Generating Football Match Summaries with NMT

Miguel Ayala
Department of Computer Science
Stanford University
mayala3@stanford.edu

Abstract

In sports journalism, football (soccer) match summaries are handwritten, despite their formulaic nature. There seem to be major patterns that govern this style of writing. Given the improvement in text generation algorithms, this study aims to see whether or not football match summaries can be produced with neural network models. An NMT framework is used to generate textual summaries from game event vectors. For this reason and several others, BLEU scores remain low. While semi-coherent text is produced by the neural network, it lacks specificity regarding match in question. What is promising, however, is that the generated text seems to be salient to some degree. In other words, if the game event vector described a goal, then the generated summary was often about a goal-scoring chance. Future work will require specific entity handling to ensure that generated sentences are effective.

1 Introduction

1.1 The Problem

With so much demand for sports content in this age of digital consumption, there have been numerous attempts to procedurally generate summaries of sporting events. Football (soccer) still relies on journalists to cover the thousands of matches that happen every week. Demand for instant summaries is incredibly high, however.

1.2 The Goal

The goal of this study is to produce a neural network that can produce instant textual summaries of football matches. More concretely, the objective is to take a vector of comprehensive match event data and generate a coherent narrative that mentions only the salient events in a fluid manner.

1.3 Related Work

Different approaches seem to work for other sports. Successful examples include the WordSmith program which, among other things, produces basketball game summaries for the Associated Press [7]. As of 2019, there is no such parallel for football (soccer) matches. Nevertheless, there have been attempts over the past 3 decades. Each approach is fundamentally different. The input data, for instance, is incredibly varied. André et al tried to transform visual images of football matches into text as far back as 1988 [1]. In 1998, abstracted representations were being used for soccer commentary generation [3]. In recent years, however, many studies have focused on using game statistics to create match summaries[4] [5]. These approaches are still largely rule-based, however, and the summaries produced fail to capture the fluidity and flair of human written summaries.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

One framework that might be able to provide this level of linguistic flexibility is the NMT [9]. NMT models are largely based on the encoder-decoder paradigm [6] [14]. This is in contrast to most traditional, phrase translation models that consist of many individual sub-components that have to be trained and tuned separately. In NMT, the encoding phase consists of a bidirectional Recurrent Neural Network that encodes a vector representation of the source sentence. The decoder then takes this encoder output to produce a target sentence. This relatively simple process now underpins some of the most effective translation services currently available[15] [8]. NMT combines the strengths of word embeddings with memory models to create a system that considers entire sentences and learns deep links between words.

Most of the current research still lies around the original task of translation. Different institutions are tweaking the various components of the basic encoder-decoder model. Some are trying to work on problems like the fixed vector length problem [2] or the vanishing/exploding gradient problem. Other optimization comes in the various attention mechanisms that can be used to condense or remember lengthy source vectors [10]. While some applications exist, there is still potential for novel NMT experimentation. This problem may be one such use case.

2 Approach

2.1 The NMT Approach

The task of producing match summaries from statistics is not explicitly a translation task. It does not map strings to strings. Instead it maps numbers to strings. The NMT is flexible enough to handle this, however. Instead of passing a word-embedding representation of a source sentence to the decoder, we pass a mix of vectorized match statistics and word-embeddings, with appropriate padding to ensure uniform length. To be more specific, our input is a flattened vector consisting of event vectors [1].

2.2 Architecture

The NMT system's architecture is inspired by one built for Winter 2019 Natural Language Processing with Deep Learning at Stanford (see Figure 1). The system consists of a bidirectional LSTM encoder and a unidirectional LSTM decoder.Words in a source sentence are converted into its embeddings. These are passed through to the encoder in exchange for hidden and cell states. The decoder's LSTM is initialized with hidden and cell states from this process. Multiplicative attention is used in conjunction with a linear layer, a tanh layer and a dropout layer to produce a combined output vector. From this vector we can produce a probability distribution of likely target words. From this, we can generate likely sentences.

