Detecting Interesting Event Sequences for Sports Reporting

François Lareau **Mark Dras** **Robert Dale**

Centre for Language Technology
Macquarie University, Sydney, Australia
francois.lareau|mark.dras|robert.dale@mq.edu.au

Abstract

Hand-crafted approaches to content determination are expensive to port to new domains. Machine-learned approaches, on the other hand, tend to be limited to relatively simple selection of items from data sets. We observe that in time series domains, textual descriptions often aggregate a series of events into a compact description. We present a simple technique for automatically determining sequences of events that are worth reporting, and evaluate its effectiveness.

1 Introduction

We are developing a Natural Language Generation (NLG) system for generating commentary-style textual descriptions of Australian Football League (AFL) games, in both English and the Australian Aboriginal language Arrernte. There are a number of research questions to be tackled: one is how to handle a resource-poor, non-configurational language, the inherent complexities of which are outlined by Austin and Bresnan (1996); another, the focus of this paper, is the issue of content selection in the sports domain. More precisely, we are concerned with a kind of content aggregation that we call *aggregative inference*. Below is an extract from a typical human-authored commentary for a game:[1]

> Led by Brownlow medallist Adam Goodes and veteran Jude Bolton, the Swans kicked seven goals from 16 entries inside their forward 50 to open a 30-point advantage at the final change—to that point the largest lead of the match.

There is a corresponding database which contains quantitative and other data regarding the game: who scored when, from where, and so on. In the example given above, the phrase *the Swans kicked seven goals from 16 entries* goes beyond simply putting similar facts together; it involves an inference on the score progression to identify a strong moment of arbitrary duration in the game. In human-authored commentaries, we observed that this kind of aggregation is common; but existing content selection and aggregation techniques will not suffice here.

After surveying some related work on data-to-text generation and content selection (§2), we characterise our notion of aggregative inference, and present an analysis of our AFL data to demonstrate that it is a significant phenomenon (§3). We then propose a method for this task that can be used as a baseline for future work, and examine its adequacy for content selection (§4).

2 Related work

Time series Previous work has dealt with time series data and the particular problem of segmenting them meaningfully. Time series are typically continuous processes monitored at regular intervals; ours, in contrast, are irregular sequences of discrete events. The main difference is the number of data points: for example, a pressure sensor can produce thousands of readings in a day, but we only need to consider about 50 events in a game (see §3).

The SUMTIME project (Sripada et al., 2003b; Yu et al., 2004) aims to produce a generic time series summary generator. It has been applied to weather forecasts (Sripada et al., 2002; Sripada et al., 2003a), neo-natal intensive care (Sripada et al., 2003c; Portet et al., 2009), and gas turbine monitoring (Yu et al., 2006). For weather forecasts, Keogh et al. (2001) used a bottom-up segmentation technique that required thresholds to be set. In SUMTIME-TURBINE, a search was made for patterns that had to be identified in a semi-automatic way using expert knowledge.

[1] All texts and data in this paper are from www.afl.com.au and stats.rleague.com/afl. For an explanation of the game, see en.wikipedia.org/wiki/Australian_rules_football.

We want to do without thresholds and experts, using instead paired data and text (as in machine learning approaches, discussed below). In the domain of neonatal intensive care, Gao et al. (2009) focused on detecting unrecorded events in time-series; in contrast, we want to detect clusters of events rather than individual events. In the domain of air quality, Wanner et al. (2007) do not explain in detail how they segmented their curves, but they appear to have detected peaks and then considered the intervals between these peaks, assessing their slope. We need to be able to assess the slopes between any two data points, as human-authored texts refer to intervals other than those between peaks (cf. §3). Boyd (1998) used a signal processing technique called *wavelets* to detect trends in weather data. This is similar to a Fourier transform, except that it is not constrained to a specific time window, an important feature for detecting trends of arbitrary lengths. In her evaluation, 17 out of 26 trends (65.4%) mentioned by experts in human-authored texts were predicted by her system. Again, she did not have paired data and text.

Sports In general, content selection in the sports domain has so far amounted to selecting individual events in the game (Oh and Shrobe, 2008; Bouayad-Agha et al., 2011), with the exception of the work of Barzilay and Lapata (2005), discussed below. Some previous NLG systems for the sports domain were live speech generators (Herzog and Wazinski, 1994; André et al., 2000) that faced problems inherent to incremental NLG which are not relevant for us, in particular the fact that content selection must take place before the full course of the game is known. Robin (1994) focused mainly on *opportunistic generation*, i.e., the addition of background information, which is not the subject of our current work.

