SGSum: 一个面向体育赛事摘要的人工标注数据集

王佳安[1] 张汀依[1] 瞿剑峰[1] 刘庆升[3] 陈志刚[2] 李直旭[1,2⋆]

[1] 苏州大学计算机科学与技术学院 江苏 苏州
[2] 科大讯飞苏州研究院 江苏 苏州
[3] 安徽淘云科技有限公司 安徽 合肥
jawang1@stu.suda.edu.cn, zhixuli@suda.edu.cn

摘要 体育赛事摘要是一类特殊的文本摘要任务，旨在从体育赛事的实时评论文本中生成对应的新闻报道。因各类体育赛事的高被关注度，该任务具有较高的实用价值。然而，现有的相关数据集存在大量噪音问题。SportsSum 是此前唯一的大规模体育赛事摘要数据集，因仅使用了较为简单的规则进行数据清洗，该数据集仍然有至少 15% 数据为噪音数据。为了进一步推进该任务的研究，我们系统性地分析了这些噪声，制定了严格的人工清洗流程，并收集了更多数据，以此构建了一个规模更大质量更高的 SGSum 数据集。数据已公开在https://github.com/krystalan/SGSum

Keywords: 体育赛事摘要 · 文本摘要 · 自然语言生成.

1 介绍

文本摘要旨在为给定的一篇或多篇文章生成较为简短的文本表述。人们可以通过文本摘要更快地了解到文章的核心内容，以此来有效缓解大数据时代下文本数据过载问题。文本摘要现有的绝大部分进展通常是在通用新闻领域的数据集上取得的，例如 CNN/DailyMail[1, 4]、XSum[5] 等。而垂直领域的各类文本摘要任务往往由于数据集不足，研究的关注度较低。

体育赛事摘要是一类特殊的垂域领域文本摘要任务，如图 1所示，其从体育赛事的实时评论文本中生成对应的新闻文章。因体育自身的极大魅力，体育赛事摘要近年来受到了较多关注。体育赛事中通常会有现场解说员为观众们带来实时讲解。比赛结束后，网络上也会更新出相应的新闻报道以便于人们回顾整场比赛。然而由于体育赛事众多，并不是所有比赛都有相应的新闻报道。如果能从实时讲解中自动生成新闻文章将会大大提升媒体编辑们撰写新闻的效率并便捷人们获得更多的赛事资讯。

⋆ 李直旭是通讯作者。

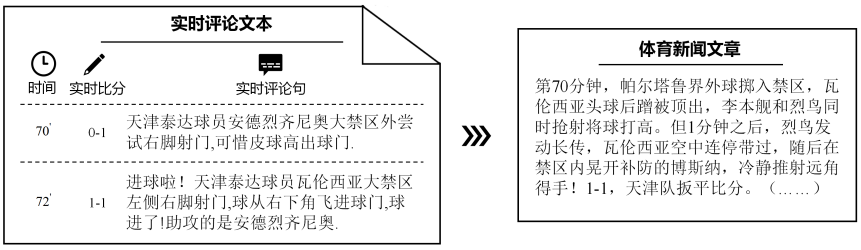

图 1. 体育赛事摘要任务示例。

相较于通用新闻领域的文本摘要任务，体育赛事摘要具有以下特点：1）体育赛事的实时评论文本记录了整场比赛发生的全部事件，通常会达到上千字，远超通用新闻文章长度; 2）实时评论文本与新闻文本之间存在文本风格差异，实时评论文本较为口语化。因此体育赛事摘要更具挑战性。Zhang 等人 [8] 首次讨论了该任务并构建了一个包含 150 场足球比赛的数据集。另一个在 NLPCC 2016 共享任务中使用的数据集包含了 900 场赛事 [6]。这些数据集虽然促进了体育赛事摘要的研究，但却因为数量有限并不满足深度学习算法研究的需求。为了解决这个问题，Huang 等人 [2] 提出了第一个大规模体育赛事摘要数据集 SportsSum，该数据集包含了 5428 场足球赛事的实时评论文本与对应新闻文本。然而该数据集存在大量噪音：部分新闻文本中包含广告信息；部分新闻文本中存在不能通过实时评论文本得出的历史信息；还有部分新闻文本与实时评论文本的赛事场次不对应。这是由于 SportsSum 仅使用了一些简单规则来清洗数据，只移除了数据集中的少部分噪音。根据我们的观察与统计，SportsSum 中至少仍有 15% 的数据为噪音数据。

