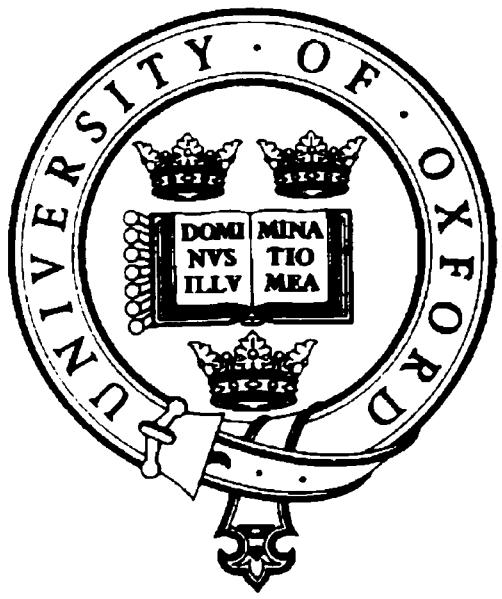
Personalised Information Filtering using event causality

Catherine Dolbear
Keble College



Supervisor: J. M. Brady

Robotics Research Group
Department of Engineering Science
University of Oxford

Michaelmas Term, 2004

This thesis is submitted to the Department of Engineering Science, University of Oxford, in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The thesis is entirely my own work and, except where otherwise stated, describes my own research.

Catherine Dolbear, Keble College, Oxford

Catherine Dolbear is the recipient of a Motorola University Partnerships in Research grant.

Abstract

Previous research on multimedia information filtering has mainly concentrated on key frame identification and video skim generation for browsing purposes, however applications requiring the generation of summaries as the final product for user consumption are of equal scientific and commercial interest. Recent advances in computer vision have enabled the extraction of semantic events from an audio-visual signal, so it can be assumed for our purposes that such semantic labels are already available for use. We concentrate instead on developing methods to prioritise these semantic elements for inclusion in a summary which can be personalised to meet a particular user's needs. Our work differentiates itself from that in the literature as it is driven by the results of a knowledge elicitation study with expert summarisers. The experts in our study believe that summaries structured as a narrative are better able to convey the content of the original data to a user.

Motivated by the information filtering problem, the primary contribution of this thesis is the design and implementation of a system to summarise sequences of events by automatic modelling of the causal relationships between them. We show, by comparison against summaries generated by experts and with the introduction of a new coherence metric, that modelling the causal relationships between events increases the coherence and accuracy of summaries. We suggest that this claim is valid, not only in the domain of soccer highlights generation, in which we carry out the bulk of our experiments, but also in any other domain in which causal relationships can be identified between events. This proposal is tested by applying our summarisation system to another, significantly different domain, that of business meeting summarisation, using the soccer training set and a manually generated ontology mapping.

We introduce the concept of a *context-group* of causally related events as a first step towards modelling narrative episodes and present a comparison between a case based reasoning and a two-stage Markov model approach to summarisation. For both methods we show that by including entire context-groups in the summary, rather than single events in isolation, more accurate summaries can be generated. Our approach to personalisation biases a summary according to particular narrative plotlines using different subsets of the training data. Results show that the number of instances of certain event classes can be increased by biasing the training set appropriately. This method gives very similar results to a standard weighting method, while avoiding the need to tailor the weights to a particular application domain.

Acknowledgements

We can do no great things, only little things, with great love.

Mother Theresa

For Paul, Sue and Elizabeth, who kept me sane through this thesis.

A heartfelt thank you to all my teachers: Mike Brady, my supervisor; Paola Hobson, Jonathan Teh, Genevieve Conaty, Gus Reid, Holly Kelleher, Raquel Navarro-Prieto, Alyson Evans and Kevin Brooks at Motorola; my examiners Barry Smyth and Janet Efstatiou for making my viva as painless as possible; Timor Kadir, Iead Rezek, Marco Rigolli, Byung-Woo Hong, Styliani Petroudi, Maud Poissonier, Vicky Mortimer, Melissa Terras, Veit Schenk, Mark Gooding, Sarah Bond and all the MVL students; Kathryn Blacker at the BBC and Gary Jacob from the Times, for making the time to help with my research; Jan, David, Tony, Eddie, Ben, Alison and Alistair for allowing me to hijack a Board of Trustees meeting; Stephanie Hunter, Alice Thomas, Jessica Meade, Matt McDonnell, David Gwynn, Henri Martius, Felix Hoffmeir, Anastasia Nijnik, Dil Joseph and the other members of Keble MCR; the girls of the Keble College Women's Association Football Club, for saving my joy of football; Emma Dedman, Zishan Hussain, David Fisher and Trio Watson, for listening; the staff, volunteers and guests at the Gatehouse, for Wednesday nights of White Lightening and Chardonnay; and finally, to my family: my parents John and Ruth, and Helen and Paul Bryce, for pretending to understand my thesis, and for sage football advice.

Contents

List of Figures

Chapter 1

Introduction

Intelligence is what you use when you don't know what to do.

Jean Piaget

Information overload is becoming an increasing problem for users of digital multimedia communications, as they are presented with an ever-increasing flow of data via email, the web, video and audio sources. Filtering and summarisation techniques, which provide information relevant to a user's current task, are becoming more and more important in combating this problem. This thesis provides a new approach to the information filtering problem using a novel synthesis of ideas drawn from the areas of narrative intelligence, summarisation and machine learning.

There are many issues to resolve in the implementation of a personalised information filtering system, and we cannot hope to address them all in this thesis. The rationale for the focus of our work within the larger information filtering problem is as follows. While there have been many advances in recent years in the areas of natural language processing of text for summarisation and low-level object and event recognition from audio and video signals, the work in this thesis operates at a higher semantic level than most previous attempts. We assume that semantic labels extracted from multimedia data are already available for our use, and instead concentrate on developing methods to prioritise these semantic elements for inclusion

in a summary that can be personalised to meet a particular user's needs. Our work differentiates itself from previous research as it is driven by the results of a knowledge elicitation study with expert summarisers. The experts in our study believe that summaries structured as a narrative are better able to convey the content of the original data to a user. Since causality is one of the most significant relationships between elements in a narrative, we use this as the basis for our narrative modelling. The central claim of this thesis is that by modelling the causal relationships between events, the coherence and accuracy of summaries can be increased. We suggest that this claim is valid, not only in the domain of soccer highlights generation, in which we carry out the bulk of our experiments, but also in any other domain in which causal relationships can be identified between events. This proposal is tested by applying our summarisation system to another, significantly different domain, that of business meeting summarisation, with considerable success.

This chapter sets the scene for the thesis, beginning in section 1.1 with the project objectives and system design criteria. The reasons for choosing the example application scenarios of soccer highlights generation and business meeting summarisation are given in section 1.2. Some terms used in this thesis are defined in section 1.3 and the technology is set in context in section 1.4. Finally in section 1.5, the structure of the thesis and its most important contributions are outlined.

1.1 Aims, objectives and design criteria

The aim of the thesis is to gather evidence to support our claim that information can be filtered more coherently and accurately when the causal relationships between events are represented in the summarised output.

This is broken down into specific project objectives as follows:

- To elicit knowledge about the summarisation process from experts in the field and apply this to our automatic system.

- To develop an automatic summarisation system in the example domain of soccer, which can be robustly tested against summaries generated by experts.
- To evaluate and compare the performance of summarisation methods using Case Based Reasoning and Markov models.
- To personalise the summary whilst retaining coherence.
- To test the generality of the Markov model method by applying the system to another, significantly different application domain: business meeting summaries.

Our summarisation system is designed with a number of criteria in mind:

- It should be possible to eventually couple the system to a real audio/video input.
- The system should allow for the mobilisation of a large amount of knowledge, the "background knowledge" of the domain, which may not come directly from the input data to be summarised, but is simply known by people versed in the domain. (For example, in the soccer domain, fans would know which players played for which clubs.)
- The system should demonstrate how narrative causal structures improve coherence and context.
- The system should be extensible and adaptable to multiple domains.

1.2 Application scenarios.

It is a characteristic of human intelligence that the wide range of knowledge employed by specialists in a particular domain (such as medicine, the stock market,

etc.) can be rapidly and effortlessly mobilised, as well as easily applied across domains through their natural ability to generalise. Artificial Intelligence has not yet developed techniques that can emulate this, rather it works best when the domain of discourse is narrowed, yet is still open enough to be useful and interesting. Therefore, in order to focus on a tractable, yet still demanding, area for study within the larger information filtering problem, we concentrate on entertainment applications, specifically the generation of football highlights from a full length match according to the preferences of a football fan. The football domain has been selected because of several notable characteristics. Football, like all sports, has comprehensive rules that govern the appearance of the teams and the players' behaviour. For example, teams, goalkeepers and referees wear distinctive colours, play nominally lasts ninety minutes in two halves and takes place on a pitch of known dimensions. This means that it is feasible to extract semantic information about the soccer events and the players involved from audio-visual data, as Ekin et al. (2003) have shown. Unlike some domains, for example medicine or the law, the background knowledge base for soccer, while large, can be limited in size so that it is practical to encode much of the immediately relevant information. Another advantage of the domain is that soccer highlights will only contain video clips in their original temporal order. Since they occur in succession, it is easier to identify causal relationships between events, than might be possible in, for example, text of a fictional narrative, where events can be presented in any order. Another major advantage of the soccer application is the ready supply of data and summarised solutions on which our system can be trained and tested.

The massive interest in soccer makes commercialisation of filtering techniques in this domain attractive. For example, the personalised soccer highlights domain has been addressed as part of the content management research of the project sponsors, Motorola, and we are able to leverage the results of their user requirements study

[Evans (2003)] in our work. Manually edited soccer highlights are already being marketed as a major application for third generation mobile phones, but since substantial expertise and time is required to edit soccer highlights by hand, an advantage of an automatic system would be to allow a user to receive personalised highlights. For instance, a fan of a particular team could express a preference for highlights featuring their team or favourite player. There are around 50 games across all divisions of the English football league each week, plus additional games played in European Championships and Internationals, as well as Scottish league games, and the football leagues of other countries, so highlights of hundreds of soccer matches could be produced every week. It takes a professional editor at the BBC between 1 and 1.5 hours to edit a highlights package; if the highlights are being produced in real-time for Interactive TV, two editors are needed per game. Taking into account the requirements of personalisation, where several different highlights packages must be produced for each football match, this adds up to around 900 person-hours of editing per week for a personalised soccer highlights service. Since most football matches take place at the weekend and users want to access highlights as soon as possible after the game, efficiency is a strong motivation for developing an automatic system to generate personalised highlights.

However, it is important to realise that the filtering techniques developed for the soccer domain can in principle be extended to other domains where summary coherence relies on accurate representation of temporal and causal relationships between events. To test this, we use the business meeting domain as a second application example, as it is far enough removed from football to present a challenge to our system's adaptability.

1.3 Definitions

Information filtering is the process of excluding data that is unwanted or irrelevant to the user, and presenting only information pertinent to the user's current task. Methods for information retrieval are closely related to those of information filtering as they both have the goal of retrieving information relevant to users' needs. However, information filtering is concerned with removing information, while information retrieval is concerned with finding information. While it is more accurate to refer to the process carried out by our system as *filtering*, we also use the term *summarisation*, as it is widely used in this context by the computer vision community.

Throughout this thesis we will be using several terms that, for clarification, will be defined here. We use both *soccer* and *football* to refer to the game of Association Football. We discuss the concept of *metadata* in this project: this is simply data about data. The term can be used to refer to any data used to aid the identification, description and location of video, audio, textual, graphical or other electronic resources. For example, a video film may be described by its title, director, principal actors, etc, which are all pieces of metadata, usually referred to as *high level metadata*, as distinct from *low level metadata* which may be derived from signal features such as motion vectors, colour histograms etc.

The other frequently used term in this thesis is *ontology*. Gruber (1993) defines an ontology as *"an explicit specification of a conceptualization"*. Ontologies specify the terms in the domain and the relations between them, and as such we design ontologies for both the soccer and the business meeting domain to clarify the semantics for the summarisation process.

We follow the conventions of the information retrieval literature in measuring our

results. That is, we measure summarisation precision and recall:

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{1.1}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{1.2}$$

A *true positive* is an event that is included in our summary that was also in the expert's summary, while a *false positive* is an event in our summary that the expert did *not* choose to include. Where it is necessary to compare two experiments which result in different precision and recall rates, since there is often a trade-off between the two, we also make use of the F measure, [van Rijsbergen (1979)] using $\beta = 1$, which allows us to combine precision and recall:

$$F_\beta = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot (precision + recall)} \tag{1.3}$$

1.4 Technology context

There has been renewed interest in content management issues from both the multimedia and artificial intelligence communities, with the advent of the MPEG-7 "Multimedia Content Description Interface" standard [Chang et al. (2001)] and the development of semantic markup languages like RDF [Lassila and Swick (1999)] and OWL [McGuinness and van Harmelen (2004)] as part of Berners-Lee's vision of the semantic web [Berners-Lee et al. (2001)]. Although some use has been made of MPEG-7 content descriptors for news and sport summarisation [Divakaran et al. (2003)], the methods are based only on low-level descriptors such as motion activity. They do not provide any deep semantic understanding of the data included in the summary, and hence remain domain specific and preclude any personalisation of the summary.

More complex semantic descriptions are provided by the ever growing number of RDF Schema and OWL ontologies published on the web. Tools [e.g. Noy and Musen (2000)] are also becoming more readily available to map between ontologies for reuse

and knowledge sharing, although as yet mapping tends to be between very similar domains.

Research on multimedia information filtering by the computer vision community has mainly concentrated on key frame identification and video skim generation for browsing purposes [Sundaram and Chang (2001), Erol et al. (2003)]. The information extraction problem has been addressed in the soccer domain by, among others, Ekin et al. (2003), Assfalg et al. (2003) and Sadlier et al. (2004), using audio and video features such as colour density analysis, slow motion replay detection, penalty-box detection and speech-band energy to identify semantic events using machine learning techniques such as Bayesian Belief Networks or Support Vector machines. With such systems, *any* event that can be recognised is deemed important enough to include in the summary. This leaves the generation of more meaningful summaries containing only *relevant* events as a promising area of research.

It has long been recognised in the natural language processing community [Lehnert (1981)] that an accurate summary includes all the narrative elements of the original text, while Trabasso and Sperry (1985) have shown that the importance of a text unit depends directly on the number and quality of causal relations that the unit has to other text units. More recently, narrative cohesion and coherence have been the basis of sentence selection algorithms for text summarisation [Mani et al. (1998)]. The causal relationships in all of these methods have been manually annotated, so an open area of research is how to identify these causal relationships automatically, and develop new ways of exploiting them in the summarisation process.

1.5 Thesis outline

The prior work mentioned briefly in the previous section is expanded on in **Chapter 2**. Our literature review covers the areas of computer vision and natural language processing techniques for audio-visual and text summarisation; narrative intelligence

research on modelling causal and episodic structure in data; and ontology mapping between domains.

Chapter 3 discusses our knowledge elicitation study, with findings taken from interviews with three BBC Sport producers and a sports journalist from the Times; a protocol analysis of football highlights editing; and data review of match reports. a television edit decision log and personalisation in fanzines. The main result of this study is that a soccer highlights editor is primarily interested in telling the story of the game, and hence this supports Lehnert's ideas about the importance of narrative in summarisation.

We use this finding to drive the design of our summarisation system, described in **Chapters 4 and 5**, where we introduce the concept of a "context group": a sequence of events that form a causal chain. In **Chapter 4** we present a soccer event ontology and method for extracting instances of the ontology from "ticker-tapes" of football games published on the web. This is used to build a case base of full-length soccer games and their corresponding summaries. Chapter 4 then describes a case based reasoning approach to soccer summarisation, which has a mean precision over 126 tests of 42% and recall of 23%, when adaptation is performed one event at a time. When a whole context group of events is adapted at a time, the results improve to 46% precision and 52% recall. This is the main contribution of Chapter 4, showing the advantages of the context group method in providing more accurate summaries when benchmarked against those produced by experts.

Chapter 5 describes the development of a novel probabilistic alternative to case based reasoning summarisation using a two-stage Markov model. Firstly, events are clustered into context groups using either a Markov chain or Hidden Markov Model and secondly, each context group is assigned a priority of inclusion using a Markov chain to calculate the joint probability of the whole group. The Markov chain transition probabilities are estimated using frequency of event co-occurrence from the

training set of football games. Events are described either by their class only, or by a 16 dimensional feature vector containing background knowledge about the football player that carried out the event and the club they play for. The results presented in Chapter 5 show that the Markov chain mechanism sets a limit on the complexity of the semantics, so we are unable to exploit this amount of background knowledge as it leads to a very high dimensionality search space. However, using the event class information only, the Markov chain method has average results of 59% precision and 65% recall, a 13% improvement on the CBR method for both measures. Also, the average F_1 results over the test set are 7.3% higher than those reported in the literature [Conroy and O'Leary (2001)], as well as our method offering the additional flexibility of generating a summary of any length. Since only the event class feature is used to describe each event in the Markov chain, another advantage of our approach is that it is easily applicable to event descriptions in other domains. As with Case Based Reasoning, the context-group based results from the Markov chain summaries are better than the single-event based summaries.

Chapter 6 addresses issues of summary personalisation, introducing an alternative to the traditional weighting method, by biasing a summary according to particular plotlines using different subsets of the training data. Results show that the number of instances of certain event classes can be increased by biasing the training set appropriately. This method gives very similar results to a standard weighting method for personalisation, while avoiding the need to tailor the weights to a particular application domain. Using profiles of two example users, we show that our summarisation system can produce a summary to within 8% of the required duration at a 5.6% compression rate (that is, much heavier compression than has usually been applied in prior work.) We also show that while single-event based summarisation produces summaries of more accurate duration than context-group based summarisation, this advantage decreases significantly with summary length. The mean percentage error

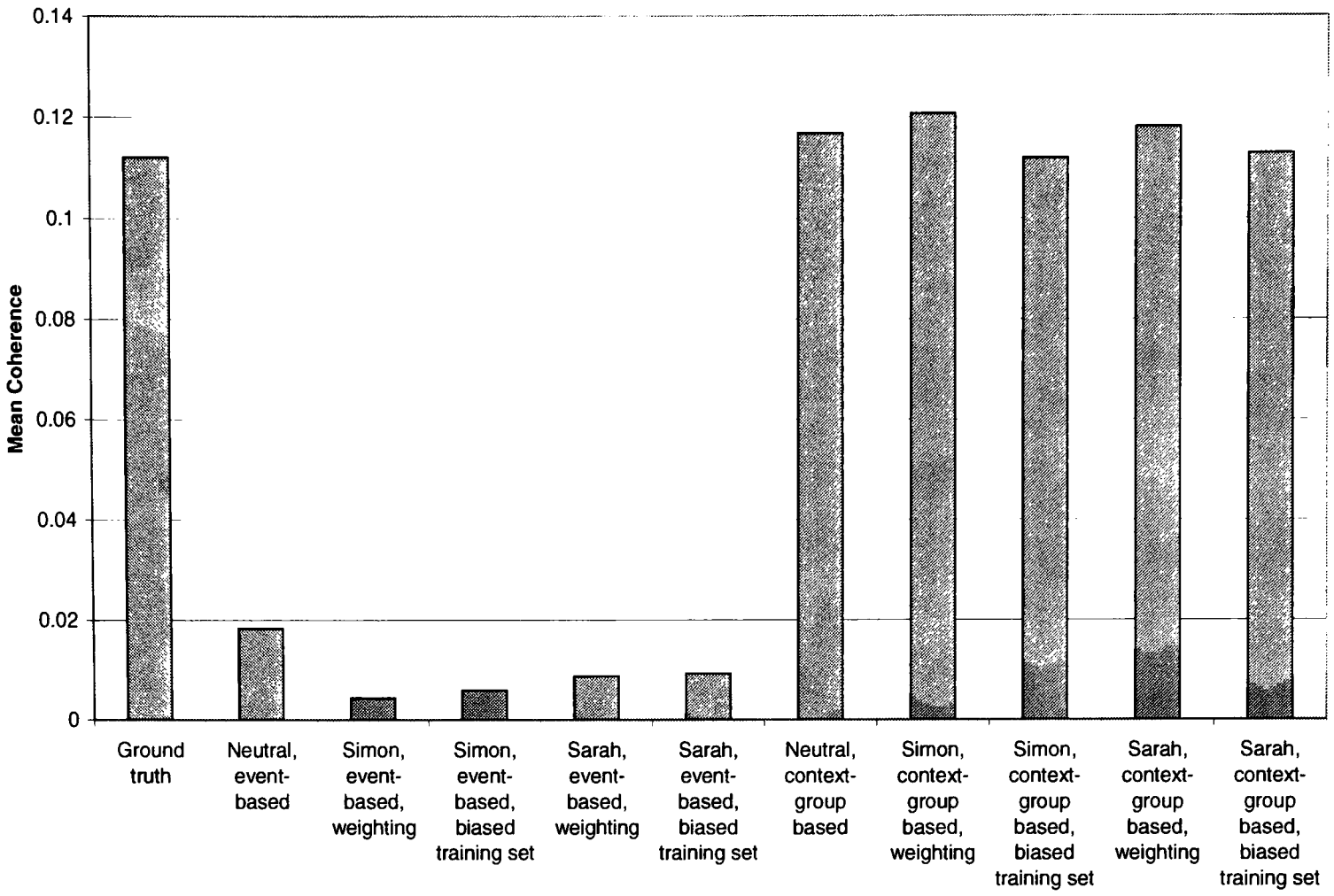


Figure 1.1: Mean coherence of various summaries; comparing ground-truth with neutral and personalised summaries for two example users *Simon* and *Sarah*.

between the actual and preferred summary length also decreases as the summary length increases.

