GOAL: Towards Benchmarking Few-Shot Sports Game Summarization

Jiaan Wang[1*]   **Tingyi Zhang**[1]   **Haoxiang Shi**[2]

[1]Soochow University, Suzhou, China   [2]Waseda University, Tokyo, Japan

{jawang1,tyzhang1}@stu.suda.edu.cn   hollis.shi@toki.waseda.jp

Abstract

Sports game summarization aims to generate sports news based on real-time commentaries. The task has attracted wide research attention but is still under-explored probably due to the lack of corresponding English datasets. Therefore, in this paper, we release GOAL, the first English sports game summarization dataset. Specifically, there are 103 commentary-news pairs in GOAL, where the average lengths of commentaries and news are 2724.9 and 476.3 words, respectively. Moreover, to support the research in the semi-supervised setting, GOAL additionally provides 2,160 unlabeled commentary documents. Based on our GOAL, we build and evaluate several baselines, including extractive and abstractive baselines. The experimental results show the challenges of this task still remain. We hope our work could promote the research of sports game summarization. The dataset has been released at https://github.com/krystalan/goal.

1 Introduction

Given the live commentary documents, the goal of sports game summarization is to generate the corresponding sports news (Zhang et al., 2016). As shown in Figure 1, the commentary document records the commentaries of a whole game while the sports news briefly introduces the core events in the game. Both the lengthy commentaries and the different text styles between commentaries and news make the task challenging (Huang et al., 2020; Wang et al., 2021, 2022a).

Zhang et al. (2016) propose sports game summarization task and construct the first dataset which contains 150 samples. Later, another dataset with 900 samples is presented by Wan et al. (2016). To construct the large-scale sports game summarization data, Huang et al. (2020) propose SportsSum with 5,428 samples. Huang et al. (2020) first adopt
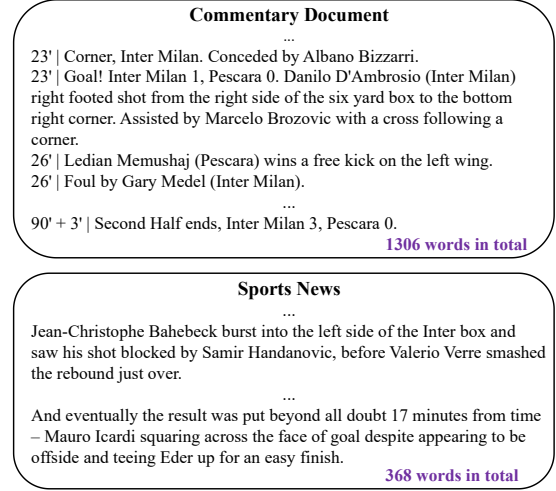
---
*Corresponding author.

Figure 1: An example of sports game summarization.

deep learning technologies for this task. Further, Wang et al. (2021) find the quality of SportsSum is limited due to the original rule-based data cleaning process. Thus, they manually clean the SportsSum dataset and obtain SportsSum2.0 dataset with 5,402 samples. Wang et al. (2022a) point out the hallucination phenomenon in sports game summarization. In other words, the sports news might contain additional knowledge which does not appear in the corresponding commentary documents. To alleviate hallucination, they propose K-SportsSum dataset which includes 7,428 samples and a knowledge corpus recording the information of thousands of sports teams and players.

Though great contributions have been made, all the above sports game summarization datasets are Chinese since the data is easy to collect. As a result, all relevant studies (Zhang et al., 2016; Yao et al., 2017; Wan et al., 2016; Zhu et al., 2016; Liu et al., 2016; Lv et al., 2020; Huang et al., 2020; Wang et al., 2021) focus on Chinese, which might be difficult for global researchers to understand and study this task. To this end, in this paper, we present GOAL, the first English sports game summariza-

tion dataset which contains 103 samples, each of which includes a commentary document as well as the corresponding sports news. Specifically, we collect data from an English sports website[1] that provides both commentaries and news. We study the dataset in few-shot scenario due to: (1) When faced with real applications, few-shot scenarios are more common. Compared with the summarization datasets in the news domain which usually contain tens of thousands of samples, the largest Chinese sports game summarization dataset is still in a limited scale. (2) The English sports news is non-trivial to collect. This is because manually writing sports news is labor-intensive for professional editors (Huang et al., 2020; Wang et al., 2022a). We find only less than ten percent of sports games on the English website[1] are provided with sports news. To further support the semi-supervised research on GOAL, we additionally provide 2,160 unlabeled commentary documents. The unlabeled commentary documents together with labeled samples could train sports game summarization models in a self-training manner (Lee et al., 2013).