为了进一步推进体育赛事摘要任务的研究，我们决定构建一个规模更大的高质量数据集 - SGSum (**S**ports **G**ame **Sum**marization) 数据集。为了构建该数据集，我们从新浪在线体育网站上收集了 2012 年至 2020 年期间全部的足球比赛数据，并制定了严格的人工清洗流程来处理收集到的新闻文本，最终获得了 7854 条高质量的体育赛事摘要数据。

我们的贡献总结如下:

- 我们提出并公开了现有规模最大、质量最高的体育赛事摘要数据集：SGSum，包含 7854 场足球赛事对应的实时评论文本与新闻文本。
- 我们系统性地分析了以往数据集中存在的噪声，并在此基础上制定了更为严格的人工数据清洗流程来获得高质量数据。
- 我们从各个角度统计并分析了 SGSum 数据集以便研究者们更好地了解该任务。

2 SGSum 数据集

2.1 数据收集

由于足球赛事的数据量充足且便于收集，因此关于此任务的研究均采用足球赛事领域的数据，尽管如此，当前基于足球领域的体育赛事摘要研究对其他类型的体育赛事也具有广泛的借鉴意义。与先前工作 [8, 2] 类似，我们选择从新浪在线体育网站[1]上收集数据。具体地，我们收集了从 2012 年（含）至 2020 年（含）期间的所有足球赛事数据，得到了 9304 场既包含实时评论页又包含新闻页的足球赛事。在进一步收集这些页面中的数据时，共有 664 场赛事中的实时评论页或新闻页已无法正常访问或无法收集到有效信息，因此我们初步得到了 8640 场足球赛事的实时评论文本与新闻文本。

2.2 数据清洗

基于对原始数据的观察，我们发现**实时评论文本**均以**结构化**表格的形式呈现，表格分为三列，分别记录 “时间”、“实时比分” 以及 “实时评论句” 信息。其中 “时间” 和 “实时比分” 分别暗示了 “实时评论句” 所描述的事件发生的时间以及当时的比分情况。大部分实时评论句都有对应的时间信息且数据质量较高，因此我们保留这部分实时评论句。而那些没有对应时间信息的实时评论句描述的内容往往与当前比赛无关，所以不被保留。而**新闻文本**则是**非结构化**的文本，具有大量的噪声信息。具体地，我们将这些噪声分为三类：1）广告与无关网页标签：我们发现新闻页中很有可能会出现与比赛无关的广告信息与网页标签等。我们对 SportsSum 中的新闻文本进行了分析，发现其中约有 9.8% 的数据存在这类噪声；2）其余比赛描述：据我们统计，在数据收集中得到的约 12% 的新闻页同时包含了多场比赛的新闻报道，而先前的工作并未考虑到这种情况。我们发现 SportsSum 中大约 2.2% 的新闻文本包含了其他赛事的描述；3）历史信息：体育新闻中往往会存在与当前赛事无关的描述，例如在一开始交代本场比赛参赛双方的历史交战情况、近期的状态等。为了缓解该问题，SportsSum 采用了基于规则的方法识别开场词（例如 “一开场”、“开场后” 等）并去除开场词之前的无关描述。然而据我们统计，SportsSum 中约有 4.4% 的新闻文本存在开场识别错误的问题，这是因为基于规则的开场识别并不能够覆盖所有的情况。除此之外，与当前场次无关的描述并不一定只会出现在新闻的开始处，许多新闻报道往往会在

