Towards Constructing Sports News from Live Text Commentary

Jianmin Zhang **Jin-ge Yao** **Xiaojun Wan**
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
Key Laboratory of Computational Linguistic (Peking University), MOE, China
{zhangjianmin2015, yaojinge, wanxiaojun}@pku.edu.cn

Abstract

In this paper, we investigate the possibility to automatically generate sports news from live text commentary scripts. As a preliminary study, we treat this task as a special kind of document summarization based on sentence extraction. We formulate the task in a supervised learning to rank framework, utilizing both traditional sentence features for generic document summarization and novelly designed task-specific features. To tackle the problem of local redundancy, we also propose a probabilistic sentence selection algorithm. Experiments on our collected data from football live commentary scripts and corresponding sports news demonstrate the feasibility of this task. Evaluation results show that our methods are indeed appropriate for this task, outperforming several baseline methods in different aspects.

1 Introduction

There are a huge number of sports games played each day. It is demanding and challenging to write corresponding news reports instantly after various games. Meanwhile, live text commentary services are available on the web and becoming increasingly popular for sports fans who do not have access to live video streams due to copyright reasons. Some people may also prefer live texts on portable devices. The emergence of live texts has produced huge amount of text commentary data. To the best of our knowledge, there exists few studies about utilizing this rich data source.

Manually written sports news for match report usually share the same information and vocabulary as live texts for the corresponding sports game. Sports news and commentary texts can be treated as two different sources of descriptions for the same sports events. It is tempting to investigate whether we can utilize the huge amount of live texts to automatically construct sports news, typically in a form of match report. Building such a system will largely relax the burden of sports news editors, making them free from repetitive tedious efforts for writing while producing sports news more efficiently.

In this work, we study the possibility to construct sports news in the form of match reports from given live text commentary scripts. As a concrete example we collect live text data and corre-

problem can hardly lead to the construction of reasonable sports news reports.

To overcome these difficulties, we explore some specific features of live text commentary scripts and formulate a system based on supervised learning to rank models for this task. In order to tackle the local redundancy issue, we also propose a probabilistic sentence selection strategy.

We summarize our contributions as follows:

- We originally study the task of sports news construction from live text commentary and we build datasets for supervised learning and evaluation for this task.
- We formulate the task in a learning to rank framework, utilizing both traditional features for document summarization and novel task-specific features during supervised learning.
- We propose a probabilistic sentence selection algorithm to address the issue of local redundancy in description.
- We conduct a series of experiments on a real dataset and the evaluation results verify the performance of our system. Results suggest that constructing sports news from live texts is feasible and our proposed methods can outperform a few strong baselines.

2 Problem Statement

2.1 Task Description

In this work, we treat the task of constructing sports news from live text commentary as a special kind of document summarization: extracting sentences from live text scripts to form a match report.

Formally, given a piece of live text commentary containing a collection of candidate sentences $S = \{s_1, s_2, \ldots, s_n\}$ describing a particular sports game G, we need to extract sentences to form a summary of G which are suitable to be formed as sports news. The total length should not exceed a pre-specified length budget B.

The overall framework of generic document summarization can still be retained for this preliminary study. We first rank all candidate sentences according to a sentence scoring scheme and then select a few sentences according to certain criteria to form the final generated news.

2.2 Data Collection

To the best of our knowledge, there does not exist off-the-shelf datasets for evaluating sports news construction. Therefore we have to build a new dataset for this study. We will focus on live text scripts for football (soccer) games as a concrete instance, since football live texts are the easiest to collect. Note that the methods and discussions described in this paper can trivially generalize to other types of sports games.

Meanwhile, live text commentary services are extremely popular in China, where sports fans in many cases do not have access to live video streams due to copyright reasons. The most influential football live services are Sina Sports Live [2] and 163 Football Live [3]. For evaluation purposes we need to simultaneously collect both live texts and news texts describing the same sports games. Due to the convenience and availability of parallel data collection, we build our dataset from Chinese websites. For most football games, there exist both live text scripts recorded after the games and human-written news reports on both Sina Sports and 163 Football. We crawl live text commentary scripts for 150 football matches on Sina Sports Live. Figure 1 displays an example of the format of the live texts, containing the main commentary text along with information of the current timeline and scoreline.