[1] see Data for more details

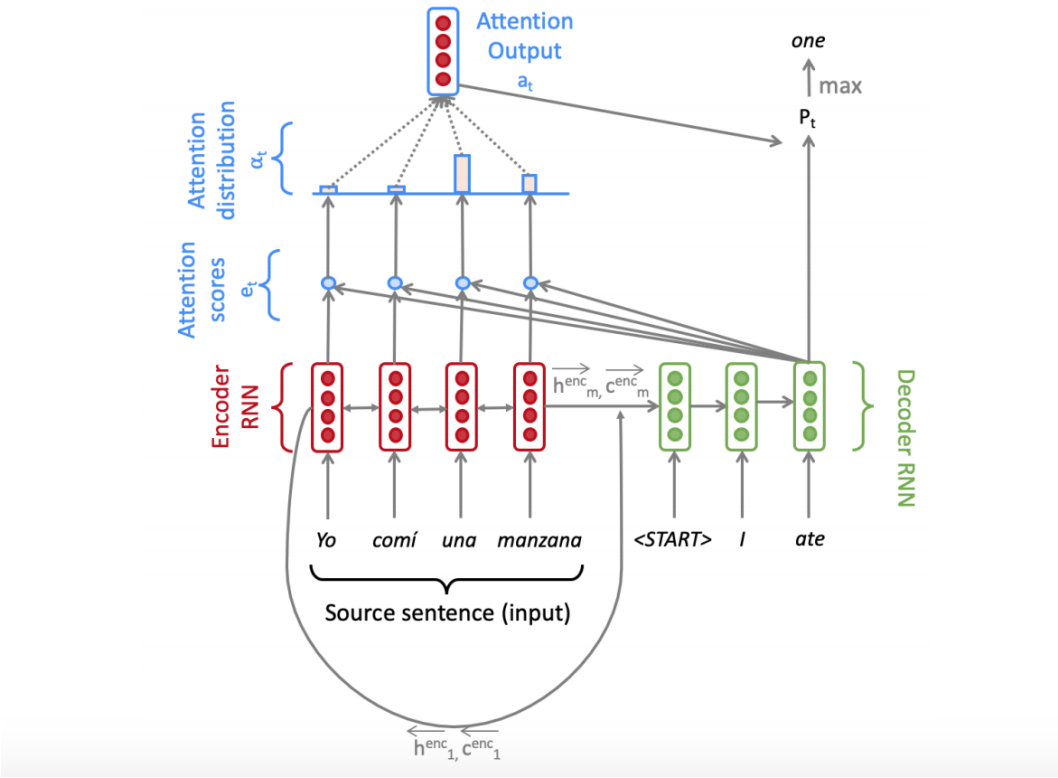

Figure 1: Architectural Diagram

3 Data

3.1 Data Collection

The dataset is bipartite. It consists of source vectors and targets. Each was collected from a different source. A variety of preprocessing tools were built to get the data into usable format and then to form associations between source sentences and target sentences.

3.1.1 Source Vectors

The source vectors come from a curated dataset [12] that details major football events in matches from the 5 biggest European leagues: The English Premier League, La Liga, Bundesliga, Serie A and Ligue 1. In contrast to many football match datasets, each datum from the Football Events dataset consists of far more than just the typical goals, red cards, yellow cards, etc. The data for each match is broken down into key events that are quantified by a series of variables such as $shot\_outcome$ and $time$. Here is the full list of categories:

[*"time", "event_type", "event_type2", "side", "event_team", "opponent", "player", "player2", "player_in", "player_out", "shot_place", "shot_outcome", "is_goal", "location", "bodypart", "assist_method", "situation", "fast_break"*]

Some of these variables are binary or numeric. See Appendix 1 for a legend.

3.1.2 Target Vectors

The target vectors consist of English language full text match summaries from Sky Sports[13]. This site was chosen because the style of summary appears to be generally consistent from year to year

and from match to match. Additionally, they have match summaries from the Top 5 leagues covered by the Football Events dataset.

3.2 Data Preprocessing

From the Football Events dataset, over 10,113 matches were found. A BS4 scraping tool was designed to pull summaries from SkySports league results pages. These summaries were matched with corresponding leagues and the corresponding seasons. Football Events information about team names, match scores and match dates were used to find the appropriate match summaries. 2,011 relevant match summaries were downloaded by the scraper. This gave us a total of 30,712 match summary sentences.

