arXiv:2301.06876v1 [cs.MM] 17 Jan 2023

CS-lol：电子竞技直播中的观众评论与场景数据集

Junjie H. Xu jhxu@acm.org 综合人类科学研究生院 筑波大学 日本茨城筑波

Yu Nakano s2030514@s.tsukuba.ac.jp 综合人类科学研究生院 筑波大学 日本茨城筑波

Lingrong Kong kong@star.rcast.u-to k yo.ac.jp 东京大学先进科学技术研究中心 日本东京

Kojiro Iizuka kojiro.iizuka@gunosy.com Gunosy公司 日本东京

摘要

数十亿直播观众会实时分享他们对正在观看的场景的看法，并通过文本评论与赛事、解说员及其他观众互动。因此，有必要探索电子竞技直播活动中与场景相关的观众评论。本文开发了*CS-lol*，这是一个新的大规模数据集，其中包含电子竞技直播中观众评论及配对的游戏场景描述。此外，我们提出了一项名为观众评论检索的任务，旨在检索与直播赛事场景对应的观众评论。基于典型信息检索评估方法构建的一系列基线检索方法的实验结果表明，这项任务颇具挑战性。最后，我们向研究社区发布*CS-lol*及基线实现，供其使用。

CCS 概念

• **以人为中心的计算** → **人机交互**； • **信息系统** → **信息检索**。

关键词

数据集、带场景的观众评论、电子竞技直播

ACM 引用格式：

Junjie H. Xu、Yu Nakano、Lingrong Kong 和 Kojiro Iizuka。2023年。CS-lol：电子竞技直播中包含场景的观众评论数据集。载于*ACM SIGIR*人类信息交互与检索会议（*CHIIR '23*），*2023*年*3*月*19*日至*23*日，美国得克萨斯州奥斯汀。ACM，美国纽约州纽约市，共5页。https://doi.org/10.1145/3576840.3578334

1 引言

随着电子竞技日益普及， [6, 21],电子竞技直播已发展成为一个市场前景可观的行业，预计将创造13.8亿美元的收入，并吸引5.32亿观众。 2022 [10, 15]了解电子竞技观众，例如他们的观点和偏好，是电子竞技直播中的一个重要议题。图1展示了电子竞技直播的一个示例。与传统的电视直播不同，电子竞技观众不仅可以观看赛事，还能通过直播平台上的互动功能，分享自己对正在观看的赛事场景的看法，并与其他观众互动，例如在*Twitch*[1] 上发送文字聊天消息，或在斗鱼[2]上发送弹幕。观众评论是了解电子竞技观众的重要资源，因为它不仅是反映真实情况的反馈数据，而且可以在不打扰观众观看体验的情况下收集。



图1：电子竞技直播中观众评论和场景描述示例。

然而，正如Ringer等人 [18], 所指出的，直播中的观众评论是一个颇具挑战性的领域，其性质和主题内容使其有别于其他领域的文本。此外，要理解观众评论的含义，必须考虑其与场景的关联，因为观众是在观看时发表评论的。尽管近期对直播数据的研究关注显著增加 [1, 11, 20] ，但由于可用资源匮乏，对直播观众评论的深入研究仍然鲜见。本文

允许免费制作本作品部分或全部内容的数字或纸质副本，供个人或课堂使用，前提是副本不得用于或分发以获取利润或商业利益，且副本须附有此声明及首页的完整引文。必须遵守本作品中第三方组件的版权规定。

