Automatic Summarization of Football Commentary Data

Anand Kamat
McGill ID: 260773313
anand.kamat@mail.mcgill.ca

Deeksha Arya
McGill ID: 260786737
deeksha.arya@mail.mcgill.ca

Abstract

Text summarization is considered one of the most active research areas with applications for both academics and industry. Extractive text summarization, in particular, is extensively used for summarizing news reports and reviews of products. In this report, we explore the ability to summarize football commentary data using multiple extractive techniques. We assess the quality of the summaries produced and attempt to conclude which methodologies are feasible and worth pursuing for this field.

1 Introduction

Text summarization systems produce concise information from a source document, upon which the user can determine the gist of the document without actually reading it(Lloret, 2008). They can be classified into extractive and abstractive summarization method.

An extractive summarization method involves collecting and selecting important sentences from the original document based on statistical and linguistic features, to generate summaries into a shorter form. Identifying the most important sentences relies on the concept of sentence salience, which is typically defined in terms of the presence of particular important words or in terms of similarity to a centroid pseudo-sentence(Allahyari et al., 2017). In this paper we explore several extractive text summarization algorithms such as LSA, LexRank, TextRank, Luhn, SumBasic and KLSum and use them to analyze football commentary data.

Association football, also known as soccer or football, is the widest played sport in the world today with about 250 million players in over 200 countries and a worldwide fan base of over a staggering 1.3 billion people as reported by FIFA. With its incredible outreach across the world, we decided to study the match commentary data. Commentary data contains minute by minute updates of the match which when summarized efficiently can provide the highlights of the game.

2 Related Work

Research in the automatic summarization of football commentary data is one that has been given very little attention. (Zhang et al., 2016) performed a study which demonstrated that sports news could be generated from text commentary. The paper describes a supervised ranking algorithm which has the ability to learn the particular sentences that should be extracted. (Nichols et al., 2012), in their paper, developed an algorithm that identifies important tweets regarding the Soccer World Cup and extracts them to produce a summary of the event. To assess the summaries produced, they used the ROUGE-n metric as well as human evaluations and concluded that their algorithm produces reasonable event summaries comprising of a few paragraphs. In both these papers, new algorithms have been proposed, however there are no measures about the performance of existing standard extractive techniques on commentary data.

In addition, football commentary can be compared in terms of its linguistics to meeting transcripts as the data is in present tense and contains minute-wise information about the on-goings of the respective event. The paper by (Murray et al., 2005) discusses evaluation techniques of different summarization algorithms on meeting transcripts. In our paper, we aim to perform a similar evaluation on minute-by-minute football commentary scripts.

3 Implementation Method

We divide our implementation methodology into 3 parts:

3.1 Data Collection and Preprocessing

Football commentary data used to analyze the summarization algorithms was obtained from SkySports, the popular and well-known sports news reporting agency. The data was retrieved in the same way it was published on the website - with the beginning of the match at the end of the page, and in chronological order as you move up the page. In order to make the commentary usable for an efficient summarization task, it was necessary to preprocess this text. A number of preprocessing steps were taken.

Removal of noisy sentences: It was observed that many advertisements were placed inline in the commentary data. Hence, a set of word phrases to be removed, such as *watch* and *click here* were identified, and any sentences containing these phrases were removed.

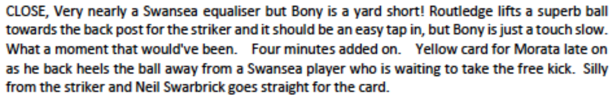
Removal of current time and score: The current time and the score was contained in the text (eg. 11:48 BRIGHTON 1-5 LIVERPOOL). As these do not contribute to the content they were removed from the text.

Reversing the commentary: Simple reversal of commentary could not be done for this data. This is because though the minute-level commentary was in descending order of time, the commentary at each minute could be composed of more than one phrase or sentence, which were in the correct order. Hence only minute blocks of the commentary were reversed, and not individual lines in the text. This was done by tokenizing minute-level collections of sentences before reversing the text.

Removing the minute of the commentary: After reversal of the commentary, the minutes of the match (eg. 89:, 90:, etc) were removed from the commentary content.

Merging keywords with next sentence: In order to ensure that the keywords that formed the sentences by themselves in the commentary, were not overlooked, they were merged with the sentence that followed them. These keywords are a set of common exclamations made by the commentators such as *GOAL!*,*YELLOW!*, and *CLOSE!*. The assumption for the same was that the sentence that followed the exclamation was related to the exclamation made.

As a result of these preprocessing tasks, a commentary data upon which effective automatic summarization could be attempted was produced. A snippet of the same is shown below.