Chapter 6 also introduces a measure of how well user requirements have been fulfilled, which we term "utility". Results show that utility increases with summary length, and, more interestingly, that some users are easier to please than others. In order to examine the trade-off between personalisation and coherence, a novel coherence metric based on the causal relationships between events in the summary is presented. As seen in figure 1.1, summaries which include one event at a time have lower coherence than those which include whole context groups, and the context group based method also produces more coherent summaries than the highlights generated by experts for broadcast on television (which we use as a"ground truth"). Furthermore, the reduction in coherence due to personalisation is smaller for the

context group based method. These experiments quantitatively demonstrate our view that the use of context groups improves coherence in summarisation.

In **Chapter 7** we present the results of a preliminary experiment to apply the soccer summarisation system to the business meeting domain. Using a manually generated ontology mapping, we summarise a meeting using the soccer training set and our Markov chain method. Although only one test is carried out, results show that the soccer summarisation system can successfully be mapped to the meeting domain, and again, the context-group based method is better than the single-event one. Finally, in **Chapter 8** we present a summary of the thesis and discuss some directions for future work leading on from our research.

1.6 Publications and patents

The following papers have been published or are pending in connection with this thesis work:

- *Soccer Highlights Generation using A Priori Semantic Knowledge* C. Dolbear and J. M. Brady. IEE International Conference on Visual Information Engineering VIE 2003 7-9th July 2003 University of Surrey, UK [relating to the work in Chapter 3]

- *Coherence and Personalization of Soccer Highlights.* C. Dolbear, J.M. Brady, J. Teh and P. Hobson 6th International Workshop on Image Analysis for Multimedia Interactive Services April 13-15th 2005 Montreaux, Switzerland [relating to work in Chapter 6]

The following patents have been filed on the work in this thesis:

- *A method for event-driven summarization using case based reasoning.* C. Dolbear and J. M. Brady. November 2003 [relating to the work in Chapter 4]

- *Apparatus and Method for Generating a Content Summary* C. Dolbear, J. M. Brady and J. Teh, February 2005 [relating to the work in Chapter 5]

- *Apparatus and Method for Generating a Personalized Content Summary* C. Dolbear, J. M. Brady and P. Hobson, February 2005 [relating to the work in Chapter 6]

Chapter 2

Literature Review

The problem tackled in this thesis lies at the intersection between several fields, namely computer vision, natural language processing, narrative intelligence and ontology research. The purpose of this chapter is therefore to outline relevant work in each of these areas, in order to motivate the particular approach we take in this thesis, to justify design decisions and to explain why we have narrowed the focus of our work along certain lines. We begin with a definition of some summarisation terms in section 2.1 and then consider different summary evaluation mechanisms in section 2.2. Both the computer vision and natural language processing communities have developed technologies that can be applied to the generation of soccer highlights and in sections 2.3 and 2.4 we examine their different approaches. Section 2.5 explains the concept of narrative intelligence and we discuss the contribution that it can make to summarisation. Section 2.6 looks at how the representation of the information in an ontology can affect the summarisation process and enable summarisation strategies developed for one domain to be mapped to another. We conclude this chapter with a discussion of the directions we take in this thesis in order to address some of the gaps in the field.

2.1 Definitions of summarisation terms

Firstly, we must make it clear that we are limiting our problem to *extraction*, rather than *abstraction* of the data. As defined in Hovy et al. (1999), "An extract is a selection of some of the material of the original, while an abstract is a condensation and reformulation of the original." For an abstract, therefore, material not in the original data set can be included in the summary. Although video synthesis has been addressed in the literature, for example inserting virtual objects into a video sequence [Smith et al. (1999)], soccer highlights packages broadcast on television generally consist of video clips which are a subset of the original video stream. Since we want to use these highlights in our learning methods, in summarisation terms we are focusing on extraction, not abstraction for our application.

Secondly, the soccer highlights application calls for an *informative*, rather than *indicative* summary. For example, video skims or browsing tools, such as Smith and Kanade (1995), provide indicative summaries, as they only give an indication of what is contained in the original. The user then has the option of accessing the full data set if they are interested in a particular piece of information. An informative summary however assumes that the user does not have access to, or time for viewing the original and so tries to give all the important information in a condensed form, as the end product for user consumption.

The distinction between *neutral* and *biased* summaries is raised in Chapter 6; the former tries to be objective, presenting action from both football teams, while the latter may concentrate on one particular narrative, for example. We also differentiate between *generic* and *user-focused* (or *query-based*) summaries. "A generic summary provides the author's point of view, while a query-based summary focuses on material of interest to the user" [Hovy et al. (1999)]. In our soccer application, the soccer highlights edited by the BBC and broadcast on TV for the benefit of many viewers and fans supporting many different teams can be described as a generic summary, as

opposed to highlights personalised for a particular football fan, which is an example of a user-focused summary.

2.2 Approaches to evaluation

There are two types of summary evaluation methods: *intrinsic* and *extrinsic*. Intrinsic methods measure a system's quality against some ideal or "ground-truth" summary, usually created by a human, while extrinsic methods evaluate the performance of a summarisation system for a given user task, such as enabling a user to categorise documents into different topics or to retrieve information [Mani and Bloedorn (1997)]. The problem with the intrinsic method is that there is no such thing as a single correct summary. Jing et al. (1998) have shown that human subjects can disagree by as much as 10% in creating summaries by hand[1]. Furthermore, Jing et al found that disagreement increases with summary length. Precision and recall (as defined in equations 1.1 and 1.2) are widely used in intrinsic evaluation of summaries. The F measure (equation 1.3) is also used to combine precision and recall and hence compare results at different points on the Receiver Operating Characteristic curve (the graph of precision against recall). Results reported for precision and recall of text summarisation methods are generally quite low, in absolute terms. That is, 100% accuracy is not yet a realistic goal for summarisation technology. For example Mani et al. (1999) report F_1 results between 47% and 72% for methods at 20% compression[2].

Jing et al argue that the binary nature of precision and recall measures makes them inappropriate to use. For example, in text summarisation, if one sentence is very similar in meaning to a second sentence, yet only one was selected by the majority of humans for inclusion in the summary, if a system chose the other sentence, it ought not

[1]Percentage agreement for a particular person is defined as the number of times they agreed with the majority opinion, divided by the number of people in the study.

[2]A percentage compression is calculated as the ratio of number of sentences in the summary, to the number of sentences in the original. That is, 20% compression means that only 20% of the original *remains*.

to be classified as completely wrong. Hatzivassiloglou and McKeown (1993) extend the definition of precision and recall to allow for the degree with which two items agree. We use this insight to develop our "correct classes" measure, which considers events in the same ontology class to have a degree of agreement. In this measure, an event is considered to be a true positive if it is from the same class as the event that was selected by the expert for the summary. For example, if a certain shot on goal was included in the soccer highlights shown on television, but our system selected a different shot on goal for its summary, this would contribute to the "correct classes" precision and recall measures, as the viewer would probably have found the alternative shot interesting too.

Crampes et al. (1998) describe a number of criteria which influence summary quality:

- Conciseness. The summary must contain the smallest number of the most concise excerpts, for example to fit within a certain time limit.
- Pertinence. The excerpts must be relevant to the viewer.
- Completeness. The summary must give enough information to fulfil the user's wishes.
- Coherence. The excerpts and the implicit or explicit links between them must be coherent. For example, in the soccer domain, if one excerpt shows a player shooting a goal, and the next clip is of the same player saving a goal, the viewer is bound to be confused.
- Reliability. The final meaning of the summary must not induce interpretations that are not present, or that have the opposite meaning to that present in the original narrative. For example, in the soccer domain, if one clip shows a player shooting at the goal, and the next clip shows a save, we would assume that it

was this shot that was saved, so the summary must not contain a shot-save pair in sequence if they aren't causally related.

Due to limited resources, we will be evaluating our system intrinsically, using football matches and their highlights annotated from television broadcasts as our "ground truth". As well as precision and recall measures, we also introduce the "correct classes" measure described earlier in this section to overcome the binary nature of precision and recall metrics. While we are unable to measure pertinence or completeness without time-consuming subjective testing, we look at the issue of conciseness in Chapter 6. Our initial summarisation methods described in Chapters 4 and 5 aim for a summary of the same length as the ground truth we have recorded, but in Chapter 6 we investigate personalisation of summaries according to user requirements, including desired summary length. Coherence and reliability are modelled using the concepts of "context groups" of events, discussed further in sections 4.2.1 and 5.2 and we evaluate our results against the coherent and reliable ground truth summaries edited by professional sports producers.

2.3 Vision based summarisation

Current low level vision techniques already allow us to extract many types of metadata from video signals. Reid and Zisserman (1996) have shown that if we know the position of two or more cameras, we can produce the trajectory of the ball from the viewpoint of a virtual camera directly overhead of the field of play, even when the projective transform and camera calibration are not known, by using the geometry of the football pitch and the goalpost positions. We can track the ball and individual players by assigning a Kalman or Condensation trajectory to each object with an associated measure of error for the position and velocity, as implemented by Needham and Boyle (2001) for five-a-side football. Zelnik-Manor and Irani (2001) propose an algorithm using histograms of features at multiple temporal scales to identify

different events such as Kick, Throw, Run etc. The Físchlár project [Smeaton et al. (2003)] links together similar news stories to provide a personalised digest of the original broadcast, using text dialogue, name identification, shot boundary detection, speaker segmentation and matching, advertisement detection and speech versus music discrimination, and includes work [Sadlier et al. (2004)] on identifying salient portions of a Gaelic football match using the audio pitch and volume. Ekin et al. (2003) employ colour density analysis, shot boundary and slow motion replay detection, shot classification (e.g. close-up or long shot) and referee and penalty-box detection. A rule-based approach is then used to recognise soccer events such as goals, referee decisions, penalties, shots on goal and saves. Assfalg et al. (2003) have a similar technique, identifying eight different soccer events, such as the kick off, counterattacks and shots on goal, using finite state machines. Although both Ekin et al and Assfalg et al describe their systems as soccer summarisers, they are only addressing one of the two summarisation problems: information extraction. They do not go on to address the second problem: generation and presentation of the best possible summary to convey the salient points in the optimum time.

These developments in computer vision, which offer answers to the recognition problem, are encouraging, but are complementary to the essentially cognitive task that we address. That is, how does a sports editor decide *which* soccer events are important enough to include in a time-limited highlights package? The results of our knowledge elicitation study with television sports editors presented in Chapter 3 show that the filtering process mobilises substantial knowledge and requires more than simply including in the summary any event that can be semantically recognised.

Flow of play and event causality, for example the events leading up to a goal, are considered by sports editors to be very important elements of a good highlights package (see section 3.4). Our approach therefore differs from previous approaches to video summarisation since it considers event causality rather than presenting isolated

events or entire topic segments to the user.

Video summaries are often presented using key frames or video skims for browsing. For example, Sundaram and Chang (2001) group shots according to consistency in chromacity, lighting and ambient sound. The authors only retain the first, middle and last groups in the scene in the skim, with no regard to content semantics, under the assumption that the content of skipped shots can be inferred. This might mean that the run up to the goal was shown, immediately followed by the crowd cheering afterwards: while the goal itself could be inferred by most people, they wouldn't be happy it hadn't been shown! This example demonstrates why video skim technology is insufficient for our application, and why semantic understanding of the content is vital.

2.4 Text summarisation

The other main body of work on summarisation comes from the natural language processing community. Most current text summarisation systems determine the importance of a sentence within a document by using various heuristics such as sentence position, cue phrases, word or phrase frequency, lexical cohesion and discourse structure. A learning algorithm, such as a Bayesian classifier, is then used to weight the scores of these various heuristics.

As described in McKeown et al. (1999), text summarisation requires the solution of two problems simultaneously: interpretation of the natural language text to identify the facts, and generation to produce a paragraph that conveys the important facts concisely. We follow the lead of McKeown et al, who use semantically labelled data rather than full text (or motion video) for their system to generate summaries of basketball games, in order to concentrate solely on the second of these two issues. The input to our soccer summarisation system is described more fully in section 4.2, but in brief, to generate semantically labelled data, we have chosen a simple template-

mining approach to information extraction from text. This allows us to concentrate on the problem of generating summaries, rather than complex information extraction from raw video or text.

McKeown et al's summariser uses a set of rules to construct a draft of the essential input facts, then additional information is inserted in between the facts, or using richer descriptions to replace more general words (for example "loping" or "stumbling" could replace "walking".) McKeown et al divide information into two classes: information that *must* appear in the summary, and information that *could* appear if there were space. Rather than this binary classification, we assign a priority to each event, so that we can include more events according to resource (time) constraints. As well as the most important facts, which McKeown et al found appearing in the first sentence of every match report (the game result, teams involved, location, date and the most remarkable statistic of a winning team player), they also suggest using historical information. We would also like to incorporate such historical information, which we refer to as *background knowledge*, but since we are not assuming a textual output, we need to use the background knowledge to influence the selection of events in the highlights video, rather than including the archive material itself. Although McKeown et al propose the use of background knowledge in their basketball summarisation system, they have not implemented it, as they were unable to deal with the combinatorial explosion of facts. They suggest searching systematically for maxima, minima and consecutive sequences of similar results, or applying domain-specific rules. For example, if a basketball player scores more than 20 points, or the most on his team, this fact should be included in the summary. Also of interest is McKeown et al's testing of their basketball system on a text corpus in the stock-market domain. Their domain mapping procedure consists of creating syntactic transformational rules between the two domains: the words in the basketball domain that should map to words in the stock-market domain. For example, a 'defeat' in basketball maps to a

'fall' in the stock-market. The drawback of this approach is that the mappings must all be carried out by hand, and are not guaranteed to be applicable in every test case. However, they report some success: the coarsest level of the basketball summarisation rules were all applicable in the stock-market domain and at the finest grain, 51.9% of them were. Although no evaluation of the resulting stock-market summaries are reported, it is encouraging that summarisation similarities have been found in two domains far more disparate than those used in most ontology mapping experiments (see section 2.6.1).

Event-based summarisation has been addressed by Maybury (1995), where key information is selected from an event database via rule-based reasoning using event frequencies, frequencies of relations between events and domain-specific importance measures. In this way, the causal relationship between events that frequently occur together can be determined. However, Maybury's approach requires manual specification of the reasoning axioms, which we would like to avoid. Vanderwende et al. (2004) also employ an event-centric approach to summarisation using a graph scoring algorithm to identify highly weighted nodes and relations in the graph. Each node is a word from a sentence of text and a relation represents the dependency between two words. A node with more links is considered more important, using the PageRank[3] system. "Events" are equated with verbs, so important events, with a PageRank score above a certain threshold, along with the highest ranking noun they are related to, are included in the summary. Using a recall-based automatic summary evaluation method, Vanderwende et al's average results are 34%, compared to human authored summaries, an improvement on the 32% achieved when entity (noun) centric summary generation is used. Capus and Tourigny (2003) suggest the use of case based reasoning to avoid specifying summarisation rules explicitly, but their case base is lim-

[3]Commonly used by search engines to prioritise web pages, the PageRank algorithm weights pages by the number of links from other pages. The more web pages that link to a particular page, the more important it is considered to be. Also, the greater the importance a page has, the more important its links to other pages are considered to be.

ited to a few hand-crafted examples. We provide a wider evaluation of the case based reasoning methodology applied to summarisation in Chapter 4, as well as taking the issue of event causality into account.

We are interested in natural language processing research on text cohesion and coherence, as it offers a way of representing causality at the syntactic and semantic levels. "Text cohesion models text in terms of relations between words or referring expressions", such as repetition, synonymy or adjacency [Mani et al. (1998)], to determine how tightly connected the text is, that is, whether the text is all about one topic. Coherence, on the other hand, models text "in terms of macro-level relations between clauses or sentences to help determine the overall argumentative structure of the text." For example, it models whether one sentence is a logical progression from the previous one, or is caused by it. Furthermore, Mani et al found that modelling text coherence gave better summarisation results than modelling cohesion. We use this result to support our proposal that modelling causality is likely to be a better summarisation strategy than simply identifying topics. Another example of a system modelling coherence is Hearst (1994)'s text tiling, which uses domain-independent word frequency to recognise the interactions of multiple simultaneous themes, so that text is partitioned into non-overlapping blocks. Hearst's method uses synonyms and word repetition to indicate coherence, however, it does not consider whether the *order* in which words are represented contributes to causal relationships between semantic concepts.

Trabasso and Sperry (1985) have shown that the importance of a text unit depends directly on the number and quality of relations that the unit has to other text units. Events that are linked by successive causes and consequences through a story are identified as being in a "causal chain": they are more frequently summarised and given higher ratings of importance than events that are not in the causal chain. Events with a higher number of causal connections are also shown to be more important. Trabasso

and Sperry demonstrate these principles by manually assessing logical criteria to assign causality to events in a textual narrative. This is too time consuming for us, and prone to errors of judgement. We would prefer to be able to infer causal relationships directly from our training data. In Chapter 5 we look at how we can learn the causal relations between events from the order in which they frequently occur.

2.5 Narrative Intelligence

Narrative Intelligence is a term coined by Davis and Travers (1998) to describe the intersection between artificial intelligence, literary theory and media technology. Narrative is used as a way of understanding the world, by harnessing the human ability to organise experience into story form. The primary result of our knowledge elicitation study described in Chapter 3 is that the highlights package must tell the story of the soccer match. Video clips are only filtered out when they do not contribute to the semantic continuum of the story. However, shots are included when they illustrate points in the story visually. This result suggests that narrative can be used as a system design principle to allow viewers to exercise their innate skill in interpreting information as a story.