Text Commentary	Timeline	Scoreline
莱万多夫斯基右路传球给到穆勒 (Lewandowski passes the ball to the right and finds Müller)	上半场 42' (first half 42')	2-0
穆勒停球后直接射门 (Müller stops the ball and gets a direct shot)	上半场 43' (first half 43')	2-0
切赫这边反应很快将球托出横梁 (Fast reaction from Cech to tip the ball over the bar)	上半场 43' (first half 43')	2-0

Figure 1: Illustration of the live text format

For every match, two different corresponding sports news reports are collected from Sina Sports Live and 163 Football Matches Live, respectively. These news reports are manually written by professional editors and therefore suitable to be treated as gold-standard news for our task. The average number of sentences in the live texts for one match is around 242, containing around 4,590 Chinese characters for that match. The gold-standard news reports contain 1,185 Chinese characters on average, forming around 32 sentences.

For both the gold-standard news and live text commentary scripts, we split them into sentences

[2] http://match.sports.sina.com.cn/
[3] http://goal.sports.163.com/

and then use a Chinese word segmentation tool [4] to segment the sentences into word sequences. For each sentence, we compute its TFIDF vector for calculating literal cosine similarity when used.

3 Constructing Sports News via Sentence Extraction

We build a system to automatically construct match reports from live text commentary. Since we have described the new challenges for this task, we may design a number of relevant features to address them. In this work, we cast the problem into supervised sentence extraction. Supervised approaches, especially those based on learning to rank (LTR), can better utilize the power of various task-dependent features (Shen and Li, 2011; Wang et al., 2013). For a given specific sports game, we extract features from all candidate sentences in the corresponding live texts and score the sentences using a learning to rank (LTR) model learned from the training data (Section 3.1). Then we select a few of them according to the ranking scores to form the constructed news (Section 3.3).

3.1 Training Data Format

Supervised sentence scoring models based on LTR require input training data in the format of (x_i, y_i) for each candidate sentence s_i, where x_i is the feature vector and y_i is the preference score. The feature vector x is described in Section 3.2. The score y will be defined to reflect the importance, or the tendency to be included in the final news report, of the candidate sentence. In this work we first calculate a group of ROUGE-2 F-scores (cf. Section 4.4.1) of the candidate sentence, treating each sentence in the gold-standard news as reference. The score y of the candidate sentence is then set to be the maximum among those ROUGE-2 F-scores. Later we will see that this scores can indeed serve as good learning targets.

3.2 Features

In this work, we extract both common features which have been widely used for generic document summarization (Shen and Li, 2011; Wang et al., 2013) and novel task-specific features aiming at proper sports news generation from live broadcast script. The features are described as follows.

3.2.1 Basic Features

Position: The position of each candidate sentence. Suppose there are n sentences in a document. For the i-th sentence, its position feature is computed as $1 - \frac{i-1}{n}$.

Length: The number of words contained in the sentence after stopwords removal.

Number of stopwords: The Number of stopwords contained in each sentence. Sentences with many stopwords should be treated as less important candidates.

Sum of word weights: The sum of TF-IDF weights for each word in a sentence.

Similarity to the Neighboring Sentences: We calculate the average cosine similarity of a candidate sentence to its previous N and the next N neighboring sentences. We set N as 1 and 2 here to get two different features.

3.2.2 Task-specific Features

The task we study has some unique properties compared with generic document summarization. For instance, in live text commentary for sports games such as football matches, the scripts not only contain descriptive texts but also the scoreline and timeline information. Such information can be utilized to judge the quality of candidate sentences as well. We extract a rich set of new features, which can be grouped into four types:

Explicit highlight markers: Explicit highlight marker words in a sentence are usually good indicators for its importance. Sentences with more marker words are more probable to be extracted and contained in news or reports for the games. For example, words such as "破门 (scores)" and "红牌 (red card)" in a sentence may indicate that the sentence is describing important events and will be more likely to be extracted. We collect a short list of 25 explicit highlight marker words [5]. For each marker word we create a binary feature to denote the presence or absence of that markers in each candidate sentence. We also use the number of markers as one feature, with the intuition that containing more marker words typically suggests more important sentences.