Based on our GOAL, we build and evaluate several baselines of different paradigms, i.e., extractive and abstractive baselines. The experimental results show that the sports game summarization task is still challenging to deal with.

2 GOAL

2.1 Data Collection

Goal.com is a sports website with plenty of football games information. We crawl both the commentary documents and sports news from the website. Following Zhang and Eickhoff (2021), the football games are collected from four major soccer tournaments including the UEFA Champions League, UEFA Europa League, Premier League and Series A between 2016 and 2020. As a result, we collect 2,263 football games in total. Among them, all games have commentary documents but only 103 games have sports news. Next, the 103 commentary-news pairs form GOAL dataset with the splitting of 63/20/20 (training/validation/testing). The remaining 2,160 (unlabeled) commentary documents are regarded as a supplement to our GOAL, supporting the semi-supervised setting in this dataset. Figure 1 gives an example from GOAL.

Dataset	# Num.	Commentary		News	
		Avg.	95th pctl	Avg.	95th pctl
Chinese					
Zhang et al. (2016)	150	-	-	-	-
Wan et al. (2016)	900	-	-	-	-
SportsSum (Huang et al., 2020)	5428	1825.6	3133	428.0	924
SportsSum2.0 (Wang et al., 2021)	5402	1828.6	-	406.8	-
K-SportsSum (Wang et al., 2022a)	7854	1200.3	1915	351.3	845
English					
GOAL (supervised)	103	2724.9	3699	476.3	614
GOAL (semi-supervised)	2160	2643.7	3610	-	-

Table 1: Statistics of GOAL and previous datasets. "# Num." indicates the number of samples in the datasets. "Avg." and "95th pctl" denote the average and 95th percentile number of words, respectively.

2.2 Statistics

As shown in Table 1, Goal is more challenging than previous Chinese datasets since: (1) the number of samples in GOAL is significantly less than those of Chinese datasets; (2) the length of input commentaries in GOAL is much longer than the counterparts in Chinese datasets.

In addition, compared with Chinese commentaries, English commentaries do not provide real-time scores. Specifically, each commentary in Chinese datasets is formed as (t, c, s), where t is the timeline information, c indicates the commentary sentence and s denote current scores at time t. The commentaries in GOAL do not provide s element. Without this explicit information, models need to implicitly infer the real-time status of games.

2.3 Benchmark Settings

Based on our GOAL and previous Chinese datasets, we introduce supervised, semi-supervised, multi-lingual and cross-lingual benchmark settings. The first three settings all evaluate models on the testing set of GOAL, but with different training samples: (i) *Supervised Setting* establishes models on the 63 labeled training samples of GOAL; (ii) *Semi-Supervised Setting* leverages 63 labeled and 2160 unlabeled samples to train the models; (iii) *Multi-Lingual Setting* trains the models only on other Chinese datasets.

Moreover, inspired by Feng et al. (2022); Wang et al. (2022b); Chen et al. (2022b), we also provide a (iv) *Cross-Lingual Setting* on GOAL, that lets the models generate Chinese sports news based on the English commentary documents. To this end, we employ four native Chinese as volunteers to translate the original English sports news (in GOAL dataset) to Chinese. Each volunteer majors in Computer Science, and is proficient in English.

The translated results are checked by a data expert. Eventually, the cross-lingual setting uses the 63 cross-lingual samples to train the models, and verify them on the cross-lingual testing set.

2.4 Task Overview

For a given live commentary document $C = \{(t_1, c_1), (t_2, c_2), ..., (t_n, c_n)\}$, where t_i is the timeline information, c_i is the commentary sentence and n indicates the total number of commentaries, sports game summarization aims to generate its sports news $R = \{r_1, r_2, ..., r_m\}$. r_i denotes the i-th news sentence. m is the total number of news sentences.

3 Experiments

3.1 Baselines and Metrics

We only focus on the supervised setting, and reserve other benchmark settings for future work. The adopted baselines are listed as follows:

- Longest: the longest k commentary sentences are selected as the summaries (i.e., sports news). It is a common baseline in summarization.
- TextRank (Mihalcea and Tarau, 2004) is a popular unsupervised algorithm for extractive summarization. It represents each sentence from a document as a node in an undirected graph, and then extracts sentences with high importance as the corresponding summary.
- PacSum (Zheng and Lapata, 2019) enhances the TextRank algorithm with directed graph structures, and more sophisticated sentence similarity technology.
- PGN (See et al., 2017) is a LSTM-based abstractive summarization model with copy mechanism and coverage loss.
- LED (Beltagy et al., 2020) is a transformer-based generative model with a modified self-attention operation that scales linearly with the sequence length.