Studies in cognitive psychology have shown that stories are more memorable (in terms of ease of recall) and quicker to process than disjointed statements [Mandler (1984)]. We believe that a summary structured as a narrative will therefore be more user-friendly than previous summarisation methods. In this section we study some of the approaches the narrative intelligence community has taken: firstly to story structure, and secondly to narrative filtering and summarisation.

2.5.1 Narrative structure

From childhood, we have all listened to stories being told, and while instinctively we may grasp that certain elements are needed for the telling of a good story, defining precisely what is required in the structure of a narrative is more difficult. Propp (1968) analysed the structure of Russian folktales and identified a limited number of elements that reoccur in all tales. For example, elements can be characters such as a hero, villain or family member; or functions such as "family member absents himself from home", "villain's trickery" or "return of the hero". Propp's morphology can, to a certain extent, be applied to football stories. For example, depending on user preferences, a player can be cast in the hero or villain's role, and the "return of the hero" after a spell away for injury or misconduct is a common theme in soccer. However, Propp proposes a global framework, where all events have to be interpreted as one of the story elements: this has the disadvantage of requiring a top-down rule-based approach.

Brooks (1996) defines stories as "a system of associations between elements, composed of events, people and things" while Meech (1999) proposes that narrative can be viewed as a framework for situating knowledge in a particular context. That is, it provides a framework for an audience to understand the background to current events which constrains their expectations of what will happen and what the events that do occur will mean. Mittal and Paris (1993) identify the following as components of context:

The problem-solving situation or tasks. For example, to present the viewer with information they require, as outlined in their user profile, or suggested by an expert sports editor.

The participants involved, their expertise, beliefs, aims etc. For example, the players, who have different roles in a team in order to score more goals than their opponents.

The mode of interaction in which communication is taking place. For example, soccer highlights being viewed on a mobile device with a small screen.

The discourse taking place. That is, the actual events shown as the highlights.

The external world. For example, the importance of a match's outcome depends not only on its final score, but on the results of the other matches being played which contribute to a team's league position.

Bal (1978) describes narratives in terms of layers, as shown in figure 2.1. The raw facts and chronological collection of events in any particular tale is called the *Fabula*. In our soccer highlights application, this would be the events of the soccer match to be summarised, along with the background knowledge relating to that match. For any given Fabula, the facts can be presented from different perspectives, and in some applications, though not for soccer highlights, in different sequential order, to produce a *Story*. This can then be rendered into a number of different forms, using different media, such as a film or a novel, which are then termed the *Narratives*. In the soccer highlights application, one Narrative might consist of key frames with captions, while another might be the motion video highlights.

Cinematography can be divided into categorical and narrative forms: categorical films base each segment of the film on one category, for example gardening, news or travel programmes. That is, any video clip within a categorical film has the same semantic meaning, irrespective of which video clip or event is placed before or after it. Narrative films on the other hand, create new meanings by the sequential association of initially distinct video sequences. This is known as the Kuleshov effect, following the discovery by the Russian cinematographer Kuleshov that a shot of an actor's expressionless face, intercut with shots of a bowl of soup, a woman in a coffin and a child playing with a toy, would lead the audience to interpret the actor as being correspondingly hungry, sad and happy in turn [Kuleshov (1974)].

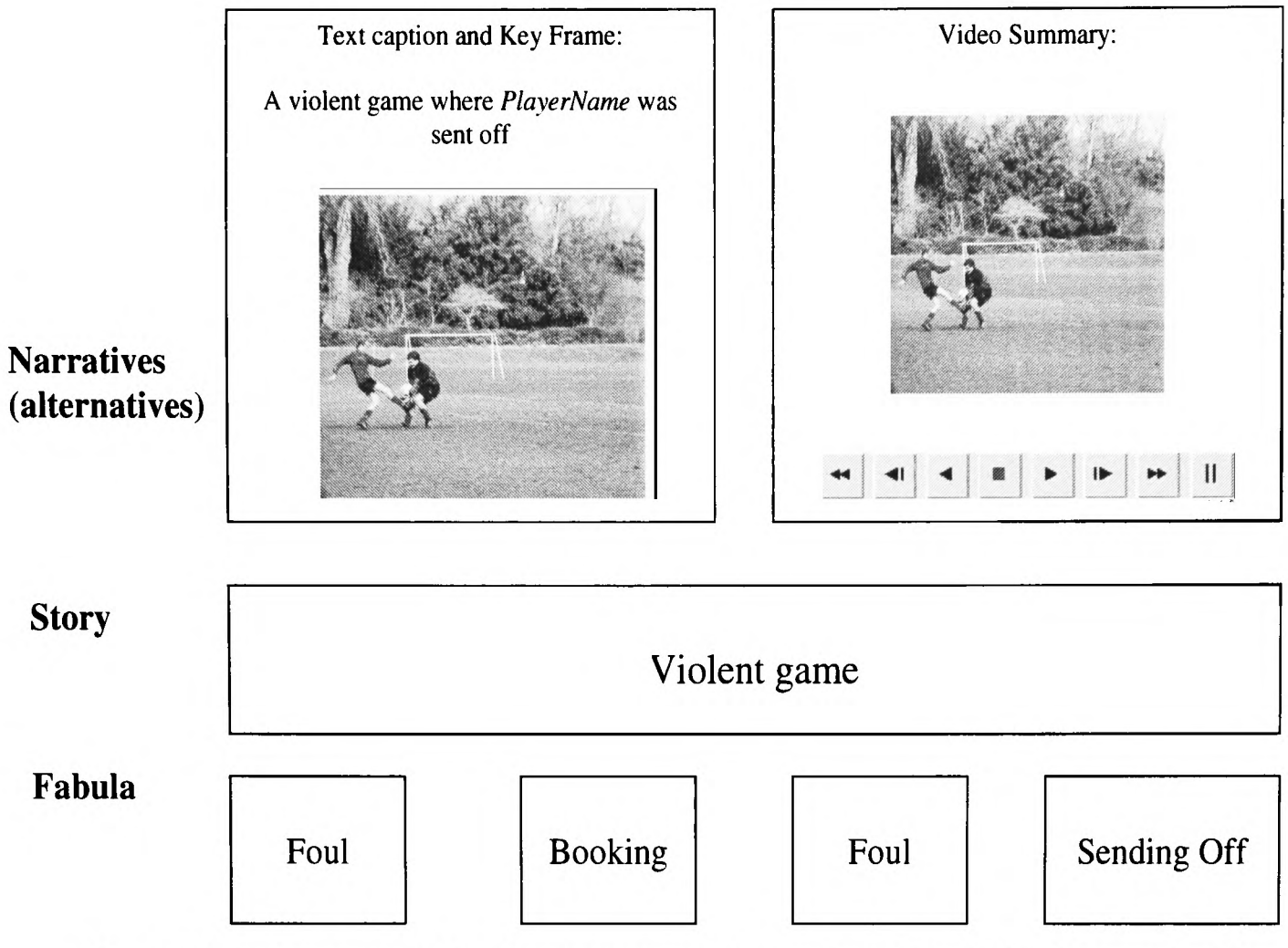


Figure 2.1: An example of a soccer game rendered in Bal's Narrative Layers

Several different methods have been proposed for linking narrative elements together. One of the earliest was Schank and Abelson (1977) who suggested that narrative could be understood in terms of scripts. For example, the script for a restaurant visit would be first to enter the restaurant, wait to be shown to your seat, order drinks, then food, eat the food, pay and then leave. From these different elements, various stories could be constructed. However, this paradigm is brittle: it cannot cope with any deviation from the script, because it assumes stories are made up of only one set of rules codified in the script.

Rather than inflexible, linear scripts, later work on narrative construction and sequencing has used software agents to allow a non-linear narrative to be generated. Brooks (1996) models the storytelling task by splitting it into what he terms the Structural, Representational and Presentation Environments. The structural environment contains a story framework or structure of abstract story element descrip-

tions known as narrative primitives: Speaker Introduction, Character Introduction, Conflict, Resolution, Diversion and Ending. Conflicts can have multiple Resolutions and Resolutions multiple Conflicts. Diversion is a story element which deviates or digresses from the plot, for example a comic incident[4]. The representational environment captures the relationships between story events such as *follows*, *precedes*, *must include*, *supports*, *opposes* and *conflict*↔*resolution*. Finally, the presentation environment makes the sequencing choices using autonomous agents to represent different story styles.

Another approach, by Lindley and Nack (2000) combines the categorical and the narrative approaches. A narrative planner uses a number of strategies to combine video clips into a sequence, based on actions, themes, goals and events to create a pattern of cause-effect relationships. For example, one strategy states: "If the action portrays an intention (goal), interrupt the action in a way that is unexpected by the character, so that the goal is unfulfilled and the character's mood is downgraded or he suffers in some way." This type of narrative generator is limited by the number of rules the system contains, so Lindley and Nack use their categorical generator to then use associative matching of the current clip to one in the database to shift context. For example, the next clip is chosen based on a weighted similarity measure to the clip metadata. The problem with this is that the sequence loses coherence as the storyline is broken. Overall, the narrative planner still needs to contain a significant number of rules for each theme or domain.

All the systems described thus far have required the narrative to be available in the database. If we consider the soccer highlights application however, we are limited by the actual soccer match as to what story we are able to tell. For example, we cannot tell a "David and Goliath" story if the two teams are equally matched. Furthermore,

[4]What is "comic" or "tragic" may depend on user preferences, for example an incident when the Manchester United goalkeeper failed to retrieve a normally harmless shot, which rolled between his legs into the goal, was seen as funny by the club's opponents, but was tragic to her own fans.

since we have plenty of previously summarised highlights available, we would like to be able to learn what kind of stories are usually told and how events relate to each other, rather than impose a fixed number of rule-based schema on to our output.

2.5.2 Narrative for filtering and summarisation

Lehnert (1981) was the first to use narrative to generate text summaries. Rather than a fixed sequence of events within a script, she proposed the concept of plot units, graphs of linked "affect states", which can be positive events (+), negative events (-) or mental states (M), which have neutral affect. Plot units avoid previous top-down approaches to story grammars, by starting with small units and building upwards. Each of the affect states occurs with respect to a single character, and events involving multiple characters require multiple affect states. Pairs of affect states are connected with pairwise causal links, which are one of four types: motivation (m), actualisation (a), termination (t) or equivalence (e). The three affect states and four causal links result in 15 possible combinations, under the following constraints: motivation links must point to a mental state, actualisation links must point from a mental state to an event, and termination or equivalence links must point from a mental state to a mental state, or from an event to an event. These 15 combinations are termed primitive plot units, and represent themes such as Success (M,a,+) or Hidden Blessing (-,e,+). More complicated plots can be built up using several of these primitive plot units.

Lehnert makes the experimental observation that the most accurate summary is the one that includes all the plot units from the original story. She also defines the concept of narrative cohesion, which requires connectivity across plot units, that is, tracing causal linkages from the first plot unit to the last[5]. The difficulty in applying

[5]Confusingly, 'cohesion' in text summarisation is a measure of how many different topics are described in the text, while Lehnert's definition of 'narrative cohesion' is actually more related to Mani's definition of 'coherence', in that there is a logical progression, or causal path, from the beginning to end of the summary.

Lehnert's method to our problem of automatic soccer highlights generation is that she has manually parsed the data on to specific plot units and causal links. Not only does this mean that her domain is very limited, but also that irrelevant data is not mapped, so the filtering process is actually occurring in the mapping phase. Later work [Singh and Barry (2003)] on commonsense knowledge acquisition employs templates to identify the plot units in stories, and use the general public's input from a web site to map hundreds of text stories on to plot units. However, we would prefer to avoid this type of explicit mapping at all.

Similarly, Crampes et al. (1998)'s approach to narrative summarisation also requires a manual mapping from the data to their narrative representation. However, since they are dealing with audio-visual data, they can only select extracts, rather than generate an abstraction as Lehnert did with her textual data. This means that not all causal links can be maintained, and Crampes et al note that extraction therefore not only emphasises those events which are selected, but also shifts the meaning of the events taken as a group. Sometimes a summary may be misleading: if two events are presented in succession, it is often presumed that they have a cause-effect relationship. For example, if a summary contains the sentences "Mary drove fast. Mary had an accident", we would tend to interpret this as "Mary had an accident because she drove fast" even if this causality is not expressed in the initial account where the two events are not linked. This may be regarded as the semantic equivalent of the Kuleshov effect.

Crampes et al identify three types of causality. Firstly, when every event is the result of all previous events; secondly, a causal event can be determined by some hidden, but universally recognised rule, for example, the rules of soccer mean that a Free kick event is caused by a Foul or an Offside. Thirdly, causality can simply be the result of a viewer's interpretation. For example, a disputed penalty could have very different causes depending on whether you asked a fan of one team or a fan of the

other. Crampes et al model changes in semantic meaning due to some video clips being excluded from the summary using manually constructed causal graphs. However, we would prefer to design a system that avoids the need for human intervention, but can harness the causality implicit in the data.

Kim et al. (2002) have developed a story schema layer on top of an ontology to create dynamic stories about artists' lives. Information is extracted from the web, using an artist ontology, and stored in a knowledge base. This is then queried using narrative construction tools to retrieve relevant facts or paragraphs of text to generate a specific biography, which can be personalised to the interests of a particular reader. However, the narrative construction tools consist of human-authored biography templates containing queries into the knowledge base, which again has the limiting requirement of manual intervention.

The drawback of all these attempts is that the narrative structures have been designed by hand, and the data has been hand-mapped (Lehnert, Crampes et al) or mapped using rule-based queries (Kim et al). A question we would like to address in this thesis is whether these mappings and narrative structures can be learnt from a corpus of data and its corresponding summaries.

2.6 Ontologies and domain independence

Gruber (1993) defines an ontology as *"an explicit specification of a conceptualization"*. Ontologies define the terms in the domain and the relations between them, and as such we have developed an ontology for our application domain in order to define what an event is and to decide what properties (metadata) should be attached to each event. As outlined by Noy and McGuinness (2001), an ontology makes domain assumptions explicit, so that they can be easily changed if knowledge about the domain changes. Furthermore, it separates the domain knowledge from the operational knowledge, so that the knowledge of the filtering process is separate from the knowledge about the

sporting domain. In Chapter 7, this allows us to apply the same filtering algorithm in a different domain. An ontology also allows us to analyse the domain knowledge, which is much easier once a declarative specification of the terms is available. Formal analysis of terms is very valuable if we want to reuse existing ontologies, extend, merge or maintain them, or map from one domain to another [McGuinness et al. (2000)]. For our application, structuring our knowledge base using a formal ontology also means that we impose a notion of similarity on instances in the knowledge base. For example, two Midfield Players are more similar than a Midfielder and a Goalkeeper. Concepts that are not in the ontology, such as information about the personal life of a player in the soccer example, rule out certain stories from being told in the summary, such as when a player is absent due to a court appearance or playing poorly because of worries off the pitch. This limitation, known as the "frame problem", is discussed further in section 4.2.

An ontology consists of *classes* which describe the concepts in the domain [Noy and McGuinness (2001)], *slots* (also known as *roles* or *properties*), which describe various features and attributes and *facets* (or *role restrictions*) which are limitations placed on slot values. An ontology, together with a set of individual *instances* of classes, constitutes a knowledge base. Developing an ontology consists of:

- Defining the classes in the ontology. These are the major concepts in the domain, such as Player, Team, Goal or Shot.

- Arranging the classes in a hierarchy of superclasses and subclasses, which represents an *is-a* relation. The class A *is a* subclass of B if every instance of A is also an instance of B. A subclass of a class represents a concept that is a *kind-of* the concept that the superclass represents. Therefore, although it might seem attractive to form Player, Teams and Matches into a hierarchy, they do not form one, as a Team is not a *kind of* Player, even though a Team may be made up of many Players. However, a Goalkeeper is a *kind of* Player, so could be one

of its subclasses.

- Defining slots, for example, *plays for, name, duration* etc. and restrictions on each slot's values. For example, the ontology can define a list of allowed values or specify the value type (String, Number, Boolean or Instance type). The Instance type allows the definition of relationships between individuals e.g. a slot *plays for* for the class Player may have instances of the class Team as its values, to denote which Team(s) the Player plays for. An instance of Player, *David Beckham*, could have two Team instances in its *plays for* slot: *Real Madrid* and *England*.

2.6.1 Ontology mapping

The need for ontology mapping tools to enable knowledge sharing, re-use and semantic information processing is increasing as more and more ontologies are published on the web. Ontology mapping is the process of establishing semantic correspondences between the classes and properties of two or more ontologies. A mapping may be one-to-one, or a more complex relation. For example, *name* can be mapped to the concatenation of *first-name* and *last-name*. Carried out manually, this is a very labour-intensive and error-prone task. Efforts to automate the process have focussed on ontology structural comparisons and linguistic similarity. For example, the ontology merging and alignment tool, PROMPT [Noy and Musen (2000)] and the information-flow based mapping approach [Kalfoglou and Schorlemmer (2002)], which offers a formal logical framework for these structural mappings.

Doan et al. (2002)'s GLUE system finds concept mappings by using similarity measures between each class in two ontologies. They implement a number of similarity

measures, including the Jacard measure:

$$Sim(A,B) = \frac{P(A \cap B)}{P(A \cup B)} = \frac{P(A,B)}{P(A,B) + P(A,\overline{B}) + P(\overline{A},B)} \quad (2.1)$$

where A and B are the set of instances of class A, in ontology *O1*, and class B, in ontology *O2*, respectively. *P(A,B)* can be defined as:

$$P(A,B) = \frac{\text{(number of instances in KB1 that belong to both A and B} \\ \text{+number of instances in KB2 that belong to both A and B)}}{size(KB1) + size(KB2)} \quad (2.2)$$

where instances of *O1* are contained in Knowledge Base *KB1* and instances of *O2* are contained in Knowledge Base *KB2*.

The Jacard measure is 1 when A and B are exactly the same, and 0 when they are disjoint. To determine whether s, an instance of class A, is also an instance of class B, a classifier is trained on all instances of class B in *KB2* as positive training examples, along with all instances in *KB2* that are not instances of B as negative training examples.

Most ontology mapping techniques in the literature have been tested on very similar ontologies [Kalfoglou and Schorlemmer (2003)], for example PROMPTDIFF [Noy and Musen (2002)], which works on two versions of the same ontology; GLUE [Doan et al. (2002)] which has been evaluated between two university course catalogues; and ONION [Mitra and Wiederhold (2002)], which maps between two commercial airline websites. This is because they are aiming to reuse the ontology vocabulary in a different application, rather than reuse the *reasoning* carried out on the ontology. In contrast, our aim is to reuse the summarisation process we develop, by mapping to a markedly different ontology. The research question is then: to what extent can a system developed to summarise information in one domain (e.g. soccer) be reused to summarise information in another domain (e.g. business meetings). We report on experiments to address this question in Chapter 7.

2.6.2 Ontology specification languages and the semantic web

The semantic web [Berners-Lee et al. (2001)] is a vision of the future for the World Wide Web, in which the semantic meaning of data is made explicit, making it easier for agents to automatically process and integrate information available on the Web. The World Wide Web Consortium has produced a stack of recommendations to extend the present HTML format to describe content structure and provide sets of inference rules: XML [Bray et al. (1998)], RDF [Lassila and Swick (1999)] and OWL [McGuinness and van Harmelen (2004)]. Encoding our data in one of these semantic web markup languages facilitates the reuse of the soccer ontology and knowledge base and, coupled with an ontology mapping, enables our summarisation process to be applied to other domains.

