏达州明尼阿波利斯。ACL。

SiddhantGarg, ThuyVu 和 AlessandroMoschitti。2020年。TANDA：迁移并调整预训练用于答案句选择的 Transformer 模型。*AAAI* 人工智能会议论文集，34(05):7780–7788。

Dan Gillick 和 Benoit Favre。2009年。摘要全局模型：一种可扩展的摘要全局模型。载于自然语言处理整数线性规划研讨会论文集，ILP '09，第 10–18 页，美国。ACL。会议地点：博尔德，科罗拉多州。Marc-André Kaufhold。2021年。危机信息学的信息提炼技术：社交媒体和移动应用的用户期望与设计原则。Springer Fachmedien Wiesbaden，威斯巴登。

Omar Khattab 和 Matei Zaharia。2020年。ColBERT：通过 BERT 上下文化的后期交互实现高效且有效的段落检索。发表于第 *43* 届 *ACM SIGIR* 信息检索研究与开发国际会议论文集，第 39–48 页，中国线上会议。ACM。

A. Kruspe、J. Kersten 和 F. Klan。2021年。综述文章：危机事件中可采取行动推文的检测。《自然灾害与地球系统科学》，21(6)：1825–1845。

Jimmy Lin、Rodrigo Nogueira 和 Andrew Yates。2022年。用于文本排序的预训练*Transformer*：*BERT* 及其后续发展。《人类语言技术综合讲座》。Springer International Publishing，Cham。

Congbo Ma、Wei Emma Zhang、Mingyu Guo、Hu Wang 和 Quan Z. Sheng。2022年。基于深度学习技术的多文档摘要：综述。《*ACM* 计算机调查》，55(5)。

Sean MacAvaney、Nicola Tonellotto 和 Craig Mac- donald。2022年。借助语料库图进行自适应重排序。载于第*31*届 *ACM* 国际信息与知识管理会议论文集，CIKM '22，第1491–1500页，纽约州纽约，美国。ACM。

Craig Macdonald 和 Nicola Tonellotto。2020年。使用 PyTerrier 在信息检索中进行声明式实验。载于*2020*年 *ACM SIGIR* 国际信息检索理论会议论文集，ICTIR '20，第161–168页，美国纽约州纽约。ACM。会议地点：线上会议，挪威。

Jose L. Martinez-Rodriguez、Aidan Hogan 和 Ivan Lopez-Arevalo。2020年。信息抽取与语义网：一项综述。语义网，11(2)：255–335。出版商：IOSPress。

Ryan McDonald. 2007.《多文档摘要中的全局推理算法研究》。载于第*29*届欧洲信息检索研究会议论文集，ECIR'07，第557–564页，柏林，海德堡。Springer-Verlag。事件地点：意大利罗马。

Rada Mihalcea 和 Paul Tarau。2004。TextRank：为文本带来秩序。载于*2004*年自然语言处理实证方法会议论文集，第404–411页，西班牙巴塞罗那。ACL。

Rodrigo Nogueira、Wei Yang、Jimmy Lin 和 Kyunghyun Cho。2019 年。基于查询预测的文档扩展。出版方：arXiv 版本编号：2。

Yingqi Qu、Yuchen Ding、Jing Liu、Kai Liu、Ruiyang Ren、Wayne Xin Zhao、Daxiang Dong、Hua Wu 和 Haifeng Wang。2021 年。RocketQA：一种针对开放域问答的密集段落检索优化训练方法。载于*2021* 年北美计算语言学协会分会年会：人类语言技术论文集，第 5835–5847 页，线上。ACL。

Christian Reuter、Amanda Lee Hughes 和 Marc-André Kaufhold。2018 年。危机管理中的社交媒体：对危机信息学研究的评估与分析。国际人机交互期刊，34(4):280–294。

Korbinian Riedhammer、Benoit Favre 和 Dilek Hakkani-Tür。2010年。长话短说——基于关键词的会议摘要全局无监督模型。《语音通信》，52(10)：801–815。

Stephen Robertson 和 Hugo Zaragoza。2009年。概率相关性框架：BM25 及其扩展。《信息检索基础与趋势》，3(4)：333–389。出版地：美国马萨诸塞州汉诺威；出版商：Now Publishers Inc.

Takeshi Sakaki、Makoto Okazaki 和 Yutaka Matsuo。2010年。地震震动 Twitter 用户：通过社会传感器实时检测事件。载于第*19*届万维网国际会议论文集——*WWW '10*，第851页，美国北卡罗来纳州罗利。ACM。

Keshav Santhanam、Omar Khattab、Jon Saad-Falcon、Christopher Potts 和 Matei Zaharia。2022年。ColBERTv2：通过轻量级晚期交互实现高效检索。载于*2022*年计算语言学协会北美分会会议论文集：人类语言技术，第3715–3734页，美国西雅图。ACL。

Philipp Seeberger 和 Korbinian Riedhammer。2020年。运用实体掩码语言建模和多任务学习提升危机相关推文分类。出版方：arXiv 版本号：1。

Royal Sequiera、Luchen Tan 和 Jimmy Lin。2018年。TREC 2018 实时摘要赛道概述。载于第二十七届文本检索会议论文集，*TREC 2018*，美国马里兰州盖瑟斯堡，*2018*年*11*月*14*日至*16*日，第500-331卷，收录于*NIST*特别出版物。美国国家标准与技术研究院（NIST）。

MeiHan Tong、Bin Xu、Shuai Wang、Meihuan Han、Yixin Cao、Jiangqi Zhu、Siyu Chen、Lei Hou 和 Juanzi Li。2022年。DocEE：面向文档级事件抽取的大规模细粒度基准。载于*2022*年计算语言学协会北美分会会议论文集：人类语言技术，第3970–3982页，美国西雅图。ACL。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Remi Louf、Morgan Funtow- icz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020年。Transformers：自然语言处理领域的先进技术。载于*2020*年自然语言处理实证方法会议论文集：系统演示，第 38–45 页，线上。ACL。

Chenyan Xiong、Jamie Callan 和 Tie-Yan Liu。2016年。用于排序的实体袋表示。载于*2016*年*ACM*信息检索理论国际会议论文集，第181–184页，美国特拉华州纽瓦克。ACM。

Yumo Xu 和 Mirella Lapata。2020年。由粗到细的查询聚焦多文档摘要。载于*2020*年自然语言处理实证方法会议论文集，第3632–3645页，线上。ACL。

Yumo Xu 和 Mirella Lapata。2021年。从无查询资源生成查询聚焦摘要。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集，第6096–6109页，线上。ACL。

匹配指标

系统与事件日对的流式条目按照重要性分数排序，并通过排序截断值 k形成摘要 $\mathcal{S}$。基于将流式条目与事实列表 F进行人工匹配，全面性计算如下

$$\frac{1}{\sum_{f\in F} R(f)} \sum_{\{f\in F: M(f,\mathcal{S})\neq\emptyset\}} R(f) \qquad (2)$$

其中，$f \in F$是唯一事实，$M(f, \mathcal{S})$ 是与事实 f 匹配的流式条目集合，$R(f)$ 是分配给事实 f 的增益。同样，系统和事件-日期对的冗余率测量如下：

$$\frac{\sum_{\{f\in F: M(f,\mathcal{S})\neq\emptyset\}} R(f)}{\sum_{\{f\in F\}} R(f)\cdot|M(f,\mathcal{S})|} \qquad (3)$$

所有运行先按事件内的日期进行宏平均，然后再对八个事件进行宏平均。