Machine learning Duboue and McKeown (2003) were the first to propose a machine learning approach to content selection; this and subsequent work has almost exclusively looked at selecting items from the raw tabular data. Taking aligned summaries and database entries in the domain of biographical texts, Duboue and McKeown (2003) construct a classification model for selecting both database rows that match the text exactly, and others that require some clustering across their graph-based representation. Barzilay and Lapata (2005) also take a classification approach, working on American football data. Formulating the problem as one of energy minimisation allows them to find a globally optimal set of database rows, in contrast to the independent row selection of Duboue and McKeown (2003). The goal of both approaches was to extract and present items that occur in the tabular data; Barzilay and Lapata (2005) explicitly restrict themselves to selecting from this raw data. Kelly et al. (2009), applying Barzilay and Lapata's approach to the domain of cricket, go beyond looking at raw data items to a limited 'grouping' of data, for example in pairing player data for batting partnerships.

		Event		Score		
Time	**Player**	**H**	**A**	**H**	**A**	**M**
$1'40''$	Jesse White	G		6	0	6
$4'42''$	Jarrad McVeigh	B		7	0	7
$10'05''$	Patrick Ryder		B	7	1	6

Table 1: Sample scoring events data

Player	**K**	**M**	**H**	**G**	**B**	**T**
Jude Bolton	16	3	20	0	0	12
Adam Goodes	11	5	5	2	4	1
Heath Grundy	8	2	8	0	0	1

Table 2: Sample of in-game player statistics

In contrast, we are interested in presenting not just raw data, but data over which some inference has been carried out (as in the selection of time series data by Yu et al. (2004)), and the feasibility of using a machine learning approach to achieve this.

3 Correlating data and texts

Our data comes in the form of tables that focus on different aspects of the game. The most important for our current purpose is the table of scoring events, which gives information about the score progression in the game: goals (worth 6 points) and behinds (1 point) scored by the home and away teams, their respective scores, and the margin[2] (see Table 1). There is also a table with statistics for each player during a given game, with his number of kicks, marks, handballs, goals, behinds and tackles for the match, as shown in Table 2. Other data is available that we do not have space to show here.

We collected human-authored summaries to see how they relate to the available data. The particular

[2]The home team's score minus the visitors'.

summaries we used are the published commentary of the sort found in newspapers: ours came from the Match Centre of the AFL website.[3] These are typically written by professional sports journalists as the game is taking place, and posted on the web shortly after the game has finished. The writers consequently have access to video of the game, and to the extensive set of statistics available from the Match Centre during the course of the game.

Each story is around 500 words long and consists of roughly 15–20 sentences organised in short paragraphs (a couple of sentences each). A typical text starts with a summary of the game's key facts: who won by how many points at which stadium, along with an overall characterisation of the match. It then continues with a more or less chronological presentation of the course of the game, an evaluation of each team's key players in the match, and a list of the injured; and it concludes with the consequences of the game's result on the season's rankings and a teaser about the upcoming games.

The stories essentially focus on in-game events (as opposed to background information), in particular scoring events. We also observed that more than half of the information conveyed required some sort of reasoning over the data. We identified three main types of propositions expressed in the text:

Raw data: propositions that refer to data readily available from the database, e.g., the margin in *The Swans led by 33 points at the final break*.

Homogeneous aggregative inferences: propositions that require reasoning over one type of data, e.g., *the Tigers kicked eight of the last 10 goals* (where there is no database entry that corresponds to this statistic, and it is necessary to carry out an aggregation over goals for an arbitrary time period) or *the result was never in doubt* (which is a more abstract assessment of the score over a period of time).

Heterogeneous aggregative inferences: propositions that require inferences on data of different types, e.g., *Melbourne physically dominated the Swans* (which refers to a combination of tackles, contested marks, players' physical attributes, and so on).

We distinguish *surface aggregation*, where information is packaged at the linguistic level, and *deep aggregation*, which takes place at the conceptual level; compare, e.g., *Johnson marked six goals and gathered 25 possessions* with *Johnson gave a stellar performance*. We are only concerned with the latter.

[3] See `www.afl.com.au`.