Scoreline features: An audience of sports games typically pays more attention on scoreline changes, especially those deadlock-breaking scores that break the game from ties. We use three

[4] We use the ICTCLAS toolkit for word segmentation in this work: http://ictclas.nlpir.org/

[5] We include the full list of marker words in the supplementary materials due to the space limit.

binary features to describe the scoreline information of each candidate sentence:

- An indicator feature on whether there was a change of scoreline when the narrator or commentator was producing that sentence.
- An indicator feature on whether the distance between the candidate sentence and the previous closest sentence with a change of scoreline is less than or equal to 5.
- An indicator feature showing whether the game was a draw or not at that time.

To better describe these features we give an example in Figure 2, where S1-S3 corresponds to the above three binary features, respectively.

Text Commentary	Timeline	Scoreline	S1	S2	S3
Both sides take advantages of counter attacks.	32'	1-1	0	0	0
1-2!!	33'	1-2	1	1	1
Alexis!!	33'	1-2	0	1	1
Özil finds the teammate byline followed by a low cross to far post, Alexis sends the ball into the net!	34'	1-2	0	1	1
Leicester players are unhappy.	34'	1-2	0	1	1

Figure 2: An example of scoreline features

Timeline features: The timestamp on each sentence can reflect the progress of a sports game. We divide a match into five different stages as "未赛 (not started)", "上半场(first half)", "中场休息(half-time)", "下半场(second half)" and "完赛(full-time)". Then we use five binary features to represent whether the sentence was describing a specific stage. We also use the specific time-stamp (in integral minutes) of the candidate sentence in the match as an additional feature. Suppose there are n minutes of the match (typically 90 minutes for football), for sentences on the time-stamp of the i-th minute , this feature is computed as $\frac{i}{n}$.

Player popularity: Sports fans usually focus more on the performance of the star players or inform players during the games. We design two features to utilize player information described in a candidate sentence: the number of players contained in the sentence and the sum of their popularity measurements. In this work the popularity of a player is measured using search engines for news: we use the name of a certain player as input query to Baidu News [6], and use the number of recent news retrieved to measure this player's popularity.

[6]http://news.baidu.com/

3.3 Sentence Selection

Once we have the trained LTR model, we can immediately construct news reports by selecting sentences with the highest scores. Unfortunately this simple strategy will suffer from redundancy in commentary, since the LTR scores are predicted independently for each sentence and assigning high scores for repeated commentary texts describing the same key event. Therefore, special care is needed in sentence selection. In principle, any In this work we propose a probabilistic approach based on determinantal point processes (Kulesza and Taskar, 2012, DPPs). This approach can naturally integrate the predicted scores from the LTR model while trying to avoid certain redundancy by producing more diverse extractions [7]. We first review some background knowledge on the model. More details can be found in the comprehensive survey (Kulesza and Taskar, 2012) covering this topic.

3.3.1 Determinantal Point Processes

Determinantal point processes (DPPs) are distributions over subsets that jointly prefer quality of each item and diversity of the whole subset. Formally, a DPP is a probability measure defined on all possible subsets of a group of items $\mathcal{Y} = \{1, 2, \ldots, N\}$. For every $Y \subseteq \mathcal{Y}$ we have:

$$\mathcal{P}(Y) = \frac{\det(L_Y)}{\det(L + I)}$$

where L is a positive semidefinite matrix typically called an *L-ensemble*. $L_Y \equiv [L_{ij}]_{i,j \in Y}$ denotes the restriction of L to the entries indexed by elements of Y, and $\det(L_\emptyset) = 1$. The term $\det(L + I)$ is the normalization constant which has a succinct closed-form and easy to compute. We can define the entries of L as follows:

$$L_{ij} = q_i \phi_i^\top \phi_j q_j = q_i \cdot \mathrm{sim}(i, j) \cdot q_j \qquad (1)$$

where we can think of $q_i \in \mathbb{R}^+$ as the *quality* of an item i and $\phi_i \in \mathbb{R}^n$ with $\|\phi_i\|_2 = 1$ denotes a normalised feature vector such that $\mathrm{sim}(i, j) \in [-1, 1]$ measures *similarity* between item i and item j. This simple definition gives rise to a distribution that places most of its mass on sets that are both high quality and diverse. This is intuitive in a

[7]Many other approaches can also be used to achieve similar effect, such as submodular maximization (Lin and Bilmes, 2010). We leave the comparison with these alternatives for future work study.

geometric sense since determinants are closely related to volumes; in particular, $\det(L_Y)$ is proportional to the volume spanned by the vectors $q_i\phi_i$ for $i \in Y$. Thus, item sets with both high-quality and diverse items will have the highest probability (Figure 3).