The next task was to connect individual game event objects with sentences in these summaries. There were 941,010 game events in total. I developed a regex tool that allowed me to assess the relevance of a game event to a specific summary sentence. This was based on information about players involved and the event described. Through this approach, 12,673 training examples were found. For reference a training example looked like this:

Source Vector:

[*7 8 na 1 lyon st etienne yoann gourcuff na na na na na 0 2 na 0 na 0*]

Target Vector:

[*a second half yoann gourcuff strike cancelled out kurt zoumas opener for les verts as lyon missed out on a third successive league win but stayed on course for champions league qualification*]

Each sentence was standardized with case and punctuation. Also, I had to create a dictionary of team names because SkySports and FootballEvents referred to the same team with different names. This allowed for easy cross-referencing beween source and target sentences.

I then split the set of all source-target pairs into a training set, a development set and a test set of 8871, 2535 and 1267 examples respectively.

4 Experiments

4.1 Evaluation Method

It is unfair to compare the results of this experiment to some of the aforementioned rules-based approaches to match summary generation. In those cases, coherence within a sentence is likely not to be an issue because each sentence is built from a set of prefabricated linguistically correct templates. With the NMT approach in this study, however, it is highly likely that the test output will feature incoherent, nonsensical sentences. Thus, the evaluation method for this generator needs to be different.

4.1.1 BLEU Score

BLEU score is computed [11]. BLEU is useful because it is objective/deterministic and allows for a quick analysis of many test examples. Additionally, BLEU is good because it is a quantified metric meaning that comparison between various translations is easy.

To calculate BLEU a few steps are required. First, one must calculated the $n-gram$ precision for each translation. This is given by:

$$p_n = \frac{\sum_{ngram} min(max(Count_{r_i}(ngram), Count_c(ngram)))}{\sum_{ngram} Count_c(ngram)}$$

Once we have this, we calculate brevity penalty (BP). BP is equal to 1 if $c \geq r^*$ where r^* is the length of the closest reference translation. Otherwise, BP is

$exp(\frac{(}{1} - r^*c)$

Then we calculate BLEU with BP and p_n:

BLEU = $BP \times exp(\sum_{n=1}^{4} \lambda_n log(p_n))$

4.1.2 Subjective Score

BLEU also has its drawbacks. Closeness to a reference translation, is not the only goal of a translation system. A relevant and coherent sentence can be a good translation even if it does not closely resemble the reference translation. It might be different just because the reference set is not comprehensive enough. Recognizing this, I devised a small test for how effective each of my outputs was. I asked 10 participants to grade a random sample of 20 generated sentences for coherence, specificity (what statistics or names are mentioned?) and saliency (with regards to the context - i.e. do they talk about a goal scoring chance). Each of these was on a scale of 0 to 10. Each person has a different scale but with the random samples and the different viewpoints, the scores might provide insight into the effectiveness of various tests.

4.2 Experimental Details

I performed 3 main tests. I trained each one on batches of 40. The dropout I used was 0.3. Validation was performed every 200 iterations for maximum of 30 epochs and 5 patience rounds.

4.2.1 Naive

The first test was a naive approach. I flattened all the event objects into a single vector. This large vector became the source sentence for the NMT system. The target sentence that matched this source sentence was the entire match summary that corresponded to the game in question.

4.2.2 Alpha

The second experiment consisted of a source sentence that only had the alphabetical components of the event object. For instance, the indicator variables were dropped from the object, leaving only player name, team name and a few others. The target sentence was a regex-matched sentence.

4.2.3 Alpha-Numeric

The third test was the approach outlined in **Data**. Each training example consisted of a condensed Football Events event object as the source sentence (alpha-numeric) with a regex-matched sentence as the corresponding target sentence.