XML, the eXtensible Markup Language, is a syntax for structuring documents via customised tagging schemes. Since HTML is intended for the (unalterable) presentation of information as Web pages, it only contains a fixed set of markup tags, inappropriate for conveying the meaning of the data inside them. Another limitation is that the HTML tag set is not extensible. In contrast, XML has no fixed set of markup tags, rather they can be user-defined according to the data being represented. XML by itself is just hierarchically structured text, which must fit a certain ontology defined by an XML Schema. The XML Schema enforces constraints on the structure of the XML document, and new type definitions can be derived from old ones in an object-oriented fashion. Each element in the XML document must be defined under a particular *namespace*. The namespace, identified by a URI (Uniform Resource Identifier) reference, is the scope within which the element (and thus its name) is valid. We need namespaces for modularity: if elements were defined within a global scope, it becomes a problem when combining elements from multiple documents, as name collision is hard to avoid. For example, "match" has a different meaning in the soccer domain and the object recognition domain.

RDF (Resource Description Framework) is a data model for objects, or "resources" and the relations between them, using XML as the interchange syntax, while overcoming many of the problems faced by plain XML. For example, XML allows tree, graph and character string data structures to be mixed in computer memory, causing manipulation difficulties. While the order of elements in an XML document is significant and often very meaningful, RDF also avoids this problem. An RDF resource is represented by a subject, predicate and object triplet known as a *tuple*. The subject names the resource being described, e.g. *Player*. The predicate is a property of the resource, e.g. *player name* and the object contains the value of the property, e.g. *Michael Owen*. An RDF Schema (RDFS) expresses the ontology of the RDF document, defining the terms that will be used in the RDF tuples and giving specific meanings to them.

More recently, the Web Ontology Language, OWL, has been developed [McGuinness and van Harmelen (2004)], which extends RDF and RDFS by providing additional vocabulary for describing properties and classes, along with a formal semantics. For example, the relations *disjointWith*, *intersectionOf*, *oneOf* between classes can be described, along with cardinality, equality and properties of properties (uniqueness, transitivity etc.) Another advantage of OWL is in its support of ontology mapping or merging, which encourages sharing and re-use of ontologies. For example, if you wish to develop a general sports ontology, you might borrow previously developed ontologies from the football, hockey and cricket domains, thus reducing your workload. However, the three ontologies would have some concepts in common, such as Ball, Referee/Umpire, or Player, which could be denoted by the *owl:equivalentClass* or *owl:equivalentProperty* properties, whereas dissimilar concepts, such as "box" (the box surrounding the goal on a football pitch versus a cricket box), could be denoted by *owl:differentFrom*.

While OWL has been developed by web and ontology researchers, we will also

briefly mention the MPEG-7 standard [Chang et al. (2001)] that has emerged from the multimedia processing field. The MPEG-7 "Multimedia Content Description Interface" standardises specific ontologies (known as *Description Schemes*) which encode the structure and semantics of the relationships between *Descriptors* describing features of audio-visual content. For example, the standard includes low-level shape, texture, colour, motion and position descriptors for images or video, and energy, harmonicity and timbre for audio. Higher level semantic information can also be described, for example descriptors related to creation, production, management and access of content, such as coding scheme, overall data size, intellectual property rights or links to other relevant material. Currently, most MPEG-7 Description Schemes are written in XML syntax, although MPEG-7 ontologies have also been built in RDF Schema [Hunter (2001)]. This thesis will not use MPEG-7 directly, but it is feasible that event-level semantics derived from MPEG-7 audio-visual feature descriptors, encoded in RDF or OWL, could be an input to our system.

2.7 Literature review summary

Starting with the high level problem of "information filtering", we have chosen to focus in on generating an informative, extractive summary, with both generic and user-focused, neutral and biased variants. From the soccer highlights broadcast on television, we are able to use ground-truth summaries in our evaluations but, recognising the limitations on defining any summary as absolutely correct, we have a "correct event class" measure to avoid binary decisions of right or wrong.

In section 2.3 we used the literature to illustrate two points. Firstly, that computer vision techniques can deliver semantic recognition of objects and events to a sufficient degree for us to assume that we can begin our work using the data that has already been given semantic meaning. This allows us to concentrate on summary generation rather than data interpretation. Secondly, we make the point that, because we need

to produce an informative, rather than indicative summary, information extraction techniques alone cannot deliver an acceptable summary. The text summarisation literature has shown that modelling coherence improves the quality of summaries, so we go beyond simply presenting what can be recognised, to the linking of items together as a coherent whole. Hand-crafted rules or templates have often been used in summarisation: we would like to derive reasoning from the data, rather than have to impose explicit rules on the domain. The soccer highlights domain also has the advantage of having many examples of previously summarised highlights available, which we can use for training and testing. This gives us a unique perspective on the summarisation problem.

In Chapter 4 we model the causal relationships between events using Case Based Reasoning rather than a rule based approach and in Chapter 5 we demonstrate how the causal relationships can be learnt from frequency analysis of the event data. We have seen that the meaning of events can be changed due to juxtaposition in a summary, because such placement implies causality. as with the Kuleshov effect. Therefore we need to make sure that only the causal links present in the original material are represented in the summary. We propose that this can be achieved by learning from the causal relationships of the events in soccer highlights edited by experts. This assumes that an expert only edits two non-temporally adjacent events together into sequence when they either *are* causally related, or if not, there is no chance for a viewer to mistakenly infer causality.

While our knowledge elicitation study in Chapter 3 suggests a narrative structure is suitable for soccer highlights generation, prior work in the literature has shown that narratives are also applicable to summaries of other domains, indeed they may offer a mechanism for generality across domains. Since narratives deal with causal links between events, they are also suited for our application, where context must often be implicitly supplied.

Our system assigns a priority of inclusion for the event information, so that we can include more if resources permit, for example, for a user with a preference for viewing a longer summary. In Chapter 5 we also demonstrate how historical information (background knowledge) can be used to modulate data choice.

McKeown et al's work suggests that it is possible to map between domains to apply a summarisation procedure developed for one domain to a substantially different one, and we would like to test this proposal further. In the next chapters we will describe the development of an ontology for the soccer domain, both for the current events and the background (historical knowledge), which is then mapped to the domain of business meeting summarisation.

First though, Chapter 3 outlines our knowledge elicitation study, and describes the knowledge experts bring to bear on the problem of editing soccer highlights, along with the conclusions we have drawn about the influence of narrative over experts' summary selections.

Chapter 3

Knowledge Elicitation

Football is a very simple game. For 90 minutes 22 men go running after the ball and at the end the Germans win.

Gary Lineker

In order for us to implement an automatic information filtering system, it is first important to understand how information is manually selected by an expert and what criteria they use to judge a "good" summary. This chapter describes the process of knowledge elicitation from domain experts, examining the types of knowledge and reasoning they employ to solve an information filtering problem. Since our initial application focus is automatic soccer highlights generation, the experts we questioned were television sports editors and sports journalists. In later chapters we will use the outcomes of the knowledge elicitation study along with previous work in the literature to motivate the approach we have taken to solving the problem of information filtering, specifically in the soccer domain, but also. more broadly, across domains using ontology mapping techniques.

The chapter is structured as follows: first, well known approaches to knowledge elicitation are described in section 3.1, while section 3.2 looks at previous work in the literature on knowledge elicitation for football highlights. In section 3.3 we outline our knowledge elicitation study: the interviews, protocol analysis and data reviews

we carried out, the results of which are detailed in section 3.4. Conclusions are drawn in section 3.5, along with recommendations for our system design, and a summary of the chapter is presented in section 3.6.

3.1 Approaches to knowledge elicitation

As defined in Kidd (1987), knowledge acquisition involves eliciting, analysing and interpreting human expert knowledge and transferring this knowledge into a suitable machine representation. It is well known that knowledge elicitation, the primary task within the knowledge acquisition process, which involves the gathering of knowledge from an expert in the field, is beset by many problems. Experts are frequently not consciously aware of their own decision-making processes and are hence unable to articulate the reasoning behind what they are doing or the knowledge they are employing, which may be grounded in personal experience rather than based on explicit "textbook" knowledge. As the expert becomes more competent in their activity, the more automatic their use of knowledge becomes, and the less accessible it is to the knowledge engineer [Bainbridge (1986)]. Human knowledge representation is procedural and people find it difficult to describe exactly how they carry out these procedures or tasks. However, just because it can be difficult to verbalise does not mean the knowledge is not well understood by the expert. There are many approaches to overcoming these problems, of which several have been employed in our study.

Knowledge elicitation is essentially knowledge collection through different methods, such as conducting open or focused interviews, reading manuals and other documents, observing current users and experts at work, analysing past cases and many other activities. Each approach obtains different types of data for the knowledge engineer.

Diaper (1989) differentiates between various types of interview, such as open, focused, and structured interviews. Open interviews are unstructured and enable the

knowledge engineer to gain a broad overview of what the expert does and the problems they encounter. The focused interview is a specialised technique to elicit in a depth-first manner all that the expert knows about a particular topic. It concentrates on one aspect of the problem solving process and follows it throughout the session. On the other hand, knowledge elicitation in a structured interview is designed to fit a pre-planned format, such as a questionnaire, and covers a broad range of topics within the expert's field. If interviews are carried out with many people, this structured approach imposes consistency across sessions.

A hybrid of these techniques is the semi-structured interview, where the content to be covered is pre-determined, but the order in which it is elicited and the wording of the questions can vary. This allows unknown or surprising aspects of the domain to be revealed. Conversation is less stilted and the interviewer can adopt the vocabulary of the interviewee where appropriate. The additional flexibility of the semi-structured interview allows the knowledge engineer to discover the interviewee's associations between topics which would not be available from a structured interview and to note answers to questions where they are offered, even if they arise in answer to a different question. We chose to employ a mix of interviews and observational methods in our knowledge elicitation study, as interviews are useful for the understanding of general principles, rules, background material and for consideration of rare events, whereas observational methods can generate detailed contextual material.

Review of the data an expert handles is another approach to knowledge elicitation which builds support knowledge and can fill in gaps in the knowledge base. Whilst we found it useful to review documented decisions, for example an Edit Decision Log (as shown in table A.1 in appendix A) and the content of highlights packages broadcast on television, such material cannot shed light on *how* the expert assesses different parts of the data and resolves conflicts in order to reach their editorial decisions. For this we needed to employ an observational method: protocol analysis.

Protocol analysis, or the "thinking aloud" approach, involves the expert performing a task and verbalising his or her thought process, which is recorded on audio tape or in written notes for later analysis. This record, or protocol, facilitates an understanding of the expert's use of knowledge and their reasoning strategies, embedded in the work context [Kidd (1987)], but is limited by the fact that the expert may not verbalise all the knowledge that they are employing in the decision-making process. Some knowledge may be subconscious and implicit assumptions that the expert makes may not be mentioned. Although it has been suggested that verbalisation can interfere with the thought process and task performance of an expert, Ericsson and Simon (1993) present evidence to the contrary. They demonstrate that there is little difference between the knowledge that is elicited when information is collected during or after the task. According to Smagorinsky (1989), the advantage of the Think Aloud Protocol is that, unlike other knowledge elicitation tools, it can provide significant information about the internal structures of cognitive processes. We used a protocol analysis method to complement our expert interviews, as it gave us more detailed information about how and why specific events were included in a summary. Direct observation also meant that details could be more accurately recorded, rather than relying on an expert to recall information from memory.

3.2 Soccer knowledge elicitation in the literature

A few interviews and studies have been reported in the literature which have a bearing on our knowledge elicitation research. An interview with a sports commentator (the BBC's John Motson) [Deeble (2004)] showed that the knowledge employed for football match description includes players' names, shirt numbers and biographical notes, along with the scores of the two teams' previous clashes. Background research for a game takes about a day and a half, using reference books and videos of the teams' recent matches. The commentator said that people often try and trip him up with

football trivia, which suggests that such background knowledge is very important for viewers.

As part of the RoboCup project, Frank (1997) carried out a statistical analysis of the database of football match articles from the Times newspaper, in order to identify the important descriptive features of a football match. Significant features include object-based features (represented by words such as *league, manager, game, cup, players, ball, club, team* and *referee*); event-based features (like *offside, goal, header, shot, free-kick, scored, save, injury, penalty, one-two* and *cross*); locations (such as *forward, midfield, long, post, away* and *back*) ; and qualitative descriptions (like *lost, defeat, draw, won, beat* and *skill*). We use many of these features in our ontology described in section 4.2.1.

Frank also raises the issue of *beauty* in a football match by quoting Barnes (1996):

> You can move the salt and pepper pots how you like to show the chess-like manoeuvring that led to the great strike, but that is to miss the point. You do not stroke your chin and say: hmm, that was a *good* goal. It is a cry of wonder from deep inside your guts.

Although it is difficult to quantify, and Frank argues that for RoboCup, winning is more important than beauty, the newspaper quotation that is used to justify this, "there are no pictures on a scorecard" [Truss (1997)], actually makes the opposite point for us. In our football highlights application, beauty *does* matter, as users are already likely to know the score, for example via text message alerts. They will pay for a highlights service in order to see a rerun of the action, especially skilled or beautiful moments. Our soccer event descriptions developed in section 4.2.1 do not include the skill or beauty of an event, as it is beyond the scope of most event recognition algorithms at present, but it is an issue worth bearing in mind for future work in this area.

The user requirements analysis in the BUSMAN project to access data from dis-

tributed multimedia databases [Evans (2003)] was based on interviews with football fans and hence has relevance for our application. The study found that people remembered a large number of details about football matches, especially for the team or teams that people supported. Each fan in the study supported one team and sometimes had an interest in the progress or one or two others. Team history was important, in particular, fans were interested in watching material covering pivotal games where the fortunes of a club changed or were under pressure. The users expressed the wish to be able to find specific events that involved a particular player: they talked about being interested in watching memorable moments of action involving players like Pele or David Beckham. This justifies the inclusion of the *favourite player(s)* property in the user profile we develop in chapter 6. Team and player names were also found to be important enough to remember, along with which teams were playing in which leagues. Football fans often remembered scores of past matches, as well as controversial or interesting events in a match. The fans ranked events in priority order: goals were the most important, followed by major referee decisions, sendings off, fouls, the build up to and celebrations following a goal, interviews with goal scorers or man of the match, and finally controversial incidents. The "boring bits" with the lowest priority were when the ball was in open play, or the team was trying to probe the defence.

Evans found that fans are very keen to receive highlights of football matches when mobile, as the key activities are only a small part of the game. They are less willing to pay to view the less interesting parts, for example when the ball is in open play. Currently it is difficult to obtain footage of recent goals: they tend to only be shown on particular TV programmes at particular times. One person talked about the frustration he felt at not being able to see a particular goal (a scissor kick by Steve McManaman against Real Madrid) that everyone was talking about for months, after missing both the match and highlights programmes. Fans reported that they would

like to retrieve soccer clips, for example in the pub, when there was a disagreement among their group of friends about what had happened, or to show someone, "Did you see that?"

3.3 Knowledge elicitation methods

The aim of our knowledge elicitation study was to understand what types of knowledge an expert brings to bear on the problem solving process in information filtering.

The study consisted of four stages. Firstly, we carried out a face-to-face interview with a New Media Development Producer at BBC Sport Interactive. The aim of this phase was to understand the motivations, reasoning and decision-making process employed by sports editors when they compile edited highlights of a football match or series of matches such as the World Cup. A semi-structured interview approach was taken, as the flexibility of this method offered the best opportunity to gain an overall picture of the editing process and sports highlights domain. Following advice in Diaper (1989), the interview was recorded and transcribed verbatim. This ensured a better understanding of the material through listening in detail to the interviewee's responses. The interview transcription is listed in Appendix A.1.

Secondly, we interviewed a sports journalist from a newspaper, in order to elicit more information about the narrative process and background knowledge held by a domain expert. By interviewing a print journalist, a sports correspondent from The Times newspaper, rather than a television editor, we were able to concentrate on the knowledge structures involved in sports summarisation, rather than the mechanics of video editing. The aims of the semi-structured interview were to find out what types of stories were told about soccer matches, both before and after the game had taken place. We also wanted to clarify what background knowledge was used to generate the stories, and what the important points of an example story were.

The third stage of the knowledge elicitation study was a data review, with the

aim of becoming familiar with the types of decisions the expert would be faced with when editing football matches. It consisted of three types of data: firstly we analysed different match reports of the same game, to determine whether all journalists applied the same narrative structure to the match, or if different narratives could be applied to the same set of events. Ten different match reports of the same match (Manchester United vs West Bromwich Albion in the 2002-03 Premier League) from various newspapers (Scotland on Sunday, the Observer, Sunday Mail, Independent on Sunday, Sunday Mirror, News of the World, Sunday Telegraph and Sunday Times) were analysed. By using so many different reports, we could see whether there was a consistent view on the story and the important events. Comparing "before and after" reports of a match showed us what *a priori* expectations there were, and what background knowledge was already known before the game took place, as well as what alternatives could be offered in a personalised service. The second data review compared national and fanzine match reports of the same game in an attempt to understand the nature of bias in narrative, and hence to elicit knowledge about the variations in personalised highlights. The third type of data we reviewed came from television broadcasts of football matches and highlights packages of different lengths. Several matches from different broadcasters (Sky, ITV and the BBC) were analysed, along with highlights programmes of different lengths. This showed us the types of events that were included in and excluded from the highlights, as well as their durations.

The final stage of our study was a protocol analysis of two editors' tasks and decision-making process as they edited a highlights package of a Manchester United versus West Ham game in real time for the BBC Interactive Service. The Interactive Service offers various options in addition to the main broadcast, including a rolling highlights sequence, which allows viewers who tune in late to catch up with what has happened in the match so far. Our protocol method involved observing the motor and eye movement of the experts, recording the actions they carried out, and their

comments about their actions. The Interactive highlights editors had no control over camera angles, slow motion or replays, but were dependent on one live video feed and two audio feeds: the "clean effects" background sound recorded in the stadium, and the commentary from two BBC commentators. This was particularly useful for our work, as we are also generating highlights from a single information source. The task of one of the experts was to record the shots that were used, and their timings, and instruct the second editor to carry out the actual editing. This required the frame numbers of the *in* and *out* points to be typed into an editing machine, which played out a continuous loop of clips.

From the protocol analysis we learnt how events were prioritised and according to which properties (duration, player involved, skill, event class etc.), as well as investigating the narrative structures concept further. Since some events were initially included, and then dropped later on, in order to fit the time limitation of the highlights, we analysed which properties made certain events less important in relation to others. A transcript of the protocol analysis, edit decision log and screenshot of the highlights package description in the BBC's editing software is shown in appendix A.2.

3.4 Knowledge elicitation results

Rather than describing the results from each stage of the knowledge elicitation study individually, we report on the outcomes of the study as a whole, citing evidence for each conclusion as necessary, which may be drawn from more than one source or expert's view. We report firstly on the knowledge elicitation problems we faced, then discuss experts' opinions on the content that should be included in a summary. Our findings on the timing of individual clips in the summary and the overall length of a highlights package are discussed in section 3.4.3. In section 3.4.4 we explain the process of editing the soccer highlights and the problems of co-ordinating audio and

video clips. Section 3.4.5 examines the metadata descriptions and entities commonly associated with soccer, such as players or event types, in order to build our soccer ontology in chapter 4, and in section 3.4.6 we explain how the experts evaluate the quality of a summary. Sections 3.4.7 and 3.4.8 report on the knowledge experts use to construct a highlights package that is based in the sporting domain and television and narrative domains respectively, and we finish with details of how a football summary may be personalised.