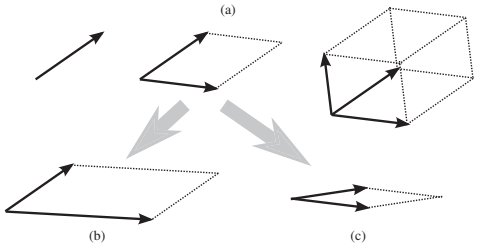

Figure 3: (a) The DPP probability of a set Y depends on the volume spanned by vectors $q_i\phi_i$ for $i \in Y$ (b) As length increases, so does volume. (c) As similarity increases, volume decreases.

3.3.2 Sentence Selection

In this work we formulate the sentence selection problem as maximum a posteriori (MAP) inference for DPPs, i.e. finding $\text{argmax}_Y \log\det(L_Y)$. It is known that MAP inference for DPPs is NP-hard (Gillenwater et al., 2012). Therefore we adopt the greedy approximate inference procedure used by Kulesza and Taskar (2011) which is fast and performs reasonably well in practice.

The remaining question is how to define the L-ensemble matrix L, or equivalently how to define itemwise quality q_i and pairwise similarity $\text{sim}(i, j)$, where each item corresponds to a candidate sentence. Since we have predicted scores for all candidates with the LTR model, we simply set q_i to be the ranking score for sentence i.

The definition of $\text{sim}(i, j)$ is more subtle since it directly address specific types of redundancy. The most straightforward definition is to use literal cosine similarity. This is used for traditional summarization problems (Kulesza and Taskar, 2011). However, the problem for constructing sports news from live broadcast script is rather different. A live broadcast script may use literally similar sentences to describe similar types of events happened at different time stamps. Simply removing sentences that are similar in content may become harmful to the preservation of important events [8].

One typical redundancy that we found in this study is local description redundancy. In live texts, an important event (such as goals) may be stressed multiple times consecutively by the commentator. Therefore in this study we use local literal similarity as a first attempt. Formally, the pairwise similarity is defined as:

$$\text{sim}(i,j) = \begin{cases} 0, \text{if } \max\{|i_p - j_p|, |i_t - j_t|\} > 1, \\ \cos(i,j), \text{ otherwise,} \end{cases}$$

where the subscripts i_p and i_t denotes position and timestamp for sentence i, respectively. In other words we treat sentences written consecutively within one minute as local descriptions and only calculate literal cosine similarity for them.

4 Experimental Setup

4.1 Data Preparation

As described earlier in Section 2.2, we evaluate the performance of different systems on our collected dataset. To utilize the dataset more sufficiently and draw more reliable conclusions, we perform cross-validation during evaluation. Specifically, we randomly divide the dataset into three parts with equal sizes, i.e. each has 50 pairs of live texts and gold-standard news. Each time we set one of them as the test set and use the remaining two parts for training and validation. We will mainly report the averaged results from all three folds. For unsupervised baselines the results are calculated similarly via averaging the performance on the test set.

4.2 Learning to Rank

For predicting ranking scores we use the Random Forest (RF) (Breiman, 2001) ensemble ranker of LambdaMart (Wu et al., 2010), implemented in RankLib [9]. We set the number of iterations to 300 and the sampling rate to 0.3. Using different values did not show real differences.

4.3 Compared Baseline Methods

Our system is compared with several baselines, typically traditional summarization approaches:

HeadTail: Using head and tail sentences only. Commentators usually describe some basic information of the two sides at the beginning and summarize the scoring events in the end of commentary. This baseline resembles the baseline of leading sentences for traditional summarization.

[8] Using cosine similarity for all similarity-dependent methods performs poorly in our experiments. Therefore we will not discuss cosine similarity in more details later.

[9] http://sourceforge.net/p/lemur/wiki/RankLib/; In preliminary experiments, we contrasted RF with support vector regression predictor as well as other pairwise and listwise LTR models. We found that RF consistently outperformed others.

Centroid: In centroid-based summarization (Radev et al., 2000), a pseudo-sentence of the document called centroid is calculated. The centroid consists of words with TFIDF scores above a predefined threshold. The score of each sentence is defined by summing the scores based on different features including cosine similarity of sentences with the centroid, position weight and cosine similarity with the first sentence.

LexRank: LexRank (Erkan and Radev, 2004) computes sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. In this model, a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences.