Type	#	%
Raw data	120	38.8
Score-based homo. aggreg. infer.	68	21.7
Other homogeneous aggreg. infer.	13	4.2
Heterogeneous aggregative infer.	112	35.8
Total	313	100.0

Table 3: Types of information conveyed in AFL stories

In a first step, we manually annotated ten of the collected texts using the above typology, leaving aside all propositions that did not refer to in-game information, and ignoring surface aggregations. Since scoring events are so important in this genre, we further divided the homogeneous aggregative inference type into two sub-categories—those based on score and those based on other data—and annotated the texts accordingly; Table 3 summarises the breakdown.[4]

Raw data accounts for just under 39% of the data expressed in these texts; the score at various points in time makes up the bulk of this category. In an AFL game, it is normal to see 30 goals and a similar number of behinds being scored. Consequently, not all are mentioned in the texts, so the problem with raw data in this context is to select the events that are mentionworthy; this problem has been explored already (cf. §2). More interesting, however, are score-based aggregative inferences, calculated from a sequence of goals and behinds. These account for almost 22% of our small corpus, and are not amenable to detection by existing approaches.

In a second step, we drew the curve for the score margin in every game, then took each expression marked as a score-based aggregative inference and identified the elements of the curve it referred to: (1) individual scoring events (points in time where the margin changes), (2) intervals between scoring events, or (3) the area under the curve (see Table 4). For the expressions that referred to intervals, we identified four subtypes: (1) those that refer to intervals where a team is on a roll (scoring points for a sustained period of time), or (2) when there is a

[4] We first annotated ten other stories with finer-grained categories, then two annotators went through three iterations of this mark-up until they agreed, before we annotated these ten stories.

Type	#	%
Intervals between events	40	58.8
Individual events	24	35.3
Area under the curve	4	5.9
Total	68	100.0

Table 4: Types of score inferences

Subtype	#	%
Team is on a roll	22	55.0
Tight struggle	7	17.5
Lead changes	5	12.5
Other	6	15.0
Total	40	100.0

Table 5: Subtypes of intervals referred to in texts

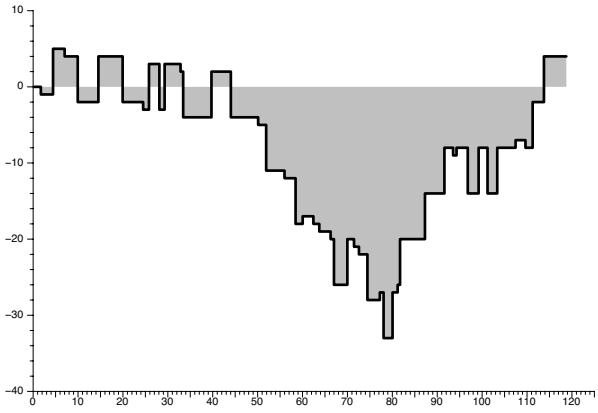

Figure 1: Sample score margin curve

tight struggle (a relatively extended period where no team is able to change the score margin significantly), (3) expressions that refer to the number of lead changes, and (4) other expressions (see Table 5).

It is clear from these observations that detecting when a team is on a roll is a very important kind of aggregative inference in this genre. We propose below a technique for doing this. Since detecting tight struggles is a closely related problem, we will also try to tackle it at the same time.

4 A curve segmentation technique

The goal is to identify clusters of events of arbitrary duration that form a unit of discourse. In contrast to the SUMTIME systems, where patterns in time series data are codified through discussions with experts or are subject to a user-defined threshold, we want to identify a measure such that content selection can be learned automatically, as an extension of techniques like those already used for homogeneous aggregative inferences (§2). We look for intervals in the score margin curve where the slope is either steep or rather flat (cf. Figure 1). What makes the problem non-trivial is that we do not know how steep or flat the curve needs to be in order to be interesting, how long the interval should be, and where it should start. There are 'natural time anchors' for intervals, namely the beginning and the end of the game or quarters, and peaks in the curve; however, human reporters also select intervals that are not bound to these anchors.

We calculate for each game the absolute value[5] of the slope between all pairs of scoring events (goals and behinds).[6] The slopes are then normalised relative to all other slopes that span the same number of events in the same game (by subtracting the mean and dividing by standard deviation); a steep slope over a short span (when a goal is scored right after another, say) is not as meaningful as an equally steep slope over a long span (which corresponds to a roll).