3.4.1 Knowledge elicitation problems

As described in section 3.1, it is well understood that the fundamental difficulty in knowledge elicitation is to access the expert's implicit knowledge. This was reinforced by the responses of the BBC producer to the questions of how highlights were chosen. Comments such as, "It comes down to knowing it" and "that [goal clip] feels a bit short" or "How long does it feel right?" were heard, along with the belief that, "there's a certain kind of personal judgement to all of it."

The protocol analysis had to be modified because the editors were under significant time pressure to edit the current clip into the on-air loop as fast as possible, so they did not have sufficient time to verbalise their thought processes fully. However, their comments and directions to each other explained their reasoning behind the activities quite clearly. Uncertainties were cleared up by post-match questions.

3.4.2 Content

The BBC producer told us that editors are given no specific verbal or written criteria about what content to include in their highlights. However, the "talking points" of the game such as major fouls, controversial incidents or decisions, especially incidents near the goal line, glaring misses and serious injuries, particularly those of significant players, should be included. For example, David Beckham's foot injury overshadowed even the goals. The key moments are said to be ones that change the course of the

match, and ones you want to relive and tell your mates about in the pub. This suggests that those events which introduce a new subplot, or change the choice of primary narrative are the ones that should have priority. "Magic Moments" demonstrating particular skill, for example a hole-in-one in golf, are enjoyed by viewers, while funny bits, such as dogs running on the pitch, and glaring misses are also included in the summary. Events such as a yellow card followed by a free kick, near-misses, a corner and a missed penalty shot in the final few minutes of the game were shown in the highlights package we reviewed. More near-misses were shown if the match was goalless. Close ups on players were shown following their goals or goal attempts. It is also interesting to note what was not shown: not all free kicks, throw-ins or corners were included, although those which followed a goal attempt were more frequently included, and no mid-distance views at pitch-level. These were mainly included in the full length match to reflect the stadium atmosphere, but it was difficult to appreciate any overview of the action from them, so they were not shown in the highlights.

All goals should be included in a summary, as "people really like to watch goals", according to the BBC producer. However, some "Wonder Goals" particularly stand out, such as Michael Owen's against Argentina, and Gianfranco Zola's against Norwich in an FA Cup match (where he kicked the ball with the back of his heel into the net behind him), which people want to watch again and again, as they are particularly unusual or skillful. Highlights programmes are content-led: a goalless draw is almost never interesting. In contrast, a programme solely consisting of goals, such as an FA Cup Goal Round Up, is very popular. Goals were always mentioned in the newspaper match reports if they occurred. In goalless draws, more shots on goal were included.

In the protocol analysis it was found that shots on goal were more likely to be included if they came from the losing side, or the side that had made fewer goal attempts; if they were short clips or if they were very near misses. The protocol analysis showed us that some events could be predictably included in the highlights:

the Line Up, the Kick Off and the Final Whistle, although the Line Up and Kick Off were later dropped. According to the highlights editors, goals, the most important event, and at least one corresponding slow-motion replay were also always included. Additional slo-mos were added if there were only a few goals.

The question was posed during the protocol analysis whether events included earlier would be taken out, especially in the game we observed which contained so many goals. There were a couple of shots on goal that were taken out, but the two Manchester United failed goal attempts were kept in, because one hit the post. According to one editor, "If it hits the post, ... it should stay in" and the other attempt was considered "pretty good" and described as "a scorcher" by the commentator. The two West Ham attempts were relatively important, because they were their only two chances.

> "We are trying to tell what happens in the game... so if you tuned in late, you get an idea, maybe not just the goals. For instance, if West Ham had scored their two goals, and Man United had missed all theirs, you'd still show a lot more Man United shots because... it was their game really. But if there are lots and lots of goals, then you have to run with that, even if they are really good shots. But as long as it's under five minutes, I think for a game that has six goals, you can keep in a couple of decent chances. And we did take out all the chat from the very beginning." (BBC Sport editor during the protocol analysis)

3.4.3 Timing

The brief of the editors at BBC Interactive was to produce approximately three minutes of highlights of the match so far, which were played to air in a continuous loop. The highlights could be longer than this, if content so dictated. The length of highlights produced during our observation, at 4 minutes 37 seconds, was considered

very long, but justified because there were so many goals. According to the editors who were being observed in the protocol analysis, events could be reduced in length as necessary, to keep within the overall time constraints, but a clip of too few frames (less than one second) is considered stuttery and messy. Goals were 16-20 seconds long and slow-motion replays 6-13 seconds.

Event length depends on the flow of the game. A goal could be preceded by a quick set piece or might follow a big movement of many players, particularly in the more skilled teams such as Arsenal. The timing of a clip should reflect the action that led to a goal. A goal often stemmed from a pass that was "a little bit out of the ordinary", according to the BBC producer, and this would be where the clip should start. This type of unusual motion was contrasted with kicks in rugby, which are very similar each time, and hence not included in the summary. Television is constrained to show some things that aren't interesting, to keep the flow. Tricks to blend the highlights together include cutting to a picture of the manager, then back to the game at a later stage. A Shot event that was shown on the TV highlights we reviewed was shortened by two seconds, compared to the replay shown in the full length match, presumably due to timing constraints on the summary. In short highlights for the News (one or two minutes long), goal events last for about ten seconds each, but in longer highlights, such as the ten to twenty minute-long packages for ITV's Premiership programme, sequences of several minutes are shown, which include the run-up to the goal, and the events that have caused it.

During the protocol analysis, an open interview was conducted with the more experienced sports editor during half time and after the match. He was skeptical about the value of automatic highlights editing that was not content based. If event length was simply a function of time and not of content, this could result in stutters and "flash frames". That is, if the *in* edit point is ten seconds before the goal, and the *out* point is 3 or 4 seconds after the goal, but the live feed has just cut again to

a close-up a few frames prior to the *out* point, we see only the flash of the close up for a few frames, which is regarded as very poor editing. If event length is fixed to a certain time, we would at the very least need to make sure that we were only dealing with one camera shot.

3.4.4 The editing process

For a live broadcast, the director on location will choose which camera to cut to next, depending on how the ball is moving around the pitch and which camera angle gives the best view of the action. There are around twelve cameras at a football match, depending on which sports ground it is, how many cameras are allowed in and the importance of the match. For example, an FA Cup Final will warrant more cameras than a third division game. Slow Mo or Super Slo Mo are selected if they allow fast-moving events to be analysed more carefully. An isolated shot (where the camera is trained on one person) is chosen to emphasise certain players, or a wide shot to illustrate an overview of the action.

The time available for editing affects what can be done, with an afternoon match highlights package for a 10:30 pm programme being more finely edited. These packages can be ten or twenty minutes in length, with one or two minute sequences used on the News. Editing takes an hour to an hour and a half, with more time being spent on packages that include voice-overs, as the timings must then be rehearsed carefully so that the pictures and flow of motion match the words. The alternative is to use clean effects or the original match commentary. The difficulty in real-time editing lies in co-ordinating the pictures and sound, when there is no time to script a voice-over. For example, the live commentator at the match will be talking to the live pictures, some of which the highlights editor wants to cut out. The editor cannot use the commentator's audio as they are in mid-sentence at the video cut point. When there is no real-time constraint on editing, the editor has time to change the various sound tracks, to fade out the commentator's voice at the cut point and so on, but for the

Interactive Service, where speed is of the essence, clean effects have to be used over the video edit point. A major difficulty, beyond the scope of our research, concerns the editing of the audio. Canned clean effects cannot be used, because commentary interspersed with the wrong type of crowd noise would sound wrong, and would not represent the stadium atmosphere following a goal. Other sounds, for example the crowd booing at an unpopular player, would also not be represented correctly.

3.4.5 Metadata descriptions

When the footage is logged in the BBC Archives, metadata is added to allow searches, but it is manually generated by the editor and does not conform to any formal standard. Tags observed on tape labels were as simple as Club Name, Date and a brief content description such as "Passing and Catching" or "Scrum" (for a rugby match). While creating match highlights, the editor will fill out an edit log by hand, which contains information such as: Editor Name, Date, Names of Football Players and Names of Commentators. Each edit is logged with the timecode and a brief description of the clip's content and shot type, such as slo-mo, isolated shot (of a single player) or long shot.

3.4.6 Professional judgement of summaries

The experts were questioned on their opinions of the highlights shown on different television stations, so that we could gain a better understanding of how they evaluated the quality of a summary. The journalist prefers the BBC highlights as they have much better flow. However, sometimes they miss key events, and he is infuriated when a specific incident that he wanted to look at again isn't shown. This is compounded when the studio commentators discuss that particular event which wasn't included in the highlights. This indicates that there is no single correct summary, and any measurements we make comparing our output to broadcast highlights need to take this uncertainty into account. ITV is "too jumpy", "picks moments you wouldn't

pick" and often does not reflect the game as it happened, according to the journalist.

3.4.7 Sporting domain knowledge

It is clear that the knowledge base must consist of information from both the sporting and the television domain. The producer interviewed employed *a priori* knowledge about different football teams, significant players in a team and the importance of certain matches to each team. She also used her knowledge of soccer tactics to recognise standard team formations, when the build up to a goal had begun and set-piece tactics following restart of play, for example, Kick Offs, Goal kicks, Throw ins, Free kicks, Corners and Penalties. She had an implicit understanding of what was "expected" during play, such that glaring misses, major fouls, lucky goals or other unusual incidents would be included because they were unexpected.

It was found that the journalist holds a huge amount of background knowledge in memory, but tends to forget details of events within matches that are unlikely to have relevance for his current work, such as events from the previous World Cup. He researches details such as how many goals were scored in each team's previous match, and other information about their recent performance. If he has worked on a team recently, he will remember their background, so will only have to look up information about their current opponents. For example, when covering a match between Coventry and Brighton, he'd seen Coventry the week before, so he knew most of their background. Therefore he only needed to check the local paper on Brighton, to find out the most recent news on their team. He found that one player was about to transfer out of the club, so after the game he interviewed him. The background information is more useful for the feature articles written two days after the match, than the match reports themselves. For example, information about a club's financial difficulties could merit a whole article on the Monday following a Saturday match.

Longer term background knowledge is known implicitly (it does not have to be

looked up, but is retained in memory) and remains static. Only recent changes to this knowledge must be checked by the journalist, such as injuries, suspensions, current team line-up or transfer rumours. The sources of the background information are unstructured, for example the wire copy, articles in the newspaper's archives or web pages. Although pre-match expectations are generally not written up as a speculation story, unless for a principal match, e.g. the World Cup or FA Cup Final, they do exist. For example, the data review of Manchester United vs West Bromwich Albion showed that there was an expectation that Albion would lose: their early goal was described as "surprising" and their early lead "shocked just about everyone".

For the protocol analysis, there was no background research carried out beforehand. However, the editors did know who all the players were, and which ones were more significant, as well as recognising the managers of each team.

3.4.8 Television and narrative knowledge

While professionals do utilise a large amount of background knowledge of the sporting domain in the editing process, a major finding of the knowledge elicitation study was that sports producers consider their primary task to be journalism, and their knowledge of how to tell a story is more significant than knowledge of a particular sporting domain.

A producer is given training in how to structure the edited package and illustrate the story visually, as well as how to plan the storyline and decide which shots they will capture and use. It is well understood in television production that "Editing is an essential part of the storytelling art, for it is the process through which scenes and sounds are selected, arranged, and timed in order to impose certain rhythms, meanings and moods on the final result" [Shook (2000)]. The BBC producer supported this opinion by stressing the significance of structure in the highlights package, and the importance of using pictures to illustrate the story. Television highlights have to represent and "give the flavour of the game". The example was given of a team near

the bottom of the league playing against a team at the top of the league. If they lost 5-Nil, but played surprisingly well, preventing ten more goal attempts, it would not be enough to show the five goals as the highlights. The point would also have to be made, through clips of the failed goal attempts, how well the poorer team played. In other words, the story is more than merely the goals and final score. The highlights programme should capture the atmosphere at the ground and reflect the flow of the game. The producer spoke about pacing the clips to represent this flow, for example a goal clip should not be too short.

When we questioned the journalist about various potential football narratives however, he expressed the opinion that matches couldn't be assigned to a particular narrative. Despite using the words "angle" and "line" taken on a story, he would not view a match as being a "lucky win" or a "grudge match". However, he did describe the Spain vs South Korea 2002 World Cup match as being controversial, and mentioned the French matches as being memorable because they had lost when they were expected to win (the "disaster" scenario). Furthermore, the ten match reports we reviewed of the Manchester United vs West Bromwich Albion game frequently used the description "one-sided match". Comments such as "The gulf in class was painfully evident", "Manchester United, away to inferior opposition", "Albion outclassed", "The gulf in class was too great" were widespread. Even the West Bromwich manager was quoted as saying, "United were a class above us." Similarly, the match between Manchester United and West Ham, observed during the protocol analysis fitted the "one-sided match" narrative model. Bets were taken on the outcome of the match, with predictions being 1-all, 2-1 and 3-1 to Manchester United. That is, it was generally expected that the stronger team would be Manchester United. The expected story prior to the game, according to the commentator, was "Can lightning strike twice?" One team had won 14 of their last 17 games, while the other was bottom of the league and had a "dodgy defence". This expected mismatch was borne out by

the events, as Manchester United scored six times. The post-match report was of "an embarrassingly one-sided game" where "Manchester United thrash West Ham" and descriptions such as "carnage" were used.

The finding that particular narratives are commonly present in soccer highlights, is inconsistent with the views expressed by the journalist, and may be attributed to two causes. Firstly, it may be due to the difficulties in eliciting such knowledge from the expert, as this is not their usual mental model for their work. Perhaps it is also not pleasant to think that one's work can be distilled down to a few simple categories. Secondly, and more significantly, it may be due to the difference between thinking of narrative structures as simple categories, versus thinking of them as the top-level story, containing a hierarchy of several potential sub-plots. That is, the narrative of a football match is indeed more complex than a simple categorisation. For example, other angles which were taken on the Manchester United vs West Bromwich Albion match included:

- The consequences of Manchester United's win (taking them to within two points of Arsenal at the top of the Premier League) and Albion's loss (making them vulnerable to relegation.)

- The weaker side beginning well with an unexpected "surprising" goal, taking them into the lead.

- A top player (Ryan Giggs) on the superior team being left out. The reasons were unclear: despite the manager claiming it was due to a hamstring injury, this was considered controversial. However, the player was not missed, and that led to questions that he might be dropped from the line-up in subsequent games.

- A top, but unpopular player (Roy Keane) returning to the superior team after a hamstring injury, and being nominated "Man of the Match".

- Criticisms of the referee's inconsistent decisions.

Despite the journalist's opinion, narratives such as Confirmation of a Team's status, Return of an Important Player, or Expected Behaviour were all apparent in the data review. However, multiple storylines were present in the reports, rather than one single narrative. In chapter 6 we examine how the summary changes if there is a person-centric rather than event-centric interpretation of the game. (That is, if the summary was focused on a particular player rather than important events, does the narrative change?)

The editing style between the main sports broadcasters (Sky, ITV and the BBC) was said to vary little, although Sky usually had more cameras available. The differences were in the commentary rather than the editing. This suggests that it is the journalistic style, the storytelling, that varies between broadcasters. Anecdotal evidence suggests that the ITV Premiership programme received complaints about its match summaries because the commentary did not explain the action well enough. In other words, pictures alone cannot fully tell the story of a match: background knowledge is needed.

3.4.9 Personalisation

The results in this section reflect what experts believe users might want to personalise, therefore the findings may not be as accurate as a large scale study with users themselves. The BBC producer suggested that viewers might want to see highlights only of their own team's play, plus goals from the other side, but no other action from the opposing team, as the interest would lie primarily with their own team. However, the BBC has no plans to offer more personalised content than is already available via the Interactive Service.

We reviewed pre and post match reports to understand the differences in information content between the two, and what alternatives could be offered in a personalised

service. Previews of the game, written the day before, published on the morning of the match, tend to focus on one person, a player or the manager. Speculation about what is expected to happen in the game is rare in these articles (that is, the reports are person-centric rather than potential-event-centric). Such speculation is limited to finals or international matches. This does not mean that the experts don't have an expectation in their minds about what will happen, as seen by the pre-match banter during the protocol analysis.

Post-match reports fall into two time categories, one the day immediately following the game, and the other two days after the game. The following day's article is a match report, concentrating on what happened in the game. The first part of a match report will review the result, mention the goals and the goal scorers, and include a quote from a manager. The body of the report will describe what happened, and will conclude with comments on the result and the skill of play. Since it is dull to just list events in print, only the most salient will be included. Key moments are defined to be those that change the dynamics of the game, for example, goals, near misses, or events that fire a team up, such as controversial incidents, suspensions, punch ups, bad tackles and fouls. Articles appearing on the Monday following a Saturday game contain little about the actual game, and even the goals will not be described, as there is the expectation that fans will already know the result and have seen the game. Instead, these articles will take another angle on the game, such as concentrating on the club's financial problems.

Discussing the recent launch of football highlights for third generation mobile phones, the journalist was reminded of when SMS goal alerts were first introduced. Users were receiving the alerts several days later, after they had already heard the results of the matches. The highlights must be timely, or their value is lost. Many football club web sites offer video clips, mainly just showing the goals and near misses. Although some do show full video highlights, it was pointed out that the content of the

highlights you would want to see would change over time, and thus time lag following a match should also be a feature in the user profile. Although full highlights wouldn't always be necessary, especially for mobile applications with limited bandwidth, it is an option requested by users according to the requirements study in Evans (2003).

Different reporters may "take different angles on the match": some are more "flowery" and "wax lyrical" in the way they describe the goals. Although some reporters may be negative about one or the other team, it is not considered good reporting: national newspapers should report the game objectively. A neutral report reflects the proportion of shots each team had during the game. For example, if one side dominated the game and had many more attempts on goal than their opponents, more of these shots should be included in the highlights. However, a biased report would spend longer on their team's goals or shots, and only mention the other team's shots in the context of their own goalkeeper's success in keeping those shots out.

We also studied match reports of the Fulham versus Birmingham City match in the 2003 FA Cup third round to understand the nature of fans' requirements for personalisation. The neutral report from the BBC web site was compared against a Fulham and a Birmingham City report from two fanzine web sites.

The narrative in the BBC report was of a one-sided match, as Fulham "overwhelmed" Birmingham City and the "result never looked in doubt". A secondary narrative was that a significant player, Louis Saha, was returning to Fulham after four months off for a hamstring injury. The report mentioned the consequences of Fulham's win: they were through to the fourth round of the FA Cup. The three Fulham goals, and one Birmingham goal, a goal attempt by each side, and a save by the Birmingham goalkeeper were all mentioned, as well as Sava's unusual celebration by wearing a black mask following his goal.

The match report by a Fulham fanzine began by talking about the three Fulham goals by Saha, Goldbaek and the "masked man" Sava. It was described as

Birmingham City's "worst game for 6 months". Also included was a description of Birmingham player Robbie Savage complaining to the referee about his decision to award a corner to Fulham: he "jumped up and down on the spot and would have started crying", after which he was booked. Other events listed were two saves by Fulham, two Fulham goals before halftime, another Fulham goal ruled out for offside, and a Fulham near miss. Birmingham's goal was mentioned briefly, as being "too late".