ILP: Integer linear programming (ILP) approaches (Gillick et al., 2008) cast document summarization as combinatorial optimization. An ILP model selects sentences by maximizing the sum of frequency-induced weights of bigram concepts [10] contained in the summary.

Highlight: This method is designed to show the effect of using merely the explicit highlight markers described in Section 3.2.2. The importance of a sentence is represented by the number of highlight markers it includes.

For fair comparisons the length of each constructed news report is limited to be no more than 1,000 Chinese characters, roughly the same with the average length of the gold-standard news. Note that we do not use the traditional MMR redundancy removal algorithm based on literal similarity (Carbonell and Goldstein, 1998) since we find only ignorable differences between using MMR or not for all systems.

4.4 Evaluation Methods and Metrics

4.4.1 Automatic Evaluation

Similar to the evaluation for traditional summarization tasks, we use the ROUGE metrics (Lin and Hovy, 2003) to automatically evaluate the quality of produced summaries given the gold-standard reference news. The ROUGE metrics measure summary quality by counting the precision, recall and F-score of overlapping units, such as n-grams and skip grams, between a candidate summary and the reference summaries.

We use the ROUGE-1.5.5 toolkit to perform the evaluation. In this paper we report the F-scores of the following metrics in the experimental results: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based) and ROUGE-SU4 (based on skip bigrams with a maximum skip distance of 4).

4.4.2 Pyramid Evaluation

We also conduct manual pyramid evaluation in this study. Specifically, we use the modified pyramid scores as described in (Passonneau et al., 2005) to manually evaluate the summaries generated by different methods. We randomly sample 20 games from the data set and manually annotate facts on the gold-standard news. The annotated facts are mostly describing specific events happened during the game, e.g. "伊万被黄牌警 J" (Ivanovic is shown the yellow card) and "内马尔开出角球" (Neymar takes the corner). Each fact is treated as a Summarization Content Unit, (SCU) (Nenkova and Passonneau, 2004). The number of occurrences for each SCU in the gold-standard news is regarded as the weight of this SCU.

5 Results and Analysis

5.1 Comparison with Baseline Methods

The average performance on all three folds of different methods are displayed in Table 1.

Method	R-1	R-2	R-SU4
HeadTail	0.30147	0.07779	0.10336
Centroid	0.32508	0.08113	0.11245
LexRank	0.31284	0.06159	0.09376
ILP	0.32552	0.07285	0.10378
Highlight	0.34687	0.08748	0.11924
RF	0.38559	0.11887	0.14907
RF+DPP	0.39391	0.11986	0.15097

Table 1: Comparison results of different methods

As we can see from the results, our learning to rank approach based on RF achieves significantly (< 0.01 significance level for pairwise-t testing) better results compared with traditional unsupervised summarization approaches [11]. The ILP model, which is believed to be suitable for multi-document summarization, did not perform well in our settings. Head and tail sentences are informative but merely using them lacks specific descriptions for procedural events, therefore not

[10] We also tried words rather than bigrams but found slightly worse performance.

[11] We also conducted experiments on using our proposed features to calculate LexRank, but did not observe real difference compared with normal LexRank. This suggest that the performance gain comes from supervised learning to rank approach, not merely from the features.

providing competitive results either.

The comparison between RF and RF+DPP shows the effectiveness of our sentence selection strategy. However, the increase is still limited [12]. This may become reasonable later when we discuss more about the errors from our systems.

Merely using highlight markers to construct news also provides competitive results, but inferior to supervised models. This suggests that the highlight marker features are relatively strong indicators for good sentences while merely using these features may not be sufficient.

Table 2 shows the average pyramid scores for the systems in comparison. The "Gold-standard" row denotes manually written news report and is listed for reference. We can see our learning to rank systems based on RF constructs news with the highest pyramid scores.

Method	Pyramid scores
HeadTail	0.13657
Centroid	0.30663
LexRank	0.28756
ILP	0.20867
Highlight	0.41121
RF	0.53766
RF+DPP	0.62500
Gold-standard	0.88329

Table 2: Average Pyramid scores

Overall, the experimental results indicate that our system can generate much better news than the baselines in both automatic and manual evaluations. We include examples of our constructed news reports in the supplementary materials.

5.2 Feature Validation

Different groups of features may play different roles in the LTR models. In order to validate the impact of both the traditional features and the novel task-specific features, we conduct experiments with different combinations by removing each group of features respectively. Table 3 shows the results, with "w/o" denotes experiments without the corresponding group of features.