As an illustration, Figure 2 gives the matrix for the curve in Figure 1. Scoring events are numbered 1 to 49, and each cell corresponds to the interval between two events, with darkness indicating the normalised value. The shortest intervals appear along the diagonal edge, and as we move away from the edge and towards the upper-right corner of the matrix, we get longer intervals. The interval with the highest value in this matrix is the one between events 32 and 35 (at row 32, column 35). Indeed, it is the interval between the 78th and 82nd minutes of play, when the home team kicked back into the game. Notice that all the cells in row 32 and column 32 have a high value. This is because the 32nd event is the lowest point of the curve, so the slope between any point and this one is likely to be higher than normal. Hence, such dark bands identify important peaks in the curve. Notice also the contrast between the generally low values in the columns 1 to 17, and the generally higher ones in columns 18 and up. This contrast identifies another kind of inflection point in the curve: the event 17 is the one at the 50th minute of play, just before the curve plunges deep into negative values.

[5] The direction in which the margin changes is irrelevant.

[6] There are around 50 such events in a typical match, so there is a matrix of roughly 1200 pairs to consider (for n events in a game, there are $n \times \frac{n-1}{2}$ possible intervals).

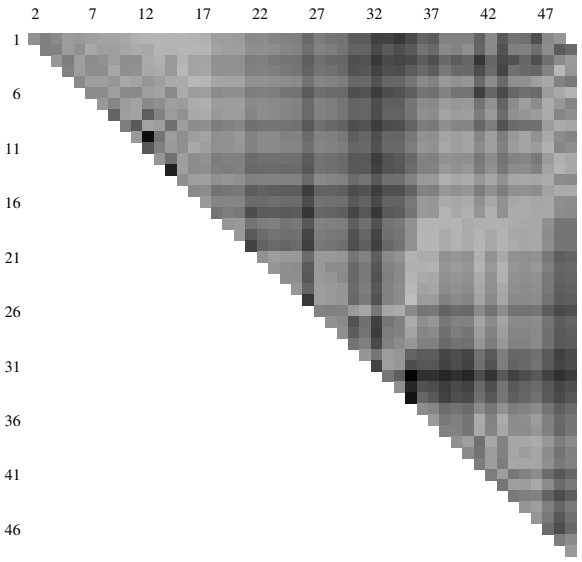

Figure 2: Sample matrix of normalised interval slopes

Rolls			Struggles		
Rank	**#**	**%**	**Rank**	**#**	**%**
≥ 0.9	15	68.2	≤ 0.1	3	42.9
≥ 0.8	17	77.3	≤ 0.2	3	42.9
≥ 0.7	20	90.9	≤ 0.3	4	57.1
Total	22	100.0	Total	7	100.0
Median: 0.956			Median: 0.204		

Table 6: Percentile ranks for normalised interval slopes

Finally, the normalised values are ranked in comparison with the other values for the game, and the ranks are expressed as percentiles. One would expect that when a team is on a roll, the slope for the corresponding interval will be comparatively high, and should rank towards the top, while in contrast, when the game is tight, the curve should look rather flat, and therefore the corresponding interval's normalised slope should have a low rank. The fact that the slopes are normalised relative to other slopes of equal intervals makes it possible to compare intervals of any duration and to rank them regardless of length.

We tested this technique on the data that corresponds to the texts we had annotated, and checked how many of the rolls and struggles mentioned in the texts received a rank that made sense (high ranks for rolls, low ones for struggles); see Table 6.

The technique works well for rolls, and could be used as a baseline and as a starting point for a stochastic reranking approach: taking the top 30%, say, and reranking based on other local score context.

For the rolls where the rank was lower than 0.9, most were cases where either it was not clear what interval was referred to in the text, or there was a reversal in the trend (and this was communicatively more important than the roll itself), or a roll was mentioned precisely because it was mild in contrast with another interval mentioned elsewhere.

The results are not as promising for struggles, probably because struggles tend to be in games with a generally flat curve, so that any segment of the game is likely not particularly more flat than the rest of the match, and therefore hard to detect. One possible alternative is to use a different score-related measure, e.g. a matrix of lead changes per time period. A second is to compare intervals with other intervals of the same duration in all games, rather than in the same game, as in the 'measures of interestingness based on unusualness' of Yu et al. (2004).

With respect to other work, our segmentation technique does not fit into any of the three types mentioned in Sripada et al. (2002): sliding window, top-down or bottom-up. It is not a pattern matching technique either, as in Yu et al. (2006). The normalisation of the segments aims to handle the variability of granularity that we need; this is the same goal as the wavelet technique of Boyd (1998), but our approach is technically much simpler. However, this method is only viable for curves with a limited number of data points, since it must take into account all possible sub-segments of the curve.