The report in the Birmingham City fanzine described their team's "pitiful performance" with only one goal during injury time. The relatively high number of Birmingham fans that had travelled, compared with the Fulham home team's supporter turnout was highlighted. Fulham's goal by Sava with his "inexplicable and rather childish Mask Celebration" was mentioned. A controversial decision against Birmingham was described, then Fulham's next two goals. A Birmingham near miss was mentioned, and the many near misses of Fulham were criticised as "wasted". The Birmingham City's captain was praised, and their only goal described, at more length than in the Fulham fanzine.

Different events will be shown in a personalised fan's match highlights. Although some, such as goals, consequences of the result, and the most unusual incident, will be included in all, different lengths of time will be assigned to them, depending on the bias of the reporting. For example, a fan may prefer longer clips of their favourite player. Event length can also be modified to fit the total duration of the highlights, as specified in a user's personal profile. The less significant events may be included or left out, according to the bias. Schadenfreude, the enjoyment of the opposition's suffering, motivates the inclusion of the other team's red or yellow card events, or other referee rulings that go against that team. For example, the inclusion of the opposition captain complaining to the referee and getting booked for his efforts (an incident that was left out of the fanzine report from the captain's own club), and

the controversial incident where the referee ruled against the home team, was also included in order to criticise the decision.

3.5 Discussion and design recommendations

The knowledge elicitation study has shown that the manual editing process is labour-intensive, taking up to an hour and a half per match. If we consider that about fifty matches take place per week in the English leagues alone, and, as we observed during the protocol analysis, two editors are needed to generate each highlights package, along with additional requirements for personalisation of highlights, this adds up to a substantial effort if manual editing is used. It can also take several days to research the background knowledge for a single live game [Deeble (2004)], so there is a strong motivation for a system to automatically generate soccer highlights.

Evans (2003) has shown that fans have a considerable interest in viewing highlights, especially when personalised to present events of their favourite teams and players. We found that the background knowledge base should consist of information such as players' names, shirt numbers, their current and previous clubs, as well as imminent transfers, along with players' recent form or injuries. The importance of certain matches to a team, their position in the league and previous scores should also be recorded. Experts implicitly use such a background knowledge base when they report on a game or edit highlights, to place the current events in context for the viewers. This is considered the mark of a good highlights package.

The knowledge elicitation study has illustrated some of the constraints we will have to impose on our system. Although the BBC does not use a formal ontology description of football metadata, experts have an implicit understanding of what types of events are expected in a game. Our event ontology will consist of these "expected" events, such as Goals, Corners, Free kicks, Offsides etc. However, it is the *unexpected* events, such as a dog on the pitch, Sava's "mask celebration" or Rooney's

shorts falling down, that are always included in the highlights. It is difficult for an information extraction system (whether text, audio or video based) to distinguish between a rare, but important event and signal noise. Therefore a constraint that we have to place on our system is to not include the concept of "unexpected event" in our ontology.

Skill, beauty and "magic moments" are highly prized in football, but are difficult to quantify, and would rely on speech recognition of the commentator's words describing the particular merit of an event, for example, the "scorcher" of a shot described during our protocol analysis. Scoring the skill of an event is beyond our scope. Our system is also unable to address a major difficulty which the experts have, the editing of the audio track to coincide with the pictures.

The knowledge elicitation study has clearly shown that some events are more salient than others. Goals are the most important, but shots on goal (especially skilled shots or those initiated by the losing side) are also salient, followed by any controversy (serious injury, major fouls, or referee decisions, especially when involving an important player). The decision of which events to include depends on the total allocation of time resources. Events can be reduced in length, or cut out completely, if more interesting events occur later in the match. There is no single correct summary, making evaluation difficult, but the experts expect the highlights to "reflect the game as it happened".

The user profile should consist of the favourite team (plus one or two others that the fan may be interested in), favourite player(s) and the types of event the fan wants to see. In chapter 6 we also use the event priority order elicited by Evans (2003) as part of the user profile. Since our database does not contain interviews with players or managers, but only the football games themselves, these will not be included as "events" in our ontology. Although our results suggest that the time between the match taking place and the highlights being shown should also be a parameter in

the personalisation of highlights, we will not include it in our experiments, as we do not have access to the additional footage (such as manager interviews or background stories on the clubs) that would usually feature in highlights programmes several days after a game.

Sports highlights generation requires far more than just background knowledge about the game of football. A producer also needs skills in television editing and an appreciation of what constitutes a well constructed highlights package. The major finding of our knowledge elicitation study is that summarisation is best performed using a narrative structure. This does not simply mean labelling the game as one category or another, with a single templated structure for that narrative, but that a game can be seen as a mixture of subplots, which can come to the fore or be demoted, if we are personalising the highlights in a particular way. Several narratives are repeated consistently, e.g. the stronger team is expected to win; there is an unexpected change in behaviour of a team or player; or the strongest player is missing from the team. The episodic structure of the narrative allows the flow of play to be represented and places events in context. As well as showing the important events themselves (e.g. goals), the events that cause them (e.g. the assisting passes) and events that they cause should also be included in the highlights. In chapter 4, we introduce a model to represent such causal relationships between events.

Since we would like our system to eventually be coupled with real audio/video input, in the next chapter we will be describing our ontology design for the soccer domain, reflecting many of the events expected by the experts. This then specifies exactly which semantic concepts the audio or video extraction module should be trying to recognise. A second criterion we have is that our summariser should be extendable and applicable to other domains. In section 7 we present preliminary experiments that map our soccer domain ontology on to the ontology of the business meeting domain, to examine the extent to which the episodic structure of a soccer

summary is applicable to a business meeting summary.

We found in this study that the experts mobilise a large amount of soccer knowledge, so we develop a knowledge base containing background knowledge of the domain, as well as using our training data to learn the causal relationships between events. The third criterion of section 1.1 is that the narrative structures should deliver coherence and context, a requirement that is supported by the experts in this knowledge elicitation study. By modelling the causal relationships between events, we are able to increase the coherence of our summaries.

3.6 Knowledge elicitation summary

In section 3.1 we introduced some knowledge elicitation methods, and presented reasons for the choice of a combination of semi-structured interviews, data review and protocol analysis in our study. Section 3.2 discussed previous results in soccer knowledge elicitation, while our own results were presented in section 3.4. Some design decisions were discussed in section 3.5, and these will be implemented in subsequent chapters.

We begin in chapter 4 with the design of a soccer event ontology, and a description of the information extraction method we use to establish our training set. We also report on experiments where groups of causally related events model coherence in a case based reasoner which summarises soccer games.

Chapter 4

Case based reasoning

> *A case-based reasoner solves new problems by adapting solutions that were used to solve old problems.* [Reisbeck and Schank (1989)]

Case based reasoning [Kolodner (1983) and (1993)] uses solutions of previously encountered situations to help solve the current problem. It appears well suited to our application as we have many such cases of soccer matches available, the 'problems', and their corresponding highlights, the 'solutions'. We believe that case based reasoning (CBR) can offer major advantages to us over other reasoning methods because it has been shown to perform well in interpreting open ended and ill defined concepts [Kolodner (1991)]. It is also good at proposing solutions for applications like ours where understanding of the domain is incomplete. For example, we can never hope to encode all personal preferences or all the information about every possible action and player in a soccer game.

Most other reasoning approaches do so from the ground up each time, minimising the contribution of prior knowledge. Therefore a second reason for choosing the CBR approach is because it incorporates a structured knowledge base within a reasoning context, allowing us to make use of the large amount of prior knowledge we have available. According to Watson and Marir (1994), CBR has an advantage over rule based systems because it does not require an explicit domain model or set of rules. This means that knowledge elicitation is simpler, as only case histories need to be

gathered. A CBR system can be maintained more easily than a rule-based system, as new knowledge can be acquired by adding new cases to the case base, rather than needing to identify which rules to update.

This chapter describes the case based reasoning system we have built using cases derived from soccer match reports on the web. Motivated by work in the literature, discussed in section 4.1, in section 4.2 we describe our soccer event ontology, along with our method for case retrieval and adaptation. We introduce the concept of a *context group* of events to represent event causality, and in section 4.3 we present the results of a series of experiments to evaluate the performance of our weighted retrieval method and context group based adaptation. A discussion of the results follows in section 4.4, while section 4.5 characterises the coverage and regularity of our case base to diagnose sources of error. Finally, in section 4.6, we present a summary of the conclusions drawn in this chapter.

4.1 The CBR cycle

Case based reasoning originated from the work of Schank and Abelson (1977) on scripts which recorded general knowledge about situations, allowing people to set up expectations and perform inferences. However, this concept did not describe human memory processes completely, as it was found that people often confused incidents which had similar scripts. Schank (1982) later proposed the dynamic memory model and Memory Organisation Packet theory of problem solving and learning, which was implemented by Janet Kolodner at Yale University in the first CBR system known as CYRUS [Kolodner (1983)]. CYRUS stores events in the life of the former US Secretary of State, Cyrus Vance, as a hierarchical structure of episodic Memory Organisation Packets, also called generalised episodes, (GEs). These are groups of specific cases sharing similar properties, along with norms (features common to all cases indexed under a GE) and indices (features which discriminate between a GE's cases). The

entire case memory is structured as a discrimination network where a node is either a GE (containing the norms), an index name, an index value or a case. A case is retrieved by finding the GE with most norms in common with the problem description. Indices under that GE are then traversed in order to find the case which contains most of the additional problem features. In this way, CYRUS can answer questions about events involving the secretary of state by reconstructing episodes using the retrieval indices in the query to traverse the GE. Later research developed the use of cases with model based reasoning [Koton (1989)], and CBR has since been applied to many different domains, such as medical diagnosis [Nilsson and Sollenborn (2004)], legal judgements [Aleven and Ashley (1996)] and recommender systems [Smyth and McGinty (2003)].

A case consists of a *problem description* and a *solution description.* For example, in our application the problem consists of a description of the football game, along with a requirement for highlights of a certain duration. The solution is the description of events that were included in the football highlights. A case can also include an explanation of why the solution was selected, or potential causes of failure.

Case based reasoning can be modelled as a four stage process. [Aamodt and Plaza (1994)] as shown in figure 4.1:

- **Retrieve:** The most similar case or cases to the current problem are retrieved.
- **Reuse:** The case or cases are reused, sometimes with adaptation, to solve the problem.
- **Revise:** The proposed solution is revised, if necessary.
- **Retain:** The elements that are likely to be useful for future problem solving are retained by adding them to the case base. This is the CBR learning method.

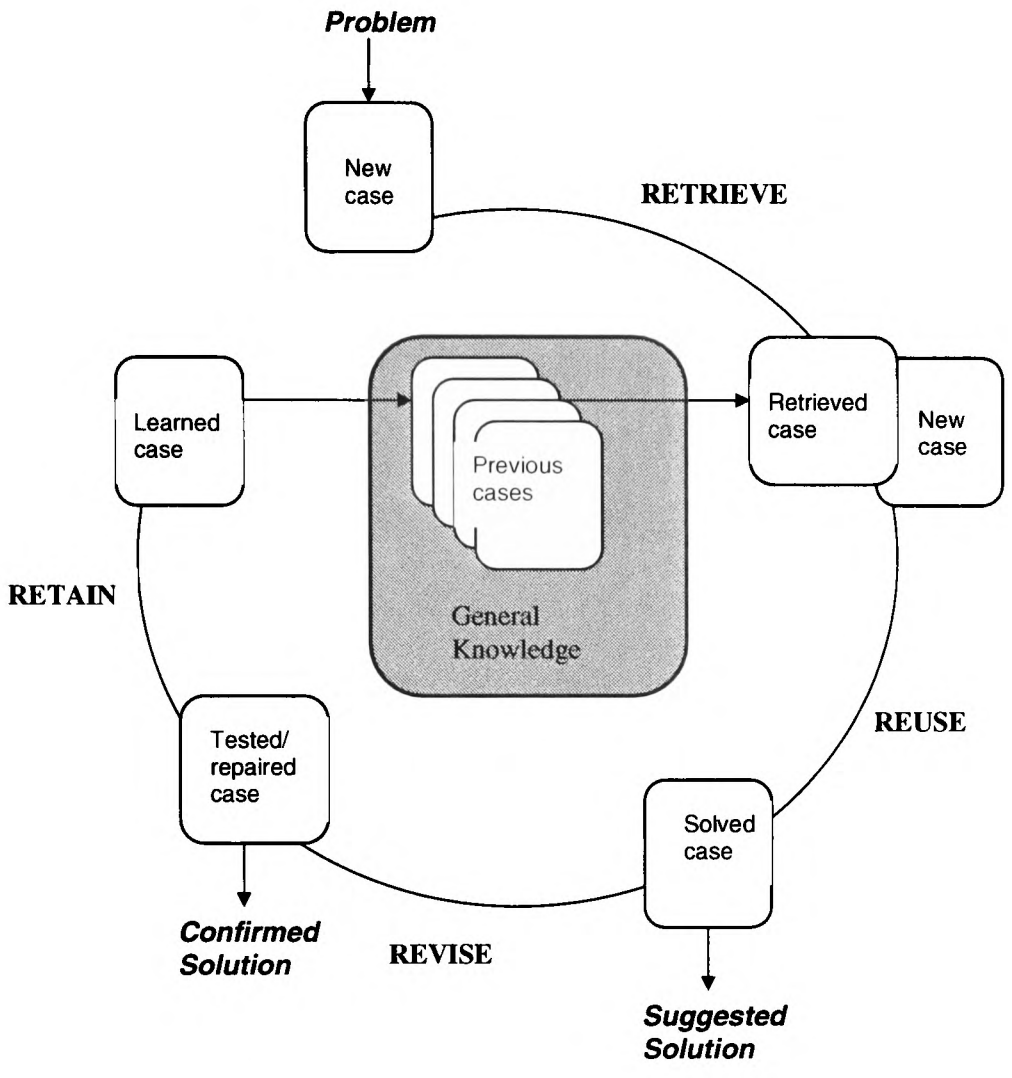


Figure 4.1: The CBR Cycle [Aamodt and Plaza (1994)]

4.1.1 Case retrieval

The case retrieval task consists of selecting one or more cases from the case base which are most similar, in some way, to the current problem. For example, in the soccer highlights domain, we want to find a case in the case base that describes a similar football game to the one we are currently trying to summarise. Large case bases may require efficient search algorithms or hierarchical case organisation to facilitate quick retrieval, such as the Fish and Shrink model [Gebhardt et al. (1997)], kd-trees [Bentley (1975)], or Case Retrieval Nets [Lenz and Burkhard (1996)]. However, since our case base will be small, we choose the simplest, albeit slowest option of a flat memory structure, using serial search. Cases are stored sequentially, and a matching function is applied to each case in turn, such that the most similar case or cases are

returned.

The best known and simplest method for case retrieval is the k Nearest Neighbour (kNN) algorithm. This assigns a class to the input datum based on the classes of the k most similar data points. The most common choice of similarity measure is to use a weighted sum of absolute differences of features. The weightings allow us to identify which features are more important in helping to predict the solution. For example, the Goal type event in a football game is likely to be an important feature, as goals are always shown in the highlights, so *Number of Goals* is likely to be a good predictor of events. (The more goals there are, the less likely other types of event are to be included in the highlights, since there is less time to spare for them.) The retrieval indices may be stored directly in the case, or, as with the *Number of Goals* example, an index value may be derived from case features. In this example, the index value is a simple summation of the Goal event features in the case. CBR retrieval differs from standard kNN classification in that an additional adaptation step often needs to be carried out on the retrieved cases before a classification decision can be made. Note that in our application, each case that is retrieved is effectively a different classification decision, as two cases rarely give exactly the same output after adaptation. We simplify this to two classes[1]: the case, or set of cases, which give the best adaptation result and the set of non-optimal cases.

The drawback of using k Nearest Neighbour is that retrieval time increases linearly with case base size; however we will be working with only a small case base and hence can easily use this method. Using the terminology of Wettschereck and Aha (1995), each case $x = \{x_1, x_2, ...x_n, x_c\}$ consists of a set of n features (which can be numeric or symbolic) and a classification x_c. Given a query q, kNN retrieves the k most similar cases to q from the case base of L cases. The most similar case is defined as the one

[1]The term "class" when discussed in relation to the k Nearest Neighbour algorithm refers to the set of optimally or non-optimally retrieved cases, and is different from the term "class" when referred to in the ontological sense of different event types or classes.

which is the shortest distance from q, where:

$$distance(x, q) = \sum_{i=1}^{n} w_i \cdot difference(x_i, q_i) \tag{4.1}$$

where w_i is a weighting assigned to feature i and:

$$difference(x_i, q_i) = \begin{cases} |x_i - q_i| & \text{if feature } i \text{ is numeric} \\ 0 & \text{if feature } i \text{ is symbolic and } x_i = q_i \\ 1 & \text{otherwise} \end{cases} \tag{4.2}$$

The probability $p(q,c,k)$ that q is a member of class c is defined as:

$$p(q, c, k) = \frac{\sum_{x \in k \ , \ x_c = c} 1/distance(q, x)}{\sum_{x \in k} 1/distance(q, x)} \tag{4.3}$$

Numeric features are normalised by subtracting the mean and dividing by the variance, to ensure each feature has the same range (and expected impact).

The k Nearest Neighbour similarity function is sensitive to irrelevant and noisy features, if features are given equal importance (i.e. $\forall i,\ w_i = 1$). By applying importance weightings to each feature, more appropriate cases can be retrieved. Wettschereck and Aha (1995) have compared a number of domain independent feature weight learning algorithms, which can be grouped into the feedback and ignorant methods. Feedback methods can be divided into two groups: incremental hill climbers and continuous optimisers. The hill climbing method modifies feature weights to increase the similarity between a query and nearby cases in the same classification, and to decrease its similarity with nearby cases in other classes. When a correct classification occurs, if feature i matches, then its weight w_i is incremented by Δ. Mismatching features have their weights decremented by this same amount. For incorrect classifications, the weights of mismatching features are incremented, while the weights of matching features are decremented. This approach is limited as it can become stuck in a local minimum, and since each case is processed only once, it is sensitive to the order of presentation. The second type of feedback method, the continuous optimiser, iteratively updates feature weights using randomly selected training cases, for example

using a genetic algorithm, which avoids the problems of presentation order and local minima.

The second type of learning algorithm investigated by Wettschereck and Aha are termed ignorant methods. These assign feature weights using conditional probabilities or mutual information, rather than modifying feature weights based on feedback from previous retrievals. This type of learning needs to have enough training examples of the different classes c and features i to make the probability estimations accurate. This requires a large case base and a limited number of classes and features. The Per Category Feature importance method (PCF) assigns high weights to features that are highly correlated with the given class: $w_i(c) = P(c|i)$. The weight for feature i for a class c is the conditional probability that a case is a member of c, given the value of i. The Cross Category Feature importance method (CCF) takes the sum of the squared conditional probabilities of the C classes, using:

$$w_i = \sum_{c=1}^{C} P(c|i)^2 \tag{4.4}$$

The CCF is not sensitive to the distribution of a feature's values across classes. Instead, it assumes a feature's weight is independent of the class, whereas the PCF assigns higher weights for features that are highly correlated to the given class. Using Shannon mutual information, feature weights can be assigned using the reduction in uncertainty of one variable's value (the class, c), given knowledge of the other variable (the feature value, v):

$$w_i = \sum_{v \in V_i} \sum_{c \in C} p(x_c = c \ , \ x_i = v) \cdot log \frac{p(x_c = c \ , \ x_i = v)}{p(x_c = c) \cdot p(x_i = v)} \tag{4.5}$$

Mutual information has the advantage that, unlike the PCF or CCF, feature values do not need to be in binary form. In our application however, it is difficult for us to establish accurate probabilities for some of the feature values which occur rarely, for example, the *by* property of each event can take an unbounded number of values (any player's name).