Method	R-1	R-2	R-SU4
RF	0.38559	0.11887	0.14907
RF-w/o novel	0.37297	0.10964	0.14021
RF-w/o trad.	0.36314	0.09910	0.13102

Table 3: Results of feature validation

[12] Significance level < 0.05 for pairwise-t testing only for ROUGE-1.

We can observe that both the traditional features and the novel features contribute useful information for learning to rank models. Due to the nature of the sentence extraction approach, features designed for traditional document summarization are still playing an indispensable role for our task, although they might be important in this work for different reasons. For example, position features are indicative for traditional summarization since sentences appearing in the very beginning or the end are more probable as summarizing sentences. For sports commentary, positions are closely related to timeline in a more coarse fashion. Certain types of key events, for example player substitutions and even scores, may tend to happen in certain period in a game rather than uniformly spread out in every minute.

5.3 Room for Improvements

5.3.1 Upper Bounds

To get a rough estimate of what is actually achievable in terms of the final ROUGE scores, we looked at different "upper bounds" under various scenarios (Table 4). We first evaluate one *reference* news with the other reference news served as the gold-standard result. The results are given in the row labeled *reference* of Table 4. This provides a reasonable estimate of human performance.

Second, in sentence extraction we restrict the constructed news to sentences from the original commentary texts themselves. We use the greedy algorithm to *extract* sentences that maximize ROUGE-2F scores. The resulting performance is given in the row *extract* of Table 4. We observe numerically superior scores compared with *reference*. This is not strange since we are intentionally optimizing ROUGE scores. And also this suggests that the sentence extraction approach for sports news construction is rather reasonable, in terms of information overlap.

Method	R-1	R-2	R-SU4
reference	0.44725	0.15265	0.18064
extract	0.43270	0.16872	0.18622
target	0.40987	0.15901	0.17941
target+DPP	0.41536	0.15994	0.18232
RF+DPP	0.39391	0.11986	0.15097

Table 4: Upper bounds on ROUGE scores

Third, we use the partial ROUGE-2 values, i.e. the targets used to train LTR models (cf. Section 3.1) for greedy selection and DPP selection, with results listed in the row *target* and *tar-*

Time	Live Text Commentary Script
55	内马尔 为 巴 萨 制造 了 一个 位置 不错 的 定位球 Neymar wins a free kick in a good position.
56	内马尔 ~~~ Neymar!!!!!!!
56	内马尔 的 定位球 直接 打入 了 球门 左上 角 的 死角 ！ ！ ！ 门将 无能为力 The free kick from Neymar goes directly into the top left corner! The keeper can do nothing.
56	球 在 飞 向 球门 的 过程 中 下坠 速度 非常 快 The ball drops quickly and flies to the goal.

Figure 4: Case I: short and noisy sentences

Time	Live Text Commentary Script
FT	切尔西中场快发任意球，科斯塔打进全场唯一入球！！！ From a quick free kick from Chelsea, Costa scored the only goal of the game!!!
FT	整场比赛切尔西占据了主动，可面对诺维奇的铁桶阵，办法不多 Chelsea dominated the game but found it difficult against Norwich's defense.
FT	下半场利用对方的一次疏忽，阿扎尔中场被放倒，威廉快发任意球，科斯塔完成致命一击 From an error from the opponent, Hazard was fouled. Willian launches a quick free kick and assists Costa for the lethal strike.

Figure 5: Case II: summarizing sentences

get+DPP of Table 4. This validates that using partial ROUGE-2 as the training target for LTR models is somewhat reasonable for this study.

5.3.2 Error Analysis

In this preliminary study, we use LTR models and probabilistic sentence selection procedure. While reasonable performance has been achieved, there exist certain types of errors as we found in the constructed news results.

Error I: First, sentences in live commentary are mostly short, and sometimes noisy. Sometimes an important event has been described using a number of consecutive short sentences. Our LTR models failed to generate high scores for such sentences and therefore will cause some lack of information. Figure 4 illustrates an example of this type of error in the constructed news report. All the sentences are describing a key scoring event. However, none of them were selected to construct the news because our LTR model assigns low scores for these short sentences. Meanwhile the second sentence can be treated as noisy.