4.3 Results

4.3.1 Evaluation Metrics

Tests			
Metric	Naive	Alpha	Alpha Numeric
BLEU	0.163	0.321	0.678
Coherence	3.4	3.8	7.8
Specificity	0.4	0.9	0.5
Saliency	3.3	3.2	4.9

Below are some outputs from the Alpha-Numeric test

4.3.2 Sample Output 1

Source:
[*4 1 na 1 elche getafe hinestroza przemyslaw na na na na 1 4 2 0 1 0*]

Reference:
[*at the other end on loan midfielder hinestroza cut in from the left but his strike was straight at elche custodian przemyslaw tyton*]

Output:
[*the spaniard picked up the ball from the edge of the half hour mark but his shot was blocked by lorient*]

4.3.3 Sample Output 2

Source:
[*25 1 12 1 elche getafe rodrigues domingo cisma na na 8 2 0 15 1 1 1 0*]

Reference:
[*guaita was kept busy at the start of the second half as he had to tip over domingo cismas header as elche picked up where they left off*]

Output:
[*but came <unk> through the first half with alexis sanchez hitting the bar after <unk> minutes with a lobbed effort from 18 yards*]

4.3.4 Sample Output 3

Source:
[*52 7 na 2 genoa juventus na na eugenio lamanna mattia perin na na 0 na na 0 na 0*]

Reference:
[*they took the lead in the 25th minute when tevez showed his fine goal past roman weidenfeller on wednesday was no fluke with a similarly stunning strike which hit the underside of the crossbar on its way past mattia perin*]

Output:
[*despite the other end of <unk> <unk> in the <unk> minute with a brilliant free kick while luis suarez scored twice in the second half to move onto <unk> pass*]

5 Analysis

Because we cannot compare to the rules-based summary generators, we have no real baselines to compare to. The 3 tests do provide interesting points of comparison, however. It is incredibly clear which training paradigm works the best. The Alpha-Numeric approach bested the other two approaches in 3 out of the 4 metrics. Even if the NMT system is not specifically designed to handle numeric inputs, the extra information does seem to improve the performance of the summary generation according to both objective and subjective measures. The only measure where it fails is specificity. All of the other metrics failed at this too, however. That is a failure on the model itself. Despite my best efforts, the model is not equipped to handle non-vocab entities. A rules based approach would probably work in the future.

In the other metrics, the approach scores relatively highly. The good coherence and saliency scores are reflected in the sample outputs. In sample output 1, we have the minor non-logical reference to the 'half-hour mark' as a physical location. Nevertheless, it details a shot that is saved by a goalie. This matches both the source vector and the reference. The specificity is non-existent as every name is butchered and replaced with something else.

Sample output 2 shows similar strengths and weaknesses. The model clearly understands the context here. The source and reference refer to a missed shot. The output, too, talks about a missed shot. The same mistakes as in output 1 are seen here. All the team names and player names are displaced or missing. We all see the presence of UNK tokens.

Sample output 3 indicates some larger structural issues. The output is moderately coherent, yet it does not capture the context of the situation and it does not mention any of the relevant specific entities. More concerning, however, is the link between reference and source in this case. These 2 sentences do not match up. If inconsistencies like this exist in the training set, then it is very hard to create an effective neural network-based summary generator. Additionally, the reference sentence details events that are outside the scope of the event data. This shows a huge problem in the approach. So much good summary writing takes into account the larger macro context of individual moments. This is something that is not captured by the event object and thus cannot be replicated by the NMT.

6 Conclusion

It is clear that sports journalists do not have to fear. Automation is still a while away. While the results are not perfect, it is a promising start. It seems like a novel way to use NMT frameworks for tasks removed from the traditional sphere of machine translation. If anything, it illustrates how versatile and robust these systems are.

There are nonetheless a litany of issues that need to be sorted out if any progress is to be made.

The model is currently not equipped to handle numerical information that well. The strength of the model comes from the use of word embeddings. With numbers, that ability is neglected. Either the data is changed to only take in strings or the model is equipped to process these numbers intelligently.

The next big issue is the specificity of the generated text. I need to come up with a way of injecting named entities into the generated text. I might have to replace team and player entities with placeholder tokens. This will probably be in the form of special substitute tokens that can be replaced with the appropriate strings after the decoder stage.

The multiplicative attention of this model may not be the best mechanism to use for this task. It will be interesting to see how different attention mechanisms affect the efficacy of the model.