5 Conclusion

We have assessed the content of human summaries of football games in terms of the source of data for the facts they express, and have observed that a significant proportion of these facts were derived from inferences made on the score progression.

One frequent type of score inference is to detect exciting segments of the game, that is, either when a team is on a roll, or when there is a tight struggle. We have proposed a baseline technique to detect such intervals based on the slope between any two scoring events on a score margin curve. Our preliminary results show that this technique tends to do quite well at detecting when a team is on a roll, and somewhat less well at detecting tight struggles. We now plan to use it as a baseline for the evaluation of machine learning techniques.

Acknowledgements

The authors acknowledge the support of ARC Discovery DP1095443.

References

Elisabeth André, Kim Binsted, Kumiko Tanaka-Ishii, Sean Luke, Gerd Herzog, and Thomas Rist. 2000. Three RoboCup simulation league commentator systems. *AI Magazine*, 21(1):57–66.

Peter Austin and Joan Bresnan. 1996. Non-configurationality in Australian aboriginal languages. *Natural Language and Linguistic Theory*, 14(2):215–268.

Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff, editors, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP'05)*, pages 331–338, Vancouver.

Nadjet Bouayad-Agha, Gerard Casamayor, Leo Wanner, Fernando Díez, and Sergio López Hernández. 2011. FootbOWL: Using a generic ontology of football competition for planning match summaries. In *Proceedings of the Extended Semantic Web Conference (ESWC'11)*, Heraklion, Greece.

Sarah Boyd. 1998. TREND: a system for generating intelligent descriptions of time series data. In *Proceedings of the IEEE international conference on intelligent processing systems (ICIPS-1998)*.

Pablo Ariel Duboue and Kathleen R. McKeown. 2003. Statistical acquisition of content selection rules for natural language generation. In Michael Collins and Mark Steedman, editors, *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, pages 121–128, Sapporo.

Feng Gao, Yaji Sripada, Jim Hunter, and François Portet. 2009. Using temporal constraints to integrate signal analysis and domain knowledge in medical event detection. In *Proceedings of the 12th Conference on Artificial Intelligence in Medicine: Artificial Intelligence in Medicine (AIME'09)*, pages 46–55, Verona, Italy.

Gerd Herzog and Peter Wazinski. 1994. VIsual TRAnslator: Linking perceptions and natural language descriptions. *Artificial Intelligence Review*, 8(2–3):175–187.

Colin Kelly, Ann Copestake, and Nikiforos Karamanis. 2009. Investigating content selection for language generation using machine learning. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 130–137, Athens.

Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. 2001. An online algorithm for segmenting time series. In *Proceedings of the IEEE International Conference on Data Mining (ICDM 2001)*, pages 289–296, San Jose, CA.

Alice Oh and Howard Shrobe. 2008. Generating baseball summaries from multiple perspectives by reordering content. In *Proceedings of the Fifth International Natural Language Generation Conference (INLG 2008)*, pages 173–176, Salt Fork.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

Jacques Robin. 1994. *Revision-Based Generation of Natural Language Summaries Providing Historical Background*. Ph.D. thesis, Columbia University, New York.

Somayajulu G. Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2002. Segmenting time series for weather forecasting. In A. Macintosh, R. Ellis, and F. Coenen, editors, *Applications and Innovations in Intelligent Systems X*, pages 193–206. Springer.

Somayajulu G. Sripada, Ehud Reiter, and Ian Davy. 2003a. SumTime-Mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10.

Somayajulu G. Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2003b. Generating english summaries of time series data using the gricean maxims. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'03)*, pages 187–196, Washington.

Somayajulu G. Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2003c. Summarizing neonatal time series data. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics (EACL'03)*, volume 2, pages 167–170, Budapest.

Leo Wanner, Bernd Bohnet, Nadjet Bouayad-Agha, François Lareau, Achim Lohmeyer, and Daniel Nicklaß. 2007. On the challenge of creating and communicating air quality information. In A. Swayne and J. Hrebicek, editors, *Proceedings of ISESS 2007*, Prague.

Jin Yu, Ehud Reiter, Jim Hunter, and Somayajulu Sripada. 2004. A new architecture for summarising time series data. In *Proceedings of INLG-04 Poster Session*, pages 47–50, Brockenhurst, UK.

Jin Yu, Ehud Reiter, Jim Hunter, and Chris Mellish. 2006. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(1):25–49.