CLOSE, Very nearly a Swansea equaliser but Bony is a yard short! Routledge lifts a superb ball towards the back post for the striker and it should be an easy tap in, but Bony is just a touch slow. What a moment that would've been. Four minutes added on. Yellow card for Morata late on as he back heels the ball away from a Swansea player who is waiting to take the free kick. Silly from the striker and Neil Swarbrick goes straight for the card.

3.2 Auto Summarization of the Commentary Data

To implement automatic summarization of commentary data, we study extractive summarization. We use 6 well-known summarization algorithms namely: *Luhn, SumBasic, KLSum, Latent Semantic Analysis (LSA), LexRank, TextRank*. These algorithms have been discussed in detail in Section 4. Summarization of the commentary data using the above-mentioned algorithms was implemented in Python using the *sumy* library.

The Plaintext parser implemented by the library is used to read the commentary from an input file. During the summarization process, stemming is done using *Snowball* from the *nltk* library ignoring the stop words. These suggestions have been made by *sumy* authors. For all summarization techniques used, a summary with 4 sentences was generated.

In addition to these algorithms, we implemented a baseline algorithm that creates a summary by picking the leading 2 and ending 2 sentences of the commentary data. This was done upon analysis of the commentary data, where it was found that in most cases, the initial 2 sentences introduce the participant teams and the final 2 sentences conclude the game results, hence providing a fair summarization of the match.

3.3 Evaluation of Summaries Generated

In order to determine the most qualitative summarization technique, we assess the summaries produced using intrinsic evaluation metrics(Steinberger and Ježek, 2012). The following methods are known as Content-Based techniques as they focus on just the composition of text in the summary and not its grammar, structure, coherence or clarity. Content-Based measures focus on the similarity between sentences. They are usually at token or word level, and compare the existence of words in two sentences. *Cosine Similarity, Unit Overlap* and *Rouge measures* were used for evaluation.

The theory behind these evaluation metrics are discussed briefly in Section 5. To determine the metric values, we used the evaluation methods available in the *sumy* library. In addition, we performed a Text Quality Evaluation, i.e. a critical human analysis of the summaries assessing the grammaticality, non-redundancy, clarity and coherence of the summaries produced.

The three phases explained were performed on live minute-by-minute commentaries of 10 different matches.

4 Automatic Summarization Models

In this paper we have experimented with the following automatic text summarization algorithms -

Luhn Algorithm: H.P Luhn in 1958 proposed a theory for predicting the topic of published papers (Luhn, 1958). The theory prescribes a *significance factor* of sentences which determines the relative significance of sentences or words. When words with highest frequency occur in greatest physical proximity to each other, the probability is very high that the information being conveyed is most representative of the article. The significance factor is then used to score the sentences and to rank the sentences to 'auto-abstract' the document.

SumBasic: The SumBasic algorithm,(Nenkova and Vanderwende, 2005) is designed with the idea that the relative frequency of a non-stop word in a document set is a good predictor of a word appearing in a human annotated summary. Each sentence is then assigned a score which reflects how many high frequency words it contains. The summary is built by adding the highest scoring sentences progressively. In order to discourage redundancy, the probabilities of words are squared, thereby drastically reducing their probabilities.

KLSum: KLSum is usually used as an extractive multi-document text summarization algorithm. Kullback-Lieber(KL) divergence represents the true distribution (the document set unigram distribution) and approximate distribution (the summary distribution). This criterion casts summarization as finding a set of summary sentences which closely match the document set unigram distribution.

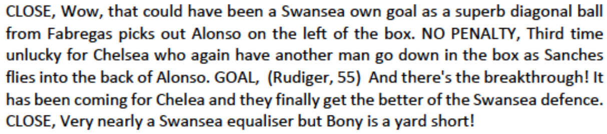
Latent Semantic Analysis (LSA): LSA is a method based on statistical calculations to extract and represent the contextual meaning of words and the similarity of sentences (Badry et al., 2013). The intuition of the model is that if the number of common words between sentences is high, the sentences are more semantically related. LSA uses Singular Value Decomposition (SVD) to identify patterns in relationships between sentences and also determining the similarity of meaning of words and sentences.

LexRank: The LexRank Algorithm assesses the centrality of the sentence in a cluster and extracts the most important ones to include in the summary (Erkan and Radev, 2004). The algorithm defines a graph representation of a document cluster. This idea is applied to extractive summarization where, to compute centrality, each edge (similarity) of the graph is treated as a vote to determine the overall centrality value of each node. The higher the number of votes that are cast for a vertex, the higher the importance of the vertex.