Error II: Second, commentators are likely to summarize important events during the game, not at the point when the event happens. Our sentence selection algorithm can only address local redundancy, while this issue is more global. Figure 5 illustrates an example of this case in the constructed news report. The only goal of the match is described during full-time (FT). Our method redundantly included this in the final constructed news even it had already selected that event.

These two issues are highly non-trivial and have not been well addressed in the method we explored in this paper. We leave them for further study in the future.

5.3.3 Readability Assessment

In this work we only consider sentence extraction. Unlike traditional summarization tasks, sports commentary texts are describing a different specific action in almost every sentence. Descriptive coherence becomes a more difficult challenge in this scenario. We conduct manual evaluation on systems in comparison along with manually written news reports (gold-standard). Three volunteers who are fluent in Chinese were asked to perform manual ratings on three factors: coherence (Coh.), non-redundancy (NR) and overall readability (Read.). The ratings are in the format of 1-5 numerical scores (not necessarily integral), with higher scores denote better quality. The results are shown in Table 5.

Method	Coh.	NR	Read.
HeadTail	3.47	3.07	3.56
Centroid	2.87	3.72	2.66
LexRank	2.90	3.23	2.43
ILP	2.87	3.23	2.50
Highlight	3.36	3.72	3.06
RF	3.23	3.64	3.13
RF+DPP	3.23	3.87	3.06
Gold-Standard	4.67	4.23	4.77

Table 5: Manual readability ratings

The differences between systems in terms of readability factors are not as large as information coverage suggested by ROUGE metrics and pyramid scores. Meanwhile, while we can observe that our approaches outperforms the unsupervised extractive summarization approaches in coherence and readability for certain level, the results also clearly suggest that there still exists large room for improvements in terms of the readability factors.

6 Discussions

The general challenges for the particular task of sports news generation are mostly addressed in those designed features in the learning to rank framework. We utilize the timeline and scoreline information, while also keep traditional features such as sentence length. Experimental results show that our framework indeed outperforms strong traditional summarization baselines, while still having much room for improvement.

We might also notice that there may exist some issues if merely using automatic metrics to evaluate the overall quality of the generated news reports. The ROUGE metrics are mainly based on

ngram overlaps. For sports texts most of the proportions are dominated by proper names, certain types of actions or key events, etc. Compared with traditional summarization tasks, it might be easier to achieve high ROUGE scores with an emphasize on selecting important entities. In our experiments, methods with higher ROUGE scores can indeed achieve better coverage of important units such as events, as shown in pyramid scores in Table 2. However, we can also observe from Table 5 that automatic metrics currently cannot reflect readability factors very well. Generally speaking, while big difference in ROUGE may suggest big difference in overall quality, smaller ROUGE differences may not be that indicative enough. Therefore, it is interesting to find alternative automatic metrics in order to better reflect the general quality for this task.

7 Related Work

To the best of our knowledge, generation of sports news from live text commentary is not a well-studied task in related fields. One related study focused on generating textual summaries for sports events from status updates in Twitter (Nichols et al., 2012). There also exists earlier work on generation of sports highlight frames from sports videos, focusing on a very different type of data (Tjondronegoro et al., 2004). Bouayad-Agha et al. (2011) and Bouayad-Agha et al. (2012) constructed an ontology-based knowledge base for the generation of football summaries, using predefined extraction templates.

Our task is closely related to document summarization, which has been studied quite intensively. Various approaches exist to challenge the document summarization task, including centroid-based methods, link analysis and graph-based algorithms (Erkan and Radev, 2004; Wan et al., 2007), combinatorial optimization techniques such as integer linear programming (Gillick et al., 2008) and submodular optimization (Lin and Bilmes, 2010). Supervised models including learning to rank models (Metzler and Kanungo, 2008; Shen and Li, 2011; Wang et al., 2013) and regression (Ouyang et al., 2007; Galanis and Malakasiotis, 2008; Hong and Nenkova, 2014) have also been adapted in the scenario of document summarization.

Since sports live texts contain timeline information, summarization paradigms that utilize timeline and temporal information (Yan et al., 2011; Ng et al., 2014; Li et al., 2015) are also conceptually related. Supervised approaches related to this work have also been applied for timeline summarization, including linear regression for important scores (Tran et al., 2013a) and learning to rank models (Tran et al., 2013b). In this preliminary work we only use the timestamps in the definition of similarity for sentence selection. More crafted usages will be explored in the future.