The other question is, how do you introduce larger narratives into these summaries? There needs to be a way for the NMT system to understand the importance of each and every moment to the larger picture. That sounds difficult and it probably is. It is a key part of good sports-writing that probably won't be mechanically reproducible for a while.

Sample output 3 really underscores how unreliable the training set maybe. My approach for creating training examples was admittedly quite naive. There needs to be a better way of relating game events with game summary sentences. It may require manual effort to do effectively. A simple regex does not suffice.

6.1 Other Datasets?

Am improved version will probably rely on data from other sources. While the Football Events Data is comprehensive, it is not well suited to this task. Many event elements are incomplete - they have 'NA' in place of details. It is an inconsistent form that significantly decreases the effectiveness of training.

Additionally, if the idea is to create a system that instantly generates summaries of football matches, the Football Events set is probably not the most reliable. It appears to be manually derived and pretty time intensive to create, which defeats the purpose of the system. Instead, the ideal system would probably rely on data streams that are already in place. Using video input or voice commentary would be interesting, though incredibly difficult. What is more feasible, and what I will probably explore next, is using the live game feed - occasional text sentences that describe events - as source sentences for a match summary NMT.

7 Additional Information

Mentor:
Michael Hermann Hahn - mhahn2@stanford.edu2

References

[1] Elisabeth Andre, Gerd Herzog, and Thomas Rist. "On the Simultaneous Interpretation of Real World Image Sequences and their Natural Language Description: The System Soccer." In: *ECAI*. Vol. 88. 1988, pp. 449–54.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[3] Kim Binsted. "A talking head architecture for entertainment and experimentation". In: *AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition*. 1998.

[4] Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. "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*. Association for Computational Linguistics. 2011, pp. 72–81.

[5] Nadjet Bouayad-Agha et al. "Perspective-oriented generation of football match summaries: Old tasks, new challenges". In: *ACM Transactions on Speech and Language Processing (TSLP)* 9.2 (2012), p. 3.

[6] Kyunghyun Cho et al. "On the properties of neural machine translation: Encoder-decoder approaches". In: *arXiv preprint arXiv:1409.1259* (2014).

[7] Andreas Graefe. "Guide to automated journalism". In: (2016).

[8] Melvin Johnson et al. "Google's multilingual neural machine translation system: Enabling zero-shot translation". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 339–351.

[9] Nal Kalchbrenner and Phil Blunsom. "Recurrent continuous translation models". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 1700–1709.

[10] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025* (2015).

[11] Kishore Papineni et al. "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2002, pp. 311–318.

[12] Alin Secareanu. *Football Events*. `https://www.kaggle.com/secareanualin/football-events`. 2017.

[13] *SkySports Football Match Reports*.

[14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[15] Yonghui Wu et al. "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144* (2016).

Appendices

Football Event Legend

	event_type
0	Announcement
1	Attempt
2	Corner
3	Foul
4	Yellow card
5	Second yellow card
6	Red card
7	Substitution
8	Free kick won
9	Offside
10	Hand ball

11 Penalty conceded

event_type2
12 Key Pass
13 Failed through ball
14 Sending off
15 Own goal

side
1 Home
2 Away

shot_place
1 Bit too high
2 Blocked
3 Bottom left corner
4 Bottom right corner
5 Centre of the goal
6 High and wide
7 Hits the bar
8 Misses to the left
9 Misses to the right
10 Too high
11 Top centre of the goal
12 Top left corner
13 Top right corner

shot_outcome
1 On target
2 Off target
3 Blocked
4 Hit the bar

location
1 Attacking half
2 Defensive half
3 Centre of the box
4 Left wing
5 Right wing
6 Difficult angle and long range
7 Difficult angle on the left
8 Difficult angle on the right
9 Left side of the box
10 Left side of the six yard box
11 Right side of the box
12 Right side of the six yard box
13 Very close range
14 Penalty spot
15 Outside the box
16 Long range
17 More than 35 yards
18 More than 40 yards
19 Not recorded

```
bodypart
1       right foot
2       left foot
3       head


assist_method
0       None
1       Pass
2       Cross
3       Headed pass
4       Through ball


situation
1       Open play
2       Set piece
3       Corner
4       Free kick
```