[1] http://match.sports.sina.com.cn/index.html

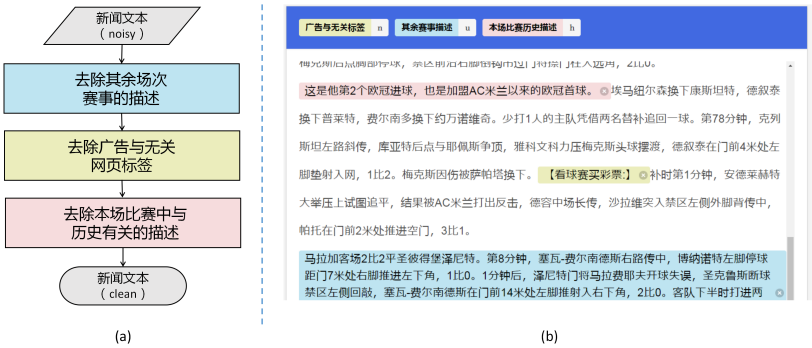

图 2. (a) 人工清洗新闻文本流程图 (b) 人工标注界面截图

中间穿插历史赛况信息，例如有球员进球，新闻有可能会介绍该球员在前几轮中同样出色的表现或者统计出这次进球是他本赛季的第几次进球。新闻也有可能在结尾处总结本场比赛的战果对参赛双方后续赛事的影响或预告未来赛事。这些信息均不能通过实时评论文本获得，但却被先前工作所忽略。

为了解决这三类噪声，我们定义了人工清洗新闻文本流程。如图 2(a) 所示，对于一篇给定的体育新闻，首先去除其余赛事的描述；其次去除广告与无关网页标签；最后识别出本场比赛与历史相关的描述并去除。为了方便控制数据质量，我们先人工标注出上述噪声词句，并在标注通过后统一进行去除。人工标注的界面如图 2(b) 所示，共有 9 名志愿者参与标注，志愿者均为我们学校体育学院的学生。我们随机挑选了 50 篇新闻文本作为测试样本，让 9 名志愿者同时进行标注，并基于标注结果确定了 7 名标注员与 2 名资深标注员。之后，每一条新闻文本会随机分配给 2 名标注员，若标注结果不一致，则由资深标注员进行确定。最终所有新闻文本的标注结果由另外 2 名数据专家进行检查，若数据专家认为标注不通过则该新闻文本需要重新标注。清洗结束后，我们发现有些新闻文本并不包含当前场次的比赛报道，或包含信息过少，例如只描述了赛事结果。我们去除了这部分新闻文本，最终得到了 7854 条符合规定的高质量新闻文本。这些新闻文本与对应的实时评论文本组成了 SGSum 数据集。

2.3 数据划分

通过对新闻文本的观察，我们发现有明星球员登场或参赛队伍粉丝较多的赛事往往备受关注，这些赛事的新闻报道通常也更加丰富。为了更好地检验模型生成丰富新闻文本的能力，我们在验证集与测试集中选用了**新闻文本**

体育赛事摘要数据集	# Examples	新闻文本						实时评论文本					
		字数		词数		句数		字数		词数		句数	
		avg.	95th pctl.	avg.	95th pctl.	avg.	95th pctl.	avg.	95th pctl.	avg.	95th pctl.	avg.	95th pctl.
Zhang 等 [8]	150	-	-	-	-	-	-	-	-	-	-	-	-
NLPCC 2016 共享任务 [6]	900	-	-	-	-	-	-	-	-	-	-	-	-
SportsSum[2]	5428	801.11	1558	427.98	924	23.80	39	3459.97	5354	1825.63	3133	193.77	379
SGSum 训练集-简短类	2618	136.23	344	78.93	199	5.16	11	2958.77	3744	1596.13	1998	116.68	148
SGSum 训练集-中等类	2618	588.36	820	337.70	476	19.67	24	2005.42	3427	1058.96	1825	213.64	431
SGSum 训练集-丰富类	1618	1097.93	1655	638.34	977	32.25	46	1801.31	3351	952.16	1825	230.26	444
SGSum 验证集-丰富类	500	1089.37	1631	634.49	965	33.12	45	1755.32	3310	927.57	1799	224.34	380
SGSum 测试集-丰富类	500	1095.42	1637	636.52	971	33.15	46	1713.89	3245	904.24	1771	230.41	426
SGSum-全部	7854	606.80	1430	351.30	845	19.38	40	2251.62	3581	1200.31	1915	187.69	388