如需用于其他用途，请联系所有者/作者。*CHIIR '23*，*2023*年*3*月*19*日至*23*日，美国得克萨斯州奥斯汀

©2023年 版权所有者/作者所有。ACM ISBN 979-8-4007-0035-4/23/03。

https://doi.org/10.1145/3576840.3578334

[1]https://www.twitch.tv/[2] https://www.douyu.com/

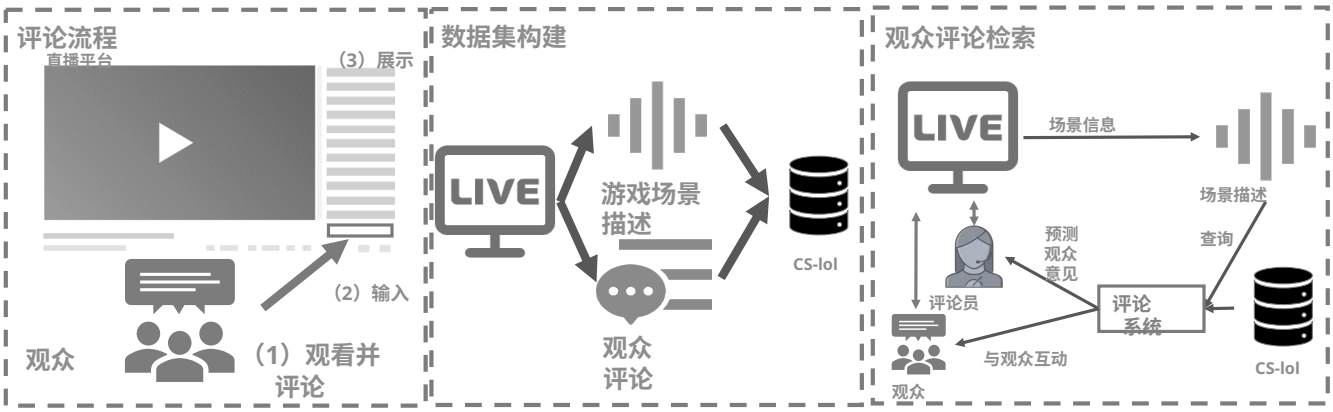


图2：直播中的评论流程、*CS-lol* 构建与观众评论检索概览。

进一步通过介绍与观众评论配对的游戏场景描述，推动对观众评论的理解与利用研究。

总而言之，本文介绍了*CS-lol*[3]，这是一个新颖的数据集，包含大规模的观众**评**论以及**场**景描述，涵盖一款电子竞技直播游戏英雄联盟之战，旨在推动未来有关观众评论理解的研究。此外，本文提出了观众评论检索任务：针对电子竞技直播中的某个游戏场景，从所有评论中检索观众评论（实际应用时则从观众评论数据库中检索），以理解观众。最后，本文介绍了*CS-lol* 的特点与可用性，并报告了通过实验得出的一系列基线信息检索方法的结果。

2 相关工作

为了提升直播观众体验，学术界和业界人士对研究电子竞技视频和直播的兴趣日益浓厚 [17]。因此，随着自然语言处理领域近期进展不断涌现，以及理解文本数据的需求日益增长，针对电子竞技游戏文本数据及游戏场景描述的研究 [5, 22, 25, 27, 30], ，或针对游戏数据流的分析研究 [1, 12, 24, 26],, 都在稳步增加。

已有大量关于直播观众的研究或文献，通过案例研究或采用设计因素的对比研究，并使用显著性检验来验证其观点 [8, 9]。然而，这些研究大多基于个案，或使用来源各异的非公开数据，缺乏可复现性，且难以将研究结果推广至实际应用 [23]。电子竞技游戏场景与观众对该场景电子竞技直播的偏好之间的关系仍不明确。

据我们所知，与我们的数据集最接近的是 Ringer等人开发的 *TwitchChat*。他们首次推出了一个包含*Twitch* 观众评论的数据集，并对其进行了分析，且该数据集可公开获取。如前所述，由于观众评论是在

[3]https://github.com/junj2ejj/CS-lol

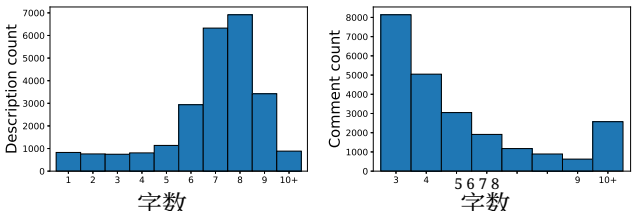


图3：场景信息（左）和观众评论（右）的字数分布。

场景发生时发布的，因此场景信息对于理解观众评论也十分必要，尤其是在电子竞技直播中。我们的工作通过将评论与场景关联起来，进一步探索观众评论。此外，我们还提出观众评论检索任务，以展示直播观众评论的可用性，并探讨未来可能的研究方向。

3 数据集构建与分析

本节介绍*CS-lol*的开发过程及其特征。我们已公开发布*CS-lol* 数据集。*CS-lol* 数据集包含三类数据：

- 场景信息：由集成在*YouTube*中的自动语音识别（ASR）系统，从游戏解说员生成的游戏场景描述（描述性文本）。
- 观众评论：视频标题以及观众发表的评论。我们将观众信息匿名化，以防止隐私问题。
- 相关性得分：评论、游戏场景信息与观众评论之间的相关性得分。