  Among them, Longest, TextRank and PacSum are extractive summarization models which directly select commentary sentences to form sports news. PGN and LED are abstractive models that generate sports news conditioned on the given commentaries. It is worth noting that previous sports game summarization work typically adopts the pipeline methods (Huang et al., 2020; Wang et al., 2021, 2022a), where a selector is used to select important commentary sentences, and then a rewriter conveys each commentary sentence to a news sentence.

Method	Validation			Testing		
	R-1	R-2	R-L	R-1	R-2	R-L
Extractive Methods						
Longest	31.2	4.6	20.0	30.3	4.2	19.5
TextRank	30.3	3.7	20.2	27.6	2.9	18.8
PacSum	31.8	4.9	20.6	31.0	5.3	19.6
Abstractive Methods						
PGN	34.2	6.3	22.7	32.8	5.7	21.4
LED	**37.8**	**9.5**	**25.2**	**34.7**	**7.8**	**24.3**

Table 2: Experimental results on GOAL.

However, due to the extremely limited scale of our GOAL, we do not have enough data to train the selector. Therefore, all baselines in our experiments are end-to-end methods.

To evaluate baseline models, we use ROUGE-1/2/L scores (Lin, 2004) as the automatic metrics. The employed evaluation script is based on py-rouge toolkit[2].

3.2 Implementation Details

For Longest, TextRank[3] and PacSum[4] baselines, we extract three commentary sentences to form the sports news. β, λ_1 and λ_2 used in PacSum are set to 0.1, 0.9 and 0.1, respectively. The sentence representation in PacSum is calculated based on TF-IDF value. For LED baseline, we utilize *led-base-16384*[5] with the default settings. We set the learning rate to 3e-5 and batch size to 4. We train the LED baselines with 10 epochs and 20 warmup steps. During training, we truncate the input and output sequences to 4096 and 1024 tokens, respectively. In the test process, the beam size is 4, minimum decoded length is 200 and maximum length is 1024.

3.3 Main Results

Table 2 shows experimental results on GOAL. The performance of extractive baselines is limited due to the different text styles between commentaries and news (commentaries are more colloquial than new). PGN outperforms the extractive methods since it can generate sports news not limited to original words or phrases. However, such a LSTM-based method cannot process long documents efficiently. LED, as an abstractive method, achieves

---

[2]https://github.com/Diego999/py-rouge
[3]https://github.com/summanlp/textrank
[4]https://github.com/mswellhao/PacSum
[5]https://huggingface.co/allenai/led-base-16384

News Sentence: Fiorentina finally made the most of their superiority eight minutes before the break – Kalinic [MASK] Buffon from a tight angle after being fed into the box by Filippo Bernardeschi.

Predict (top5): beat, caught, charged, run, hit

Label: beating

News Sentence: Juve almost levelled on the stroke of half-time, but Maxi Olivera crucially [MASK] Higuain's goal-bound effort to preserve the hosts' lead.

Predict (top5): stopped, assisted, ended, saved, broke

Label: blocked

Figure 2: Predicting key verbs during generating sports news based on LED baseline. During predicting, the model is conditioned on the corresponding commentaries. The gray text indicates where the model do not compute self-attention mechanism.

the best performance among all baselines due to its abstractive nature and sparse attention.

3.4 Discussion

Moreover, we also manually check the generated results of LED baseline, and find that the generated sports news contains many repeated phrases and sentences, and fail to capture some important events in the games. We conjecture this is because the few-shot training samples make the baseline difficult to learn the task effectively.

To further verify whether the model is familiar with sports texts, we let the model predict the key verbs during generating the sports news, which is similar to Chen et al. (2022a). As shown in Figure 2, when faced with a simple and common event (i.e., beating), the trained LED model could predict the right verb (i.e., beat). However, for a complex and uncommon event (i.e., blocking), the model cannot make correct predictions.

Therefore, it is an urgent need to utilize the external resources to enhance the model's abilities to know sports texts and deal with sports game summarization. For example, considering the semi-supervised setting and multi-lingual settings, where the models could make use of unlabeled commentaries and Chinese samples, respectively. Inspired by Wang et al. (2022c), the external resources and vanilla few-shot English training samples could be used to jointly train sports game summarization models in the multi-task, knowledge-distillation

or pre-training framework. In addition, following Feng et al. (2021a), another promising way is to adopt other long document summarization resources to build multi-domain or cross-domain models with sports game summarization.