表 1. 体育赛事摘要数据集统计信息

篇幅较长的数据。具体地，我们将 7854 条体育赛事摘要数据按照**新闻文本**的句数均匀地划分为三类：“简短”、“中等” 与 “丰富”。其中 “简短” 类别的新闻文本句数最少，“丰富” 类别最多。“中等” 则介于两者之间。每类包含 2618 条数据。更进一步，我们分别从 “丰富” 类别中随机选取了 500 条数据作为验证集，另外 500 条作为测试集。剩余的 6854 条数据划分到训练集。

2.4 数据集统计

如表 1所示，SGSum 是数据量最大的体育赛事摘要数据集。其中平均每场赛事的实时评论文本包含 2251.62 个汉字、新闻文本包含 606.80 个汉字。通过统计分析，我们发现：1）实时评论文本中：SportsSum 中实时评论文本长度超过了 SGSum 中的平均长度，这是由于在 SportsSum 中，实时评论页的所有文本数据均计算在内，而 SGSum 只统计了包含时间信息的实时评论句，未包含时间信息的实时评论句往往与本场比赛无关，因此我们将其排除在外；2）新闻文本中：由于我们在新闻文本上定义了严格的人工清洗流程，所以 SGSum 的平均新闻文本长度小于 SportsSum 的平均新闻文本长度。除此之外，SGSum 中 “简短”、“中等” 和 “丰富” 类别的平均新闻长度呈递增趋势，符合我们的设定。

3 相关工作

体育赛事摘要的研究仍处于初期，Zhang 等人 [8] 首次讨论了这个任务，随后该任务也被 NLPCC 2016 所关注并作为共享任务出现在人们的视野中 [6]。早期对于该任务的研究 [8, 9, 7, 3] 主要关注在如何从实时评论文本中选出已有的句子组成新闻文本，然而这些方法忽略了实时评论文本与新闻文本之间的文本风格差异。Huang 等人 [2] 提出了第一个大规模的体育赛事摘要数据集 SportsSum，并在该数据集上求助于 seq2seq 模型来缓解上述问题。

4 结论

我们在本文提出了 SGSum 数据集，是现有的数据量最大且质量最高的体育赛事摘要数据集。为了提升数据的质量，我们分析了以往数据集中的噪声，并制定人工清洗流程。除此之外，我们从各个角度对现有的体育赛事摘要数据集进行统计分析，以便研究者们能够更好地了解本任务。我们希望 SGSum 数据集能够进一步推进体育赛事摘要的发展。

致谢

本工作由国家重点研发计划（No. 2018AAA0101900），国家自然科学基金（No. 62072323, 61632016），江苏省自然科学基金（No. BK20191420），软件新技术与产业化协同创新中心，以及江苏高校优势学科建设工程部分资助。

参考文献

1. Hermann, K., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: NIPS (2015)
2. Huang, K.H., Li, C., Chang, K.W.: Generating sports news from live commentary: A chinese dataset for sports game summarization. In: AACL/IJCNLP (2020)
3. Liu, M., Qi, Q., Hu, H., Ren, H.: Sports news generation from live webcast scripts based on rules and templates. In: NLPCC/ICCPOL (2016)
4. Nallapati, R., Zhou, B., Santos, C.D., Çaglar Gülçehre, Xiang, B.: Abstractive text summarization using sequence-to-sequence rnns and beyond. In: CoNLL (2016)
5. Narayan, S., Cohen, S.B., Lapata, M.: Don ' t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In: EMNLP (2018)
6. Wan, X., Zhang, J., ge Yao, J., Wang, T.: Overview of the nlpcc-iccpol 2016 shared task: Sports news generation from live webcast scripts. In: NLPCC/ICCPOL (2016)
7. ge Yao, J., Zhang, J., Wan, X., Xiao, J.: Content selection for real-time sports news construction from commentary texts. In: INLG (2017)
8. Zhang, J., ge Yao, J., Wan, X.: Towards constructing sports news from live text commentary. In: ACL (2016)
9. Zhu, L., Wang, W., Chen, Y., Lv, X., Zhou, J.: Research on summary sentences extraction oriented to live sports text. In: NLPCC/ICCPOL (2016)