8 Conclusion and Future Work

In this paper we study a challenging task to automatically construct sports news from live text commentary. Using football live texts as an instance, we collect training data jointly from live text commentary services and sports news portals. We develop a system based on learning to rank models, with several novel task-specific features. To generate the final news summary and tackle the local redundancy problem, we also propose a probabilistic sentence selection method. Experimental results demostrate that this task is feasible and our proposed methods are appropriate.

As a preliminary work, we only perform sentence extraction in this work. Since sports news and live commentary are in different genres, some post-editing rewritings will make the system generating more natural descriptions for sports news. We would like to extend our system to produce sports news beyond pure sentence extraction.

Another important direction is to focus on the construction of datasets in larger scale. One feasible approach is to use a speech recognition system on live videos or broadcasts of sports games to collect huge amount of transcripts as our raw data source. Although more data can be easily collected in this case, the noisiness of audio transcripts may bring some additional challenges, therefore worthwhile for further study.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61331011), National Hi-Tech Research and Development Program (863 Program) of China (2015AA015403) and IBM Global Faculty Award Program. We thank the anonymous reviewers for helpful comments and Kui Xu from our group for his help in calculating player popularity features. Xiaojun Wan is the corresponding author of this paper.

References

Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. 2011. Content selection from an ontology-based knowledge base for the generation of football summaries. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 72–81. Association for Computational Linguistics.

Nadjet Bouayad-Agha, Gerard Casamayor, Simon Mille, and Leo Wanner. 2012. Perspective-oriented generation of football match summaries: Old tasks, new challenges. *ACM Transactions on Speech and Language Processing (TSLP)*, 9(2):3.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479.

Dimitrios Galanis and Prodromos Malakasiotis. 2008. Aueb at tac 2008. In *Proceedings of the TAC 2008 Workshop*.

Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. 2012. Near-optimal map inference for determinantal point processes. In *Advances in Neural Information Processing Systems*, pages 2735–2743.

Dan Gillick, Benoit Favre, and Dilek Hakkani-Tur. 2008. The icsi summarization system at tac 2008. In *Proceedings of the Text Understanding Conference*.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721, Gothenburg, Sweden, April. Association for Computational Linguistics.

Alex Kulesza and Ben Taskar. 2011. Learning determinantal point processes. In *UAI*.

Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3).

Chen Li, Yang Liu, and Lin Zhao. 2015. Improving update summarization via supervised ilp and sentence reranking. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1317–1322, Denver, Colorado, May–June. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics.

Donald Metzler and Tapas Kanungo. 2008. Machine learned sentence selection strategies for query-biased summarization. In *SIGIR Learning to Rank Workshop*, pages 40–47.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method.

Jun-Ping Ng, Yan Chen, Min-Yen Kan, and Zhoujun Li. 2014. Exploiting timelines to enhance multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–933, Baltimore, Maryland, June. Association for Computational Linguistics.

Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 189–198. ACM.

You Ouyang, Sujian Li, and Wenjie Li. 2007. Developing learning strategies for topic-based summarization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 79–86. ACM.

Rebecca J Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. 2005. Applying the pyramid method in duc 2005. In *Proceedings of the Document Understanding Conference (DUC 05), Vancouver, BC, Canada*.

Dragomir R Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pages 21–30. Association for Computational Linguistics.

Chao Shen and Tao Li. 2011. Learning to rank for query-focused multi-document summarization. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 626–634. IEEE.

Dian Tjondronegoro, Yi-Ping Phoebe Chen, and Binh Pham. 2004. Integrating highlights for more complete sports video summarization. *IEEE multimedia*, 11(4):22–37.

Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. 2013a. Predicting relevant news events for timeline summaries. In *Proceedings of the 22nd international conference on World Wide Web Companion*, pages 91–92. International World Wide Web Conferences Steering Committee.

Giang Binh Tran, Tuan A Tran, Nam-Khanh Tran, Mohammad Alrifai, and Nattiya Kanhabua. 2013b. Leveraging learning to rank in an optimization framework for timeline summarization. In *SIGIR 2013 Workshop on Time-aware Information Access (TAIA*.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In *IJCAI*, volume 7, pages 2903–2908.

Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie. 2013. A sentence compression based framework to query-focused multi-document summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1384–1394, Sofia, Bulgaria, August. Association for Computational Linguistics.

Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270.

Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. 2011. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 433–443, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.