3.1 数据收集与预处理

图 2 展示了电子竞技直播中观众评论的产生流程、同时包含游戏场景信息及相应观众评论的数据集构建过程，以及该数据集在直播中的实际应用情况。

我们收集了名为 英雄联盟 的电子游戏中的电子竞技职业赛事数据。根据最新研究 [15]，这款游戏是全球最受欢迎的游戏，拥有数十亿观众。为此，我们从 英雄联盟电子竞技 职业赛事视频中的 20 场比赛提取了训练有素的专业解说员所提供的评论，并将其用于构建数据集。数据集中的数据来自两个平台：场景信息（描述）来自 *YouTube*，观众评论则来自 *Twitch*。对于提取的 20 场比赛，我们分别从 *YouTube* 和 *Twitch* 中找出同一场电子竞技比赛的回放视频，并手动对齐这两个视频。为了从 *YouTube* 收集游戏场景信息，我们使用*YouTube* API [4] 获取解说员对电子竞技比赛所作描述的文字稿及其时间戳。为了收集观众评论，我们手动搜索 *Twitch* 直播回放中的片段，确保其内容与上述 *YouTube* 片段完全一致。随后，我们截取比赛从开始到结束的相应内容，以确保不同来源的视频片段内容完全匹配且时长相同。此外，*Twitch* 上有相当一部分评论（见表 2）存在以下情况：（1）极其简短，和/或（2）仅包含表情符号（*Twitch*中的特殊表情），而这些表情并未出现在描述中。这两类评论所含信息很少，因此我们删除了指代 *Twitch*[5] 的表情符号，以及少于三个词的评论。

3.2 相关性得分标注

我们通过对齐来自两个不同来源的内容，将从*YouTube*提取的场景信息与从*Twitch*提取的观众评论结合起来。根据有关在 *Twitch* 平台上观看电子竞技游戏直播的文献 [2, 3]，，当观众在电子竞技直播期间尝试发表评论时，必须考虑（1）观众输入评论所花费的时间。根据一项用户研究 [16]，，键盘上的平均打字速度 WPM_k 为每分钟 52 词，而 WPM_s 在智能手机上为每分钟 38 词，因此我们将 WPM_k 和 WPM_s 的平均值作为观众打字速度。对于每场比赛，我们有 i 条游戏场景描述和 j 条游戏场景评论，用于标注每条评论和描述。给定一条包含n_w个词的评论的时间戳 T_c，以及一条描述的时间戳 T_d，（2）观众无法预测直播中的未来场景。因此，相关性得分 *Per* 按以下公式计算：

$$Per(d_i, c_j) = \begin{cases} -(T_c - T_d - \frac{0.5*(WPM_k+WPM_s)*n_w}{60}) & T_c >= T_d \\ -\infty & T_c < T_d \end{cases} \quad (1)$$

其中，d_i 从 *CS-lol* 的所有对局中捐赠第 i 条描述，c_j 捐赠第 j 条评论。对于每个游戏场景，我们根据 *Per* 对评论进行排序，并附上 *Rel* 最高的前 20 条评论。因此，我们得到 *CS-lol*：一个由代表电子竞技直播游戏场景的电子竞技描述、与场景对应的观众评论，以及游戏场景信息与观众评论之间的相关性得分组成的数据集。统计数据见表 2，*CS-lol* 的描述和观众评论的字数分布如图 3 所示。

[4]https://developers.google.cn/youtube/v3/getting-started [5] https://www.twitch.tv/creatorcamp/en/learn-the-basics/emotes/

4 观众评论检索

4.1 任务定义

本文聚焦于电子竞技直播中，根据游戏事件场景检索观众评论的问题。本文所研究的问题定义如下：给定游戏场景描述，检索评论日志中的观众评论，并按照其与该描述的相关性进行排序。具体而言，输入由一个描述 q和一组评论 D组成，这些评论取自所有比赛。该问题的输出是一个排序后的评论列表，即 $\mathbf{r} = (r_1, r_2, ..., r_{|D|})$，其中 $r_{\pi(i)} = d_i$和 $\pi(i)$表示 D中第 i条评论的排名。

4.2 实验设置

我们采用 BM25 [19], 带 Dirichlet 平滑的查询似然（QLD） [31], 带 Jelinek-Mercer 平滑的查询似然（QLJM） [31], 和序列依赖模型（SDM） [14] 作为基线。我们使用 Anserini [29]实现了这些模型。对于这些模型的参数，我们采用 Anserini 的默认设置。