4 Conclusion

In this paper, we present GOAL, the first English sports game summarization dataset, which contains 103 commentary-news samples. Several extractive and abstractive baselines are built and evaluated on GOAL to benchmark the dataset in different settings. We further analyze the model outputs and show the challenges still remain in sports game summarization. In the future, we would like to (1) explore the semi-supervised and multi-lingual settings on GOAL; (2) leverage graph structure to model the commentary information, and generate sports news in a graph-to-text manner (Feng et al., 2021b), or even consider the temporal information (i.e., t_i) in the graph structure (Zhang et al., 2022).

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Jiashu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yijiang Chen, and Chang Su. 2022a. Probing simile knowledge from pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5875–5887, Dublin, Ireland. Association for Computational Linguistics.

Yulong Chen, Ming Zhong, Xuefeng Bai, Naihao Deng, Jing Li, Xianchao Zhu, and Yue Zhang. 2022b. The cross-lingual conversation summarization challenge. *ArXiv*, abs/2205.00379.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021a. A survey on dialogue summarization: Recent advances and new frontiers. *ArXiv*, abs/2107.03175.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. MSAMSum: Towards benchmarking multi-lingual dialogue summarization. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 1–12, Dublin, Ireland. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021b. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21,*

pages 3808–3814. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Kuan-Hao Huang, Chen Li, and Kai-Wei Chang. 2020. Generating sports news from live commentary: A Chinese dataset for sports game summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 609–615, Suzhou, China. Association for Computational Linguistics.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Maofu Liu, Qiaosong Qi, Huijun Hu, and Han Ren. 2016. Sports news generation from live webcast scripts based on rules and templates. In *Natural Language Understanding and Intelligent Applications*, pages 876–884. Springer.

Xue-Qiang Lv, Xin-Dong You, Wen-Chao Wang, and Jian-She Zhou. 2020. Generate football news from live webcast scripts based on character-cnn with five strokes. *Journal of Computers*, 31(1):232–241.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Xiaojun Wan, Jianmin Zhang, Jin-ge Yao, and Tianming Wang. 2016. Overview of the nlpcc-iccpol 2016 shared task: sports news generation from live webcast scripts. In *Natural Language Understanding and Intelligent Applications*, pages 870–875. Springer.

Jiaan Wang, Zhixu Li, Qiang Yang, Jianfeng Qu, Zhigang Chen, Qingsheng Liu, and Guoping Hu. 2021. *SportsSum2.0: Generating High-Quality Sports News from Live Text Commentary*, page 3463–3467. Association for Computing Machinery, New York, NY, USA.

Jiaan Wang, Zhixu Li, Tingyi Zhang, Duo Zheng, Jianfeng Qu, An Liu, Lei Zhao, and Zhigang Chen. 2022a. Knowledge enhanced sports game summarization. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22,

page 1045–1053, New York, NY, USA. Association for Computing Machinery.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022b. Clidsum: A benchmark dataset for cross-lingual dialogue summarization. *ArXiv*, abs/2202.05599.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022c. A survey on cross-lingual summarization. *ArXiv*, abs/2203.12515.

Jin-ge Yao, Jianmin Zhang, Xiaojun Wan, and Jianguo Xiao. 2017. Content selection for real-time sports news construction from commentary texts. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 31–40, Santiago de Compostela, Spain. Association for Computational Linguistics.

Jianmin Zhang, Jin-ge Yao, and Xiaojun Wan. 2016. Towards constructing sports news from live text commentary. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1361–1371, Berlin, Germany. Association for Computational Linguistics.

Ruochen Zhang and Carsten Eickhoff. 2021. SOCCER: An information-sparse discourse state tracking collection in the sports commentary domain. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4325–4333, Online. Association for Computational Linguistics.

Tingyi Zhang, Zhixu Li, Jiaan Wang, Jianfeng Qu, Lin Yuan, An Liu, Lei Zhao, and Zhigang Chen. 2022. Aligning internal regularity and external influence of multi-granularity for temporal knowledge graph embedding. In *Database Systems for Advanced Applications*.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Liya Zhu, Wenchao Wang, Yujing Chen, Xueqiang Lv, and Jianshe Zhou. 2016. Research on summary sentences extraction oriented to live sports text. In *Natural Language Understanding and Intelligent Applications*, pages 798–807. Springer.