TextRank: TextRank, like LexRank, uses a graph based ranking model for text processing and extractive summarization (Mihalcea and Tarau, 2004). The premise of TextRank algorithm is very similar to the LexRank algorithm differing mainly in the methods used to score and finding the similarity between the two sentences in a cluster. While the LexRank algorithm uses cosine similarity to compute how similar two sentences in cluster are, the TextRank algorithm uses a more elaborate formula for the same.

5 Evaluation Criteria

We focus on the intrinsic evaluation techniques in which a generated summary is measured against a gold standard. We requested three human annotators to jointly create 4-sentence extractive summaries for each commentary as the gold standard to be compared with auto-generated summaries. A sample golden summary is shown below.



CLOSE, Wow, that could have been a Swansea own goal as a superb diagonal ball from Fabregas picks out Alonso on the left of the box. NO PENALTY, Third time unlucky for Chelsea who again have another man go down in the box as Sanches flies into the back of Alonso. GOAL, (Rudiger, 55) And there's the breakthrough! It has been coming for Chelea and they finally get the better of the Swansea defence. CLOSE, Very nearly a Swansea equaliser but Bony is a yard short!

To evaluate the quality of the extractive summaries produced with respect to a reference summary, we used different evaluation methods described below.

Cosine similarity: Cosine Similarity measures the relatedness of two texts by calculating the dot product of their vector representations divided by the product of their norms. For our evaluation purposes, we measure the cosine similarity of the extractive summary with the reference summary as well as with respect to the original commentary data.

Unit Overlap: Unit overlap measures the number of overlapping words between the two texts. For our evaluation purposes, we measure unit overlap of the extractive summary with the reference summary as well as with respect to the original commentary data.

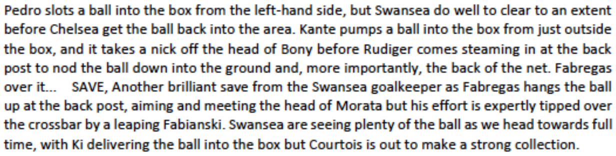
ROUGE: We use the following ROUGE scores to evaluate the summaries.

ROUGE-n: measures the overlap of n-grams between the generated and the reference summaries. For our project, we use the Rouge-1 and Rouge-2 metric for evaluation.

ROUGE-L: measures the longest matching sequence (consecutive or non-consecutive) of words (Longest Common Subsequence - LCS) between the generated summary and the reference summary.

6 Results

As mentioned before, we summarized and evaluated live minute-by-minute commentaries of 10 different matches. The following text is a sample summary generated by the LSA algorithm.



Pedro slots a ball into the box from the left-hand side, but Swansea do well to clear to an extent before Chelsea get the ball back into the area. Kante pumps a ball into the box from just outside the box, and it takes a nick off the head of Bony before Rudiger comes steaming in at the back post to nod the ball down into the ground and, more importantly, the back of the net. Fabregas over it... SAVE, Another brilliant save from the Swansea goalkeeper as Fabregas hangs the ball up at the back post, aiming and meeting the head of Morata but his effort is expertly tipped over the crossbar by a leaping Fabianski. Swansea are seeing plenty of the ball as we head towards full time, with Ki delivering the ball into the box but Courtois is out to make a strong collection.

Table 1 shows the metrics obtained from the summarization of one of the matches.

The baseline performs the poorest in terms of content similarity with respect to the gold standard. Its summary does not include important events during the match such as a goal being scored or a player being sent off the field.

SumBasic, due to its simplistic structure and tendency to favor repetition of frequent words despite the redundancy update leads to a poor performance. The summary itself is highly incoherent, and simply seems to state facts either about one of the teams (as each team name is usually a high frequency word) or retrieves sentences in which an important event, such as a goal has occurred. Very little other content exists in the summary. It also seems to focus on shorter sentences, hence omitting important details.

The Luhn algorithm retrieves moderately relevant sentences from the commentary. It is interesting to note that it seems to focus on more detailed sentences, where the number of common words is less in comparison to the number of rarer and football-exclusive lingo. Because of this, upon reading the summary, the impression given is that it is impressive, though not necessarily the best summary that could be produced. These thoughts are reflected statistically as Luhns significance factor enabled the algorithm to perform quite well in terms of cosine similarity, unit overlap and ROUGE-1.

We observed that the LSA scores are consistently high across all the evaluation metrics. This can be attributed to the use of SVD which captures the relationships between sentences and the meaning of words and sentences. LSA has the highest ROUGE-1 score of 0.415385. The LSA fails to maintain its ranks in ROUGE-2 and ROUGE-L scores, which is due to the selection algorithm implemented. We found this algorithm to produce a summary very similar to that of Luhn's algorithm. For many commentaries, there was overlap between sentences in the summaries generated in both these cases.