如 Huang 等人 [4],所述，从 *Youtube* 提取的自动生成描述并未按照评论员口述的句子原本进行标点处理。通常，仅使用一行自动生成的描述无法将完整的评论员话语纳入查询。因此，在检索 q_t 的观众评论时，t 指评论员给出描述的时间，q_{t-1} 和 q_{t-2} 也根据经验一并纳入。此外，对于 *Twitch* 评论，我们在 Ringer等人 [18] 的基础上，对观众评论应用基于规则的词语清理，以修正评论中的词语扩展和拼写错误。在检索之前，还分别对描述和观众评论执行了词干提取和停用词移除。

4.3 实验结果

表1展示了采用标准评估指标对词汇匹配进行评估的结果，包括精确率（P）、召回率（R）和 F1 分数（F1）。表3展示了基线检索方法在 nDCG@3、5、10 和 20 指标上的结果。粗体表示每项指标的最高分。根据表1，QLDM 在词语匹配方面表现更好。在采用 nDCG 指标评估的排序结果中，SDM 优于其他模型。

根据表3，基于词汇匹配的基线方法表现不佳。我们发现，BM25 并不像“通常”那样是最佳模型。我们认为，这可能是由我们任务的特殊性所致：查询相对较长，而文档则高度精简。因此，新提出的检索任务可能是“非传统”的。

5 未来方向

在本节中，我们进一步探讨 *CS-lol* 的可用性，并介绍其在信息检索和自然语言处理领域中的未来用途。

表1：检索方法中基于词匹配的实验结果。

	Top@3			Top@5			Top@10			Top@20		
模型	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BM25 [19]	3.14	4.14	3.26	**4.76**	6.11	4.98	**7.69**	9.59	8.10	**11.35**	14.00	12.11
QLD [31]	2.85	3.96	3.03	4.33	5.81	4.64	7.07	9.30	7.65	10.67	14.03	11.77
QLJM [31]	**3.18**	**4.46**	**3.38**	**4.76**	**6.40**	**5.08**	7.67	**9.89**	**8.18**	11.34	**14.21**	**12.16**
SDM [14]	2.87	3.96	3.04	4.37	5.81	4.66	7.12	9.30	7.67	10.68	14.02	11.78

表 2：*CS-lol*的统计数据。

类型	数值
一般信息	
电子竞技比赛（视频）数量	20
时长（秒）数量	39,868
游戏场景信息（描述）	
描述数量	24,770
每条描述的平均词数	6.89
每条描述的最少词数	1
单条描述中的最大词数	14
单条描述的平均时长（秒）	1.63
观众评论	
评论数	60,431
评论数（≥ 3词）	23,411
独立观众数	15,346
单条评论的平均词数	3.23
单条评论的最少词数	1
单条评论的最多词数	94

表3：检索方法的归一化折损累计增益结果。

模型	归一化折损累计增益@3	归一化折损累计增益@5	nDCG@10	nDCG@20
BM25 [19]	0.00349	0.00343	0.00343	**0.00399**
QLD [31]	0.00372	0.00352	0.00345	0.00391
QLJM [31]	0.00350	0.00333	0.00333	0.00388
SDM [14]	**0.00378**	**0.00359**	**0.00348**	0.00394

5.1 CS-lol 的信息检索

如第 4 节所示，我们为观众评论检索任务实现了信息检索基线检索方法。为确保结果可复现，我们公开了这些信息检索基线方法实现的源代码。本研究采用的基于词汇匹配的传统检索模型表现不佳，因为这些方法没有考虑词的语义含义。由于描述和评论中存在大量同义词，我们认为，使用考虑语义信息的检索模型（例如嵌入了知识图谱 [32]或外部知识 [28],的神经稠密段落检索方法 [7], ）可以显著提升性能。

5.2 CS-lol 的自然语言处理

近期一些研究聚焦于传统体育和电子竞技中的自然语言生成，但对自然语言理解的研究较少，尤其是在真实场景中。我们可以将*CS-lol* 数据集用于自然语言处理任务，例如命名实体识别（NER），该任务旨在识别提及命名实体的文本片段，并将其分类到预定义类别中。进行依存句法分析以提取实体之间的潜在关系也很有帮助 [13]。此外，未来的一个方向或许是邀请电子竞技专家，利用他们的专业知识为*CS-lol*补充标注。我们相信，此类标注能够显著加深我们对真实场景中电子竞技用语的理解，从而为电子竞技研究领域做出贡献。