For us, a drawback of all these learning methods is the limited amount of data per class that we have available. For some queries, there may only be one case that gives the best adaptation output. At most, we could only expect two or three cases that adapt the query to generate the same optimal[2] output. In our system, as well as looking at the Cross Category Feature weighting method, we also investigate a simpler approach by choosing each soccer event type as a retrieval index, and calculating the conditional probability of that event type being included in the summary, given that it has occurred in the case. That is, we have now only two classes, Included and Excluded, and a larger body of data from which to learn. In section 4.3 we compare this simpler method with the Cross Category Feature weighting.

4.1.2 Case reuse

The solution of the retrieved case can be reused in order to solve the current query, taking into account the differences between the retrieved case and the current problem. Aamodt and Plaza (1994) distinguish between transformational reuse[3], which reuses the past case solution, and derivational reuse, which reuses the method of solving the previous case. Derivational reuse needs the plan or method for solving the previous case to be stored in memory along with the solution itself, so that the plan can be replayed when solving the current problem. The drawback of this approach is that it requires sufficient understanding of the cases in order to represent the solution methods well. In transformational reuse, we assume that knowledge exists in the form of transformational operators which convert the old solution into a solution for the current case. The operators can be indexed according to the differences between the retrieved and current case.

The simplest transformational adaptation technique is null adaptation, or simple copying. This gives the same solution to the current problem as was stored in the

[2]'Optimal' is defined here in terms of the highest precision and recall rates of all candidate solutions generated by adapting every case in the case base.

[3]Elsewhere [Kolodner (1993)], this is known as structural adaptation.

retrieved case. However, in soccer summarisation this isn't possible: you cannot generate highlights for an Arsenal versus Manchester United game using only events from a Southampton versus Portsmouth game!

Instead, in our system we apply the transformational adaptation techniques known as reinstantiation and parameter adjustment. Reinstantiation is used to instantiate features of the retrieved case's solution with new feature values. For example, if the retrieved case highlights contained a Goal by Thierry Henry, we could replace this with a Goal by an equally famous player in the current case, for example, Michael Owen. Parameter adjustment is then applied where direct instantiation is not possible. The parameter adjustment technique compares certain parameters of the retrieved and query cases in order to modify the query in an appropriate direction. For example, if no Goal can be found in the current case, a similar event, such as a Shot on Goal, is used instead.

4.1.3 Case revision and evaluation

Case revision takes place when a case solution generated by the reuse step is incorrect, and we wish to learn from the failure. The theoretical case revision step first evaluates the case solution, and if it has not succeeded in solving the problem, it is repaired using domain specific knowledge. Evaluation is usually carried out externally to the CBR system, for example, asking users or experts to subjectively evaluate a summarised solution, or carry out a task such as retrieval of information using the summary (the intrinsic or extrinsic evaluation strategies described in section 2.2) and it can take some time before the result of the evaluation is known. The majority of our evaluations will therefore be carried out intrinsically, using expert-generated artifacts, i.e. comparing the soccer highlights output from our system with those edited together by a sports producer and broadcast on television. Here, for the sake of simplicity and tractability, we are making the assumption that all experts summarise a football game in the same way, or at least, over our case base of soccer

games and highlights, the editors' personal choices are averaged out. This means that our system is learning a general form of the summarisation process, rather than being tailored to a specific expert, as it was not practically possible to gather enough examples from a single expert for our experiments. So when we compare our output, which is that of "the average expert", to the highlights package edited by a particular expert which we are using as a "ground-truth", it may appear that there are errors due to our generalisation. However, since users are already used to viewing highlights generated by many different editors, and find their different styles acceptable, for a viewer, we believe that the "generalised" style of our summaries would be equally acceptable to a user. In other words, our objective evaluation method is likely to be a harsher judge than a subjective evaluation would be. For example, a particular sports editor might select a certain Foul for inclusion in the highlights, to illustrate that it was a particularly violent game, while another expert may choose a different Foul to illustrate the same point; the fan watching the highlights is mainly interested in the fact that the game was violent and would have found either choice acceptable.

In order to repair a case solution, reasons for the errors in the current solution must be suggested. This is beyond the scope of our work, as we are unable to generate the domain-specific knowledge about error avoidance that is needed for this step. It would require extensive interviews with experts to establish, for each case, *why* a particular event should have been chosen over another one. We neither have the resources to carry out this knowledge elicitation, nor, given the findings in Chapter 3, the expectation that the expert could articulate the answers to such questions anyway.

4.1.4 Case retention

Case retention is the learning stage of the case based reasoning system, where the useful aspects of a case that has just been solved are stored in the case base for future use. The process involves selecting the information to retain from the case, indexing

the case for later retrieval by similar problems and integrating the new case into the memory structure. If the new problem was solved by using a previous case, it may be added directly to the case base as a new case, or the previous case may be generalised to encompass the newly solved problem as well. If the newly solved case is situated in a well populated region of the feature space, we may not wish to retain it, as that area is adequately covered by cases already in the case base. However, if it lies in a sparsely populated part of the feature space, we are more likely to retain the case.

As well as adding new cases to the case base, we may wish to delete or modify cases or indexing information to improve the CBR system's performance: a process known as case base maintenance [Leake and Wilson (1998)]. There are a variety of policies for case base maintenance, the simplest being *standard case learning*, a policy of always adding each new case to the case base. Smyth and McKenna (1999) have suggested a method to improve the compactness of the case base whilst maintaining case base competence, "the range of target problems that can be successfully solved". The authors present a policy for case base maintenance that uses an explicit case competence model based on coverage and reachability, which measures the competence contributions of each case. The coverage set of a case is the set of all target problems that this case can be used to solve. Conversely, the reachability set of a target problem is the set of all cases that can be used to solve it. They define a measure of "relative coverage" which estimates the unique competence contribution of an individual case c, that is, cases covered by c that are not covered by other cases. To grow the case base, the cases, ordered according to competence contribution measured using relative coverage, are processed using a nearest neighbour algorithm that retains only those cases that are not solved by a case already in the case base. Portinale et al. (1999) remove cases that have not been retrieved recently, along with false positive cases, that were retrieved, but frequently failed to solve the problem. Zhu and Yang (1999), in contrast, add cases to the case base in succession based on how

much the new case adds in terms of usefulness to the current case base. We will not be looking at case retention strategies in depth, as we do not have enough cases for a separate training set and test set, so that case retention experiments could be carried out using the test set. However, in section 4.5 we look at how different sizes of case base affect results, using a policy similar to Zhu and Yang's, where cases are added to the case base in order of usefulness, specified in terms of the number of times a case is retrieved.

Wilson (2001) measures the quality of the case base using two types of regularity. *Problem-solution regularity*, which represents the assumption that if a previous problem is similar to a current problem, the solution to the previous problem will also be similar to the desired solution to the current problem; and *problem-distribution regularity*, which represents the assumption that new problems encountered in the domain bear some similarity to previous problems. That is, that the current problem is similar to at least one case in the case base. Leake and Wilson (1999) define problem-solution regularity using a set of problems $Q = p_i, p_{i+1}...p_j$. After each problem p_k is processed and its solution evaluated, it is added as the k^{th} case to the updated case base B_k. *PDist* is the problem distance function: the retrieval similarity metric that measures the distance between a new problem and the problem description of a stored case, and *RDist* is the "real distance" function, which measures the usefulness of retrieved solutions according to the evaluator's goals for the retrieval process. *RDist* can be calculated off-line to determine the retrievals that the CBR system *should* have made. For example, we can use the difference between the output of our system and the highlights shown on television as *RDist*, measured in terms of precision and recall (combined using the F_1 measure defined in section 1.3).

Leake and Wilson then define a neighbourhood of Closest Cases to Problem (CCP): all cases within a case base B whose problem descriptions are closest to the query problem: and Real Closest Cases (RCC): the cases whose solutions are within a

user-specified neighbourhood of the optimal solution. The size of the neighbourhood is determined by a user-specified non-negative parameter ϵ.

$$CCP(PDist, p, B) = \{c \in B | PDist(p, c) = min_{c' \in B} PDist(p, c')\} \quad (4.6)$$

$$RCC(RDist, p, B, \epsilon) = \{c \in B | RDist(p, c) \leq min_{c' \in B} RDist(p, c') + \epsilon\} \quad (4.7)$$

Then, the probability that a case returned as optimal by the similarity function will actually be within ϵ of an optimal case, is measured as:

$$SimPrecision(PDist, RDist, p_k, B_k, \epsilon) = \frac{CCP(PDist, p_k, B_k) \cap RCC(RDist, p_k, B_k, \epsilon)}{CCP(PDist, p_k, B_k)} \quad (4.8)$$

and problem-solution regularity is defined as the average value of *SimPrecision* over the problem sequence Q, starting with case base B_i:

$$ProbSolnReg(PDist, RDist, Q, B_i, \epsilon) = \frac{\sum_{k=i,...,j} SimPrecision(PDist, RDist, p_k, B_k, \epsilon)}{j - i + 1} \quad (4.9)$$

Problem-distribution regularity is the percentage of cases in a problem sequence $Q = p_i, ..., p_j$ for which there are sufficiently close cases in the current case base B_k built up from the seed case base B_i, according to the user-specified distance limit $\epsilon \geq 0$.

$$ProbDistReg(Q, B_i, \epsilon) = \frac{1}{j - i + 1} * \sum_{k=i,...,j} \begin{cases} 1, & \text{if } min_{c \in B_k} PDist(p_k, c) < \epsilon \\ 0, & \text{otherwise.} \end{cases} \quad (4.10)$$

These measures are employed in section 4.5 to characterise the quality of our case base.

4.1.5 Temporal and causal CBR

Work in the literature [Gupta et al. (2002)] has shown that representation of causal relationships between case parts improves retrieval. Gupta et al assume however that the CBR engineer will specify fixed causal relationships during the case design

process. We would prefer to be able to derive these automatically, and address this issue in section 5.2. Knowledge representations of temporal relationships have also been reported in the CBR literature. Dørum Jære et al. (2002) define a number of temporal relations (such as *before*, *after*, *during*, *overlaps* etc) using interval-based temporal logic. A human-monitored editor generates each case, and the similarity of temporal intervals in the case and query are calculated using a dynamic ordering algorithm. This compares the first time interval in the case with the corresponding interval in the query. If it doesn't match, the second interval in the case is compared against the query's interval, and so on until a match is found with case interval *n*. The second query interval is then compared with the intervals in the case, starting with interval *n+1*. This search is repeated for every interval in the query, so that each interval retrieved from the case is always later in time. Since similar football games do not necessarily need to have similar orderings of events throughout, we would prefer to allow the comparison of every time interval in the case with every other interval in the query, irrespective of its temporal placing in the case. We also considered following Rougegrez (1994)'s method of using the edit distance algorithm for string matching to predict what will happen next in a process, given a similar, already complete process. Edit distance is calculated between one sequence of events and another as a function of the number of Insert, Delete and Substitution operations that need to be carried out to transform the first sequence into the second. The drawback of Rougrezez's experiments are that they operate only on a global (whole case) basis, and the cases and the query all need to contain an equal number of events, which does not fit with our soccer summarisation problem.

4.2 CBR system design

Significant advances in computer vision over the past few years allow us to make the assumption that extraction of semantics from visual information will soon be at a

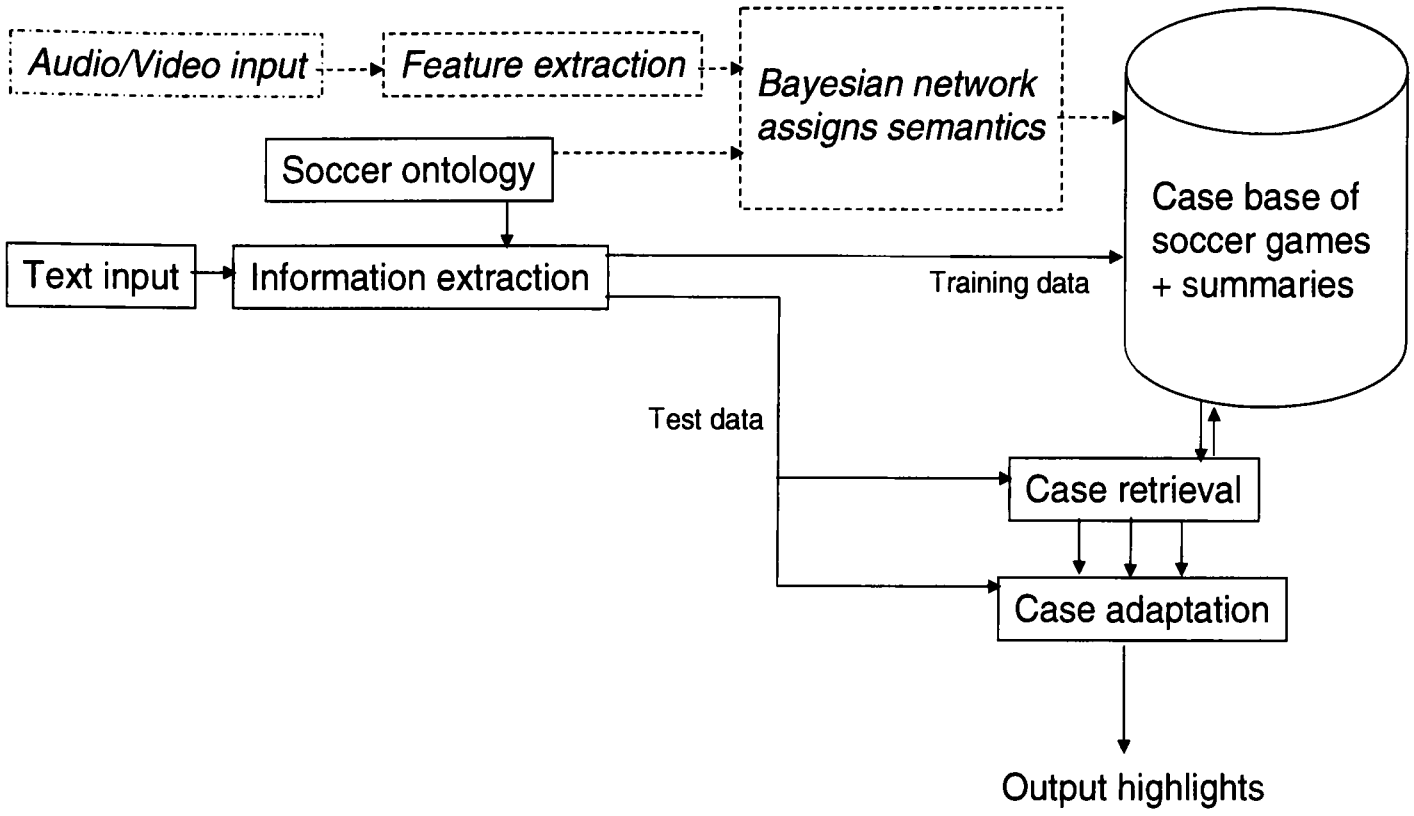


Figure 4.2: The case based reasoning system

stage to deliver metadata descriptions of sports video action. Figure 4.2 is a high level block diagram of our case based reasoning system, with the dotted boxes giving an overview of how semantics might be extracted from audio or video input. This could be achieved using, for example, extensions to the work of Ekin et al. (2003) and Assfalg et al. (2003), which are discussed in section 2.3.

However, since our primary focus is on information summarisation rather than extraction, we sidestep the need to extract information from the audio or video representation of soccer games, and use the minute-by-minute "ticker-tape" reports widely available on many sports' websites. While work has been done on information extraction from free-text soccer reports [Saggion et al. (2002)], and more generally, on information extraction for the creation of CBR cases [Brüninghaus and Ashley (2001) and Daniels (1997)], we avoid this complexity by following Lawson et al. (1996)'s method of *template mining*: extracting information directly from text where there is an automatically recognisable pattern. We use ticker-tape web reports prepared from a template, as shown in figure 4.3. The time stamp on each sentence in the ticker-

90:00 (4:03) Goal kick taken long by Paul Jones (Southampton).
90:00 (3:25) Foul by Matt Oakley (Southampton) on Robert Pires (Arsenal). Free kick taken right-footed by Thierry Henry (Arsenal) from own half, passed.
90:00 (3:05) Attacking throw-in by Fredrik Ljungberg (Arsenal).

90:00 (1:46) Booking

Foul by Michael Svensson (Southampton) on Sylvain Wiltord (Arsenal). Michael Svensson (Southampton) booked for dissent. Free kick taken right-footed by Fredrik Ljungberg (Arsenal) from right wing, passed.

Figure 4.3: A partial web ticker-tape

tape consists of two elements: the start time of the event within the game's main 90 minutes, and any extra time that might be being played. In the ticker-tapes we have used, events are described in reverse order. The time stamp correlates well with the time on the video at which the event took place. This means that our system can simply output the times and durations of all those events we have decided to include in the highlights, and an editing machine could convert this description to a video highlights package.

We acknowledge that these ticker-tapes have, to a limited extent, already been filtered by the web author. The author is implicitly excluding uninteresting events from the ticker-tape, for example, times when the ball is out of play, or moving up and down the pitch under no clear control by either team ("open play"). However, this pre-filtering by the web author is not sufficient to provide a summary of the game: a typical ticker-tape describes about 250 events in the game, so that each event lasts around 20 seconds, and the remaining filtering problem is still significant. We are also limited in the richness of our description by the ontology we use (discussed further in section 4.2.1). We cannot represent events in the summary that are not specified as concepts in our ontology, which, by their rarity, are actually *more* likely to be interesting. For example, the one-off occurrence of Wayne Rooney's shorts falling down is beyond the scope of our ontology, but was considered by the television sports editor to be a joke worth sharing in the highlights. This *frame problem* is a well

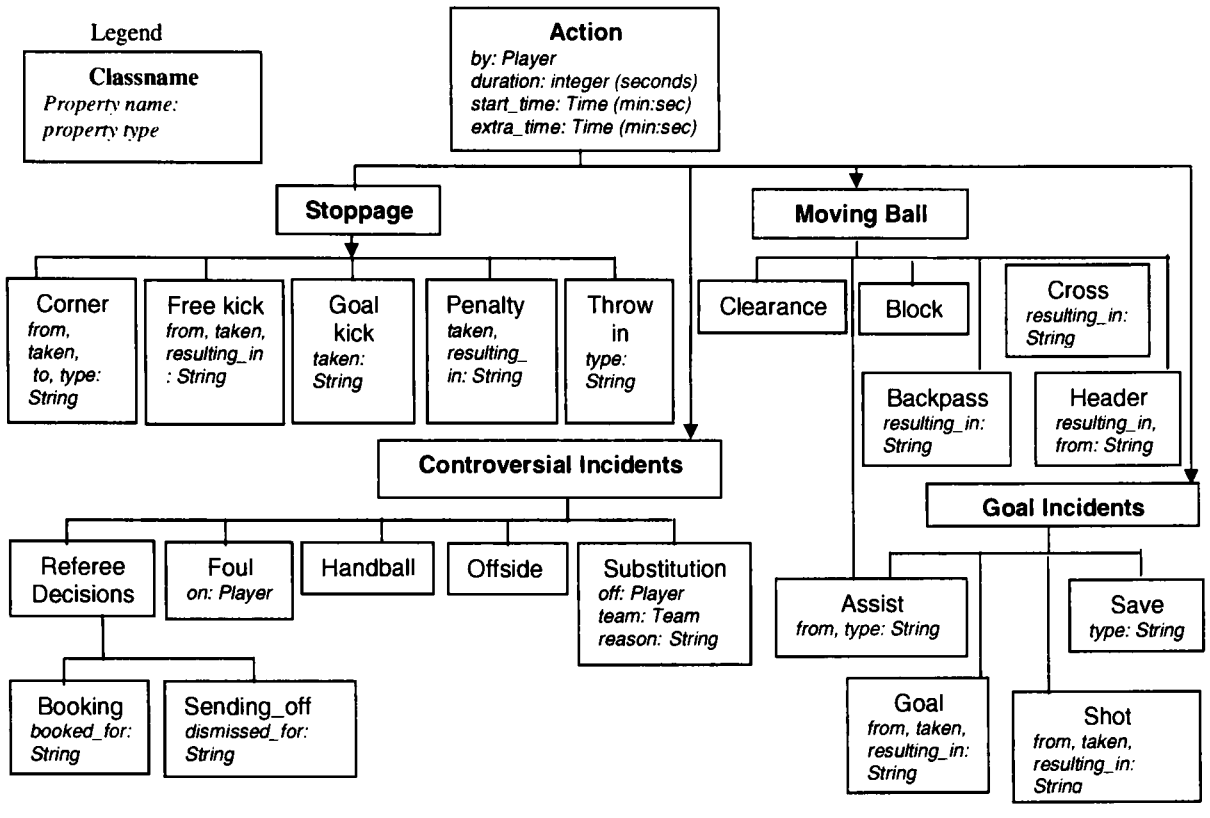


Figure 4.4: Hierarchical soccer event ontology

known trade-off in artificial intelligence: in order for the problem to be tractable, we must make a closed world assumption and limit the scope of what can be represented, rather than try to describe every concept in the world. The ontology that we use is a reasonable compromise between descriptive richness and limited complexity.