KLSum is another algorithm which showed con-

Table 1: Evaluation of different algorithms on a sample football commentary

Algorithm	Cosine similarity	Cosine similarity (document)	Unit overlap	Unit overlap (document)	ROUGE-1	ROUGE-2	ROUGE-L (Sentence Level)
Baseline	0.490598	0.657276	0.183486	0.119349	0.276923	0.064935	0.142002
Luhn	0.650815	0.857320	0.203252	0.153707	0.353846	0.077922	0.170354
SumBasic	0.550104	0.819752	0.142857	0.074141	0.200000	0.051948	0.157434
LSA	0.662919	0.889069	0.210526	0.177215	0.415385	0.090909	0.124964
LexRank	0.515115	0.758461	0.179487	0.135624	0.323077	0.051948	0.142149
TextRank	0.684199	0.883674	0.198347	0.148282	0.338462	0.090909	0.176167
KL	0.662746	0.875009	0.189655	0.135624	0.323077	0.129870	0.183726

sistently high performance. We can observe KLSum having the highest ROUGE-2 and ROUGE-L score of 0.12987 and 0.183726 respectively. The KLSum criterion of greedily selecting sentences as long as they decrease KL-divergence explains these scores.

The graph based algorithms, LexRank and TextRank show promising results. We can observe the TextRank algorithm outperforms the LexRank algorithm in all the evaluation metrics. The vertex scoring algorithm and the similarity formulas used by TextRank are superior to those used by the LexRank algorithm and better model a human connotation, hence explaining their performances. The highest cosine similarity score is given by the TextRank algorithm. Though different in terms of performance, the two algorithms were found to be similar in structure. It was interesting to note that these algorithms extracted sentences that were highly action oriented and contained phrases like *it flicks off*, and *pumps a ball into the box*. These algorithms were found to be very similar to those produced by the LSA, KLSum and Luhn algorithms, focusing on longer, detailed sentences.

Likely due to the content of the original commentary itself, there was less redundancy or repetition in any of the summaries produced. All the algorithms studied, in general, outperform standard baseline and random algorithms in both evaluation metrics as well as in terms of text quality but do not sufficiently capture the best representative sentences as modeled by the gold summary.

7 Discussion and Conclusion

Summarization of commentary data for a match should ideally provide a brief highlight of the most important events in the game along with the final score of the match. These highlights are not usually repeated through the commentary. Since most generic summarization algorithms are based on the premise that certain words/topics would be mentioned more than the other indicating their importance, they produce only moderately performing summaries. Incorporating additional data such as content context and sentiment scores should also be considered to optimize summarization algorithms.

The evaluation criteria used, though a good judge of the content resemblance between the generated summary and the golden summary, may not be the best approach to evaluation. ROUGE functions, for example, are based on the assumption that in order for a summary to be of high quality, it has to share many words or phrases with a human golden summary which in itself is prone to bias. A better approach to evaluate summarization models can be based on a premise that concepts take meanings from the context they are in, and that related concepts co-occur frequently.

In conclusion, it is evident that there is a need to develop efficient and accurate summarization models with its growing need. Generic extractive summarization algorithms may not be the solution to all kinds of texts as observed in this paper. Evaluation criteria used to analyze the summarization models fail to consider important attributes such as context which can affect its bias to the results.

8 Statement of Contributions

This project was the result of the combined effort of both the authors. Deeksha worked on retrieving and preprocessing the commentary data and using different evaluation metrics to determine the quality of the summaries produced. Anand implemented each of the summarization algorithms on the preprocessed data. He also performed an unbiased critical analysis of the extractive summaries produced. Both team members took responsibility of writing their contributions in the paper.

References

[Allahyari et al.2017] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: A brief survey. *arXiv preprint arXiv:1707.02268*.

[Badry et al.2013] Rasha Mohammed Badry, Ahmed Sharaf Eldin, and Doaa Saad Elzanfally. 2013. Text summarization within the latent semantic analysis framework: comparative study. *International Journal of Computer Applications*, 81(11).

[Erkan and Radev2004] Gunes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

[Lloret2008] Elena Lloret. 2008. Text summarization: an overview. *Paper supported by the Spanish Government under the project TEXT-MESS (TIN2006-15265-C06-01)*.

[Luhn1958] Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

[Mihalcea and Tarau2004] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

[Murray et al.2005] Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2005. Evaluating automatic summaries of meeting recordings.

[Nenkova and Vanderwende2005] Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, 101.

[Nichols et al.2012] 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.

[Steinberger and Ježek2012] Josef Steinberger and Karel Ježek. 2012. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.

[Zhang et al.2016] Jianmin Zhang, Jin-ge Yao, and Xiaojun Wan. 2016. Towards constructing sports news from live text commentary. In *ACL (1)*.