6 结论

本文介绍了一个新颖的数据集*CS-lol*，其中包含电子竞技直播赛事中观众发布的评论，以及与之配对的游戏场景描述。我们提出了观众评论检索这一任务，即检索与场景描述相对应的观众评论，以便在直播过程中实时了解观众。我们介绍了*CS-lol* 的构建过程、特点及获取方式。本文还报告了在所提出的任务上采用信息检索基线检索方法进行的实验结果，并讨论了*CS-lol* 未来的发展方向。未来，我们计划应用并扩展*CS-lol* 及其方法，为电子竞技和传统体育研究领域作出贡献。

参考文献

[1] Florian Block、Victoria Hodge、Stephen Hobson、Nick Sephton、Sam Devlin、Marian F Ursu、Anders Drachen 和 Peter I Cowling。2018年。叙事字节：电子竞技中的数据驱动内容制作。载于*2018*年*ACM*电视与在线视频互动体验国际会议论文集。29–41。[2] Enrico Gandolfi。2016年。观看还是游玩，游戏才是关键：Twitch.tv上的游戏文化——表演者、游戏与观众。《游戏与虚拟世界期刊》8，1（2016年），63–82。[3] Mu Hu、Mingli Zhang 和 Yu Wang。2017年。观众为何选择继续观看直播视频平台上的内容？双重认同框架的解释。《人类行为中的计算机》75（2017年），594–606。[4] Hong Huang、Junjie H Xu、Xiaoling Ling 和 Pujana Paliyawan。2022年。电子竞技直播中协作式评论生成的句子标点。载于*2022*年 *IEEE*消费电子国际会议（*ICCE*）。IEEE，871–872。[5] Tatsuya Ishigaki、Goran Topic、Yumi Hamazono、Hiroshi Noji、Ichiro Kobayashi、Yusuke Miyao 和 Hiroya Takamura。2021年。从视觉、语言和结构化数据生成赛车游戏解说。载于第*14*届国际自然语言生成会议论文集。103–113。[6] Mark R Johnson 和 Jamie Woodcock。2019年。直播和 Twitch.tv 对电子游戏产业的影响。《媒体、文化与社会》41，5（2019年），670–688。[7] Vladimir Karpukhin、Barlas Oguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020年。面向开放域问答的密集段落检索。载于*2020*年自然语言处理实证方法会议论文集

自然语言处理（EMNLP）方法。6769–6781。[8] Han Soo Kim 和 Minjung Kim。2020年。一起在线观看体育赛事？社交直播服务使用带来的心理影响。《体育管理评论》23，5（2020年），869–882。[9] Seyun Kim。2021年。电子竞技广播解说员属性对广播态度和持续观看意愿的影响。《国际先进智能融合期刊》10，1（2021年），209–215。[10] Young Hoon Kim、John Nauright 和 Chokechai Suveatwatanakul。2020年。电子竞技的兴起及其在后新冠疫情时期持续增长的潜力。《体育与社会》23，11（2020年），1861–1871。[11] Athanasios Vasileios Kokkinakis、Simon Demediuk、Isabelle Nölle、Oluseyi Olarewaju、Sagarika Patra、Justus Robertson、Peter York、Alan Pedrassoli Pedrassoli Chitayat、Alistair Coates、Daniel Slawson 等。2020年。Dax：电子竞技中的数据驱动观众体验。载于ACM交互媒体体验国际会议。94–105。[12] Athanasios Vasileios Kokkinakis、Simon Demediuk、Isabelle Nölle、Oluseyi Olarewaju、Sagarika Patra、Justus Robertson、Peter York、Alan Pedrassoli Pedrassoli Chitayat、Alistair Coates、Daniel Slawson 等。2020年。Dax：电子竞技中的数据驱动观众体验。载于ACM交互媒体体验国际会议。94–105。[13] Jing Li、Aixin Sun、Jianglei Han 和 Chenliang Li。2020年。命名实体识别深度学习综述。《IEEE知识与数据工程汇刊》（2020年）。[14] Donald Metzler 和 W. Bruce Croft。2005年。词项依赖关系的马尔可夫随机场模型。载于第28届ACM SIGIR信息检索研究与发展国际会议论文集。ACM，472–479。https://doi.org/10.1145/1076034.1076115。/。Newzoo。。Newzoo《2022年全球电子竞技与直播市场报告》。检索于2023年10月10日，网址：https://newzoo.com/insights/trend-reports/newzoo-global-esports-live-streaming-market-report-2022-free-version[15] [16] Kseniia Palin、Anna Maria Feit、Sunjun Kim、Per Ola Kristensson 和 Antti Oulasvirta。2019年。人们如何在移动设备上打字？一项针对37,000名志愿者的研究观察。载于第21届人机交互与移动设备及服务国际会议论文集。1–12。[17] Jason G Reitman、Maria J Anderson-Coto、Minerva Wu、Je Seok Lee 和 Constance Steinkuehler。2020年。电子竞技研究：文献综述。《游戏与文化》15，1（2020年），32–50。[18] Charles Ringer、Mihalis Nicolaou 和 James Walker。2020年。TwitchChat：用于探索直播聊天的数据集。载于AAAI人工智能与交互式数字娱乐会议论文集，第16卷。259–265。[19] Stephen E Robertson、Steve Walker、Susan Jones、Micheline M Hancock-Beaulieu、Mike Gatford 等。1995年。TREC-3中的Okapi。NIST特别出版物 SP 109（1995年），109。[20] Joseph Rubleske、Travis Fletcher 和 Brett Westerfeld。2020年。电子竞技分析：学生研究项目与课程计划的入门指南和资源。《教学法期刊》23（2020年）。[21] Max Sjöblom 和 Juho Hamari。2017年。人们为什么观看他人玩电子游戏？关于Twitch用户动机的实证研究。《人类行为中的计算机》75（2017年），985–996。[22] Tsunehiko Tanaka 和 Edgar Simo-Serra。2021年。LoL-V2T：大规模电子竞技视频描述数据集。载于IEEE/CVF计算机视觉与模式识别会议研讨会论文集。4557–4566。[23] Jan B. Vornhagen、April Tyack 和 Elisa D. Mekler。2020年。CHI PLAY中的统计显著性检验：挑战与提升透明度的机遇。载于计算机与人类游戏交互年度研讨会论文集（线上活动，加拿大）（CHI PLAY '20）。计算机协会，美国纽约州纽约市，4–18。[24] Günter Wallner、Marnix Van Wijland、Regina Bernhaupt 和 Simone Kriglstein。2021年。玩家想要什么：玩家对游戏结束后可视化内容的信息需求。载于2021年CHI计算系统中的人为因素会议论文集。1–13。[25] Junjie H Xu、Yulin Cai、Zhou Fang 和 Pujana Paliyawan。2021年。通过基于精彩片段的弹幕促进直播游戏观众的心理健康。载于第10届全球消费电子会议论文集（日本京都）。IEEE，694–696。[26] Junjie H Xu、Zhou Fang、Satoru Ohno、Qihang Chen 和 Pujana Paliyawan。2021年。利用精彩片段提示并调整音高和响度的格斗游戏解说员。载于第10届IEEE全球消费电子会议论文集（日本京都）。IEEE，673–677。[27] Junjie H Xu、Hong Huang、Xiaoling Ling 和 Pujana Paliyawan。2022年。电子竞技直播中协作式评论生成的句子标点。载于2022年IEEE消费电子国际会议（ICCE）。IEEE，867–870。[28] Ikuya Yamada、Akari Asai、Hiroyuki Shindo、Hideaki Takeda 和 Yuji Matsumoto。2020年。LUKE：采用实体感知自注意力的深度上下文化实体表示。载于2020年自然语言处理实证方法会议（EMNLP）论文集。6442–6454。

[29] Peilin Yang、Hui Fang 和 Jimmy Lin。2018年。Anserini：使用 Lucene 的可复现排序基线。*ACM J. Data Inf. Qual.*10，4（2018年），16:1–16:20。[30] Huanyu Yu、Shuo Cheng、Bingbing Ni、Minsi Wang、Jian Zhang 和 Xiaokang Yang。2018年。面向体育叙事的细粒度视频描述。在*IEEE*计算机视觉与模式识别会议论文集中。6006–6015。[31] ChengXiang Zhai 和 John D. Lafferty。2004年。应用于信息检索的语言模型平滑方法研究。*ACM Trans. Inf. Syst.* 22，2（2004年），179–214。[32] Da Zheng、Xiang Song、Chao Ma、Zeyuan Tan、Zihao Ye、Jin Dong、Hao Xiong、Zheng Zhang 和 George Karypis。2020年。Dgl-ke：大规模训练知识图谱嵌入。在第*43*届*ACM SIGIR*信息检索研究与发展国际会议论文集中。739–748。