4.2.1 Soccer ontology

To design the ontology in figure 4.4, we have used the Protégé 2000 ontology development tool [Protégé 2000] since it was found by Duineveld et al. (2000) to be the tool best suited for the conceptualisation and formalisation phases of ontology design. We then created a knowledge base by defining individual instances of the classes, filling in specific slot value information and additional slot restrictions, by parsing the ticker-tape soccer game reports.

To our knowledge, there have only been two previous soccer ontologies developed, Saggion et al. (2002) and Crampes et al. (1998). The former used 31 event classes, for the purpose of information extraction from free text, while the latter was more

complex, designed to encapsulate the rules of football, for a purpose similar to ours, narrative abstraction. (Crampes et al. (1998) is discussed in more detail in section 2.5.2 in the context of narrative summarisation.) Our event ontology is designed to allow template mining from the ticker-tapes, and hence is limited by the content of the templated ticker-tapes we use. Since the ultimate aim would be to use extracted video semantics as the basis for summarisation, our ontology is simpler than those in the literature. It does not contain the more detailed events such as Kick Off, Scissor Kick, Nutmeg or One-Two that have been included in previous soccer ontologies. As shown in figure 4.4, a second difference to past soccer ontology designs is that the twenty event classes are grouped into a hierarchy. This is for two reasons. Firstly, it allows parameter adjustment in the adaptation stage of our algorithm. For example, if a Goal is not available, it can be replaced by an instance of a sibling class, such as a Shot. Secondly, it allows reasoning on coarser input to take place, if fine detail about the events is not available. For example, this might occur if we were extracting semantics from the video, rather than using a text description. We could expect to identify a Stoppage, Moving Ball, Goal Incident or Controversial Incident from the video, but we might not be able to tell between the subclasses of Throw in and Free kick using current computer vision techniques.

Our case base consists of 126 cases of soccer games representing the 'problems' and their corresponding highlights, representing the case 'solutions'. These solutions are generated in one of two ways: firstly, by selecting all events highlighted in colour on the web page ticker-tape. For example, in the partial ticker-tape of figure 4.3, the shaded Booking event and subsequent events reported with the same start time, would be included as part of the case solution. The second method of generating the case solutions is by manual annotation of events from the highlights broadcast on television. Despite being time consuming to generate, this provides a more accurate representation of the events broadcast in a soccer highlights package, than the first

method. We also include in the case problem description the duration of the summary that is produced, i.e. its length in seconds. In an attempt to capture temporal coherence and causality, we group events in the same paragraph of the ticker-tape into groups, which we have called *context groups*. As our results demonstrate in section 4.3, this innovation improves performance and we believe it is a significant novelty of our approach. For example, in the ticker-tape of figure 4.3, the last paragraph contains a Foul, a Booking and a Free kick event, which are considered to be one context group. This enables us to include events in the highlights which are not important in themselves, but are a cause of the significant event, thus providing context within the highlights. For example we might include an assisting pass, which is insignificant in itself, but enables the receiving player to score a goal. In order to evaluate how much improvement this context-group-based method gives us, we run separate experiments, one using an algorithm operating on one event at a time, and another experiment using context-group-based retrieval and adaptation, for both types of case solution. Later in this thesis, (section 5.2) we look at how we might group causally related events automatically, but for now we rely on the web editor's paragraph groupings.

4.2.2 Case retrieval

The k cases which are the most similar to the test problem are retrieved from the case base, using the k Nearest Neighbour approach. We compare two approaches to case retrieval. Firstly, retrieval similarity, or rather, distance between the case and test, *PDist*, is measured using a weighted sum of absolute differences between the normalised number of instances of each of the twenty event classes (as shown in figure 4.4) in the case and the test problem, as shown in equation 4.11. The weightings w_i are the conditional probabilities of each event class being included in the summary, given that it has occurred in the case (equation 4.12).

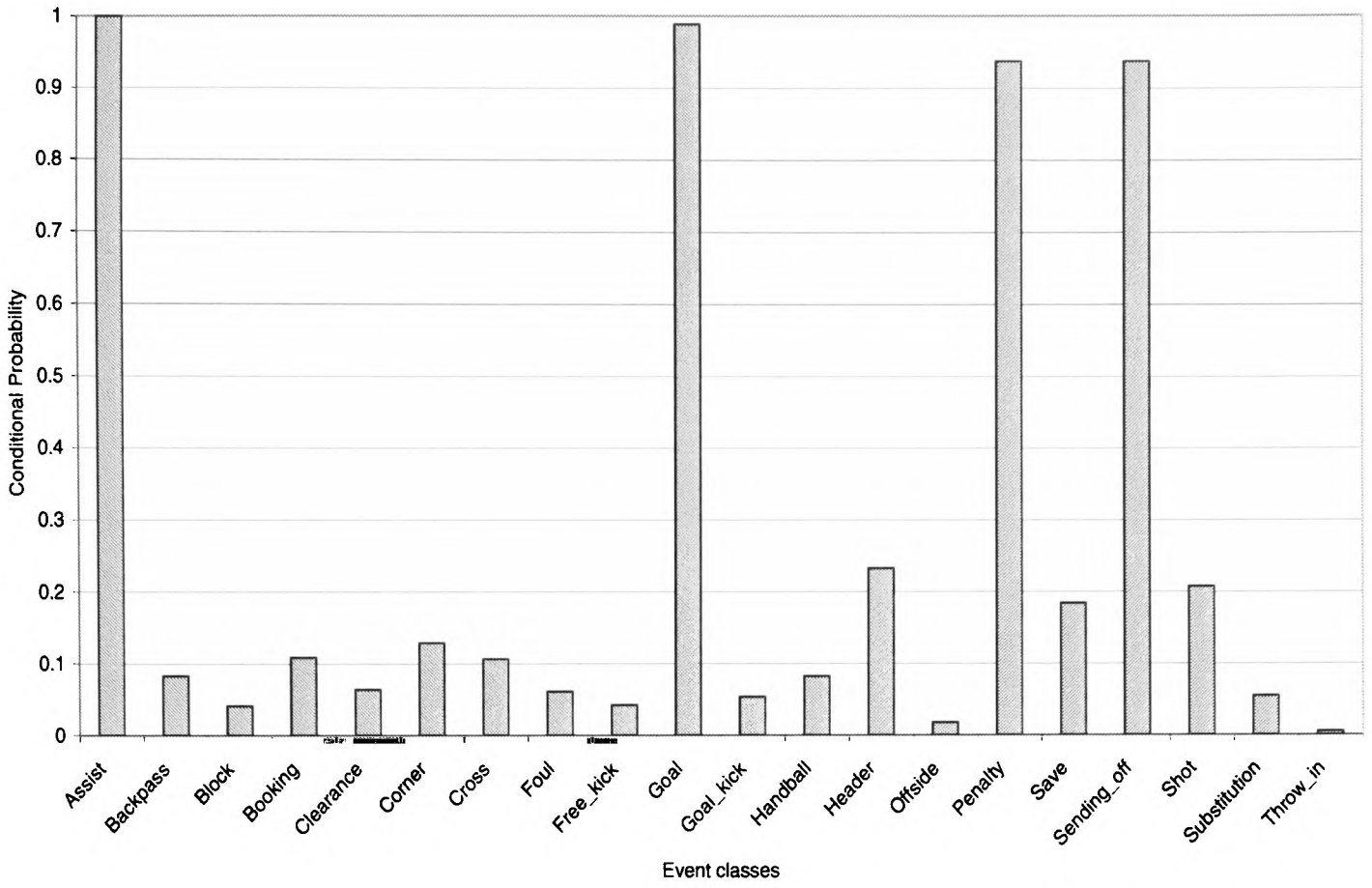


Figure 4.5: The conditional probabilities of each event class being included in the summary, given that it has occurred in the case

$$PDist = \sum_{i=1}^{N} w_i \cdot \frac{|frequency(event(i)_{case}) - frequency(event(i)_{test})|}{var(frequency(event(i)))} \tag{4.11}$$

$$\begin{aligned} w_i &= P(Included|Occurred) \\ &= \frac{P(Included.Occurred)}{P(Occurred)} \\ &= \frac{frequency(event(i) \in case\ solutions)}{frequency(event(i) \in case\ problems)} \end{aligned} \tag{4.12}$$

where *frequency(event(i))* is the number of events of class i and N is the number of event classes in the soccer ontology, (N=*20*). Figure 4.5 shows the conditional probabilities for the case base where one case is left out. (There is little dependence on which particular case is left out.)

The second method of calculating feature weightings for retrieval follows the Cross Category Feature importance method described in section 4.1, and involves similarity

measured between all event features in the case, rather than the event classes alone. As shown in figure 4.4, we have 16 different features (*event class, by, duration, from, to* etc.) Essentially, the purpose of the weightings is to bring *PDist* closer to *RDist*, and increase the problem-solution regularity as defined in section 4.1.4 by reducing the impact of irrelevant features. For each test case, we calculate weightings w_i as follows:

1. Find the set of cases that are optimally retrieved from the case base for a certain case x_1. That is, those cases which give the highest *RDist* values (using the F_1 measure combining precision and recall) when x_1 is adapted to each of them.

2. For each of our 16 features, count the number of times that feature occurs in each optimally retrieved case. Also calculate the frequency of occurrence of each feature across the whole case base. Note that the occurrence of a feature is counted only when it has the same value in both x_1 and the optimally retrieved set. (This allows us to limit the features to binary values.)

3. Then count the number of times each feature occurs in each non-optimally retrieved case.

4. Repeat steps 1-3 for all other cases x in the case base, and find the mean of the feature frequencies. The Cross Category Feature importance measure for each feature i can then be calculated as:

$$w_i = P(c = \text{optimally retrieved cases}|i) + P(c = \text{non optimally retrieved cases}|i) \tag{4.13}$$

where the two categories c are the optimally and non-optimally retrieved cases.

$$w_i = \frac{frequency(\text{feature(i) averaged over cases in the optimally retrieved set})}{frequency(\text{feature(i) in all cases})} + \frac{frequency(\text{feature(i) averaged over cases in the non-optimally retrieved set})}{frequency(\text{feature(i) in all cases})} \tag{4.14}$$

4.2.3 Case adaptation

The second stage of the algorithm is to adapt the test problem to each of the k retrieved cases in turn, and output the test solution which is closest in time length to the required summary duration.

- For each case context group, take the first event, C_1, and find all events of the same class in the test context group, $T = T_1 \ldots Tn$.

  1. If $n = 0$, create a new set T consisting of all instances of sibling classes of C_1 which are present in the test context group. and choose from step 2 or 3. If this new set T is also empty, no match can be found.
  2. If $n = 1$, T_1 is therefore the best match to C_1 in this context group, and we calculate its similarity rating using the Adaptation Similarity Measure described overleaf.
  3. If $n \geq 1$, calculate the similarity of C_1 to each event in T, using the Adaptation Similarity Measure.

  The event with the highest similarity, T_k, is the best match to C_1 in this context group.

- Once a match has been found for C_1, remove this test event, T_k, from further consideration when matching subsequent case events.

- When each event in the case context group has been matched to one in the test context group, or it has been determined that no match can be found, sum the similarities of each matching event pair to find the total similarity of the context group. This value is normalised by dividing by the number of events in the case context group, to take into account any non-matches.

- Choose the test context group which has the highest similarity to the case context group for inclusion in the highlights.

- Repeat this process for each case context group from the case summary in turn.

Algorithm for Context Group Comparison

- Similarity between two events E_1 and E_2, which may or may not be instances of the same ontology event classes, is calculated as the inverse of a difference measure, D, where:

$$D = \sum_{i=0}^{N} |\text{diff}(feature(i)_{E1}, feature(i)_{E2})| \tag{4.15}$$

  where N is the number of features in E_1. Note that this measure is non-commutative, as the summation is made over all features in E_1, which may not be the same as the features in E_2. Since both E_1 and E_2 are from subclasses of the Action class, the Action properties (*by*, *start_time*, *extra_time* and *duration*) can be compared directly.

- If the Player instances in the *by* properties of E_1 and E_2 have the same name (i.e. the strings are equal), the difference score of this feature is zero, else if the players play for the same team, the difference score is 0.5, otherwise it is 1.

- The *start_time*, *extra_time* and *duration* feature differences are normalised absolute differences of the respective property values.

- If E_1 and E_2 are from the same subclass (e.g. both Penalties), other slot values can be compared directly (e.g. String comparison of the *resulting_in* property values). This works because only certain strings (such as "left-footed", "attacking", "tactical" etc) will occur in the ticker-tape authoring template. If the strings are the same, the feature difference is zero, else it is one.

- If E_1 and E_2 are from different subclasses, but have the same properties, they too can be compared in the same way. Thus a Penalty resulting in a Goal will be considered similar to a Free kick resulting in a Goal. The total difference score is normalised by dividing by the number of properties in the case event class.

Adaptation Similarity Measure

Adaptation is carried out on a context-group by context-group basis, rather than by matching individual events, so all causally-related events within a group are included in the highlights, rather than just single, isolated events. This is the main contribution of this chapter: our algorithm provides context by allowing the flow of play up to an important event, like a goal, to also be included in the summary. More

formally, we define a context group to be a group of temporally adjacent events that form a causal chain.

Figures 4.6 and 4.7 show the adaptation process, using an example retrieved case and test problem to be summarised. In the example in figure 4.7, context group 2 is found to be the most similar, so it is included in the highlights output, and removed from further consideration in the adaptation process. That is, we are using a greedy algorithm, both for selection of events within a context group, and for selection of context groups within the whole soccer game. This means that the ordering of events in the test context group is maintained, relative to those in the case context group and similarly, the ordering of context groups is also maintained. The automatic inclusion of *all* events in a context group in the summary means that we could potentially introduce false positive errors, if an event in the context group was irrelevant to the other events in the group. It could also mean we introduce false negatives, if the event immediately after a certain context group was caused by events in the included context group, yet did not have enough importance to force the whole of its own context group to be included. This really relates to whether we have got the groupings of events correct, such that they represent the causality between events. So far, we are relying on the web editor to decide for us, by grouping events into paragraphs in the ticker-tape. In the next chapter (section 5.2) we look in more detail at how the context groups can be learnt automatically.

4.3 CBR system evaluation

As our case base is quite small, containing only 126 cases, we cannot afford to keep the test set completely separate from the case base, so we test the system using a leave-one-out approach. This means that the retrieval weightings of equation 4.12 need to be recalculated for each 'new' case base. The results are evaluated by comparing our experimental output with events in the case solutions. Since our case

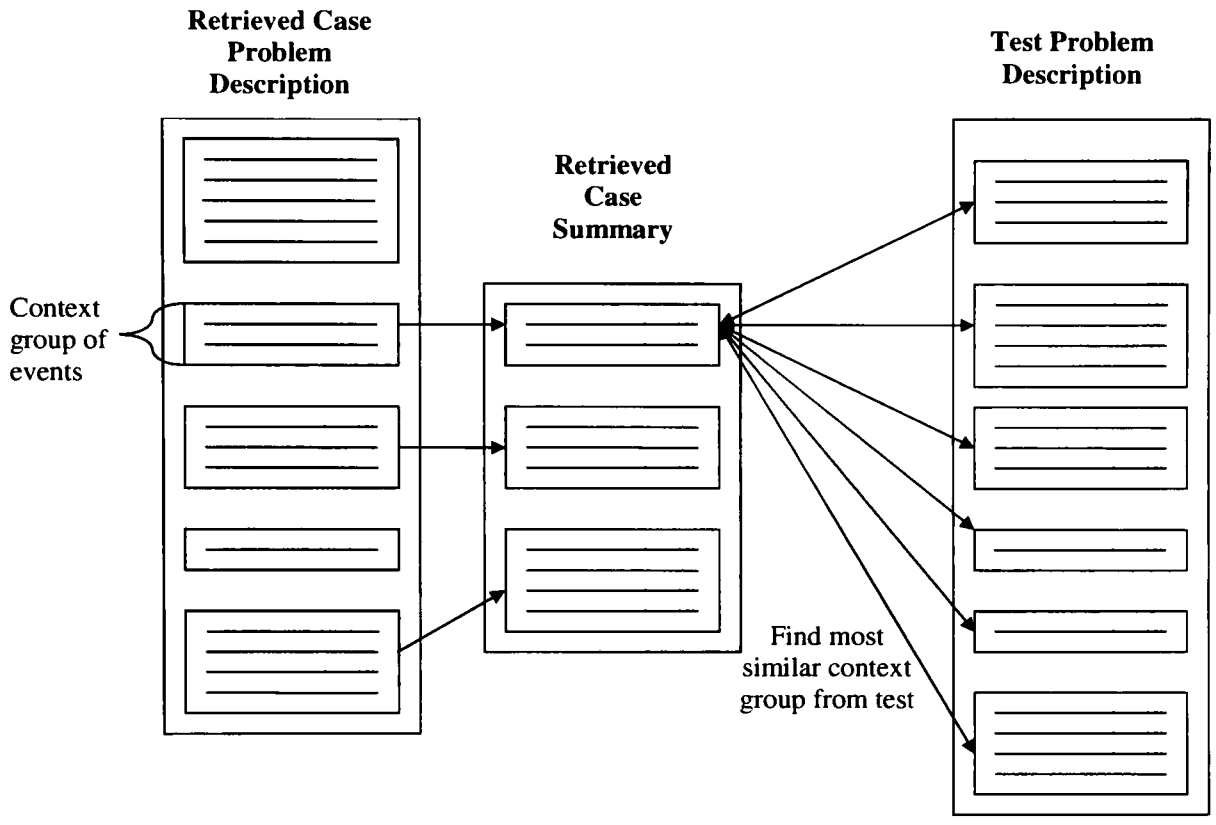


Figure 4.6: The context-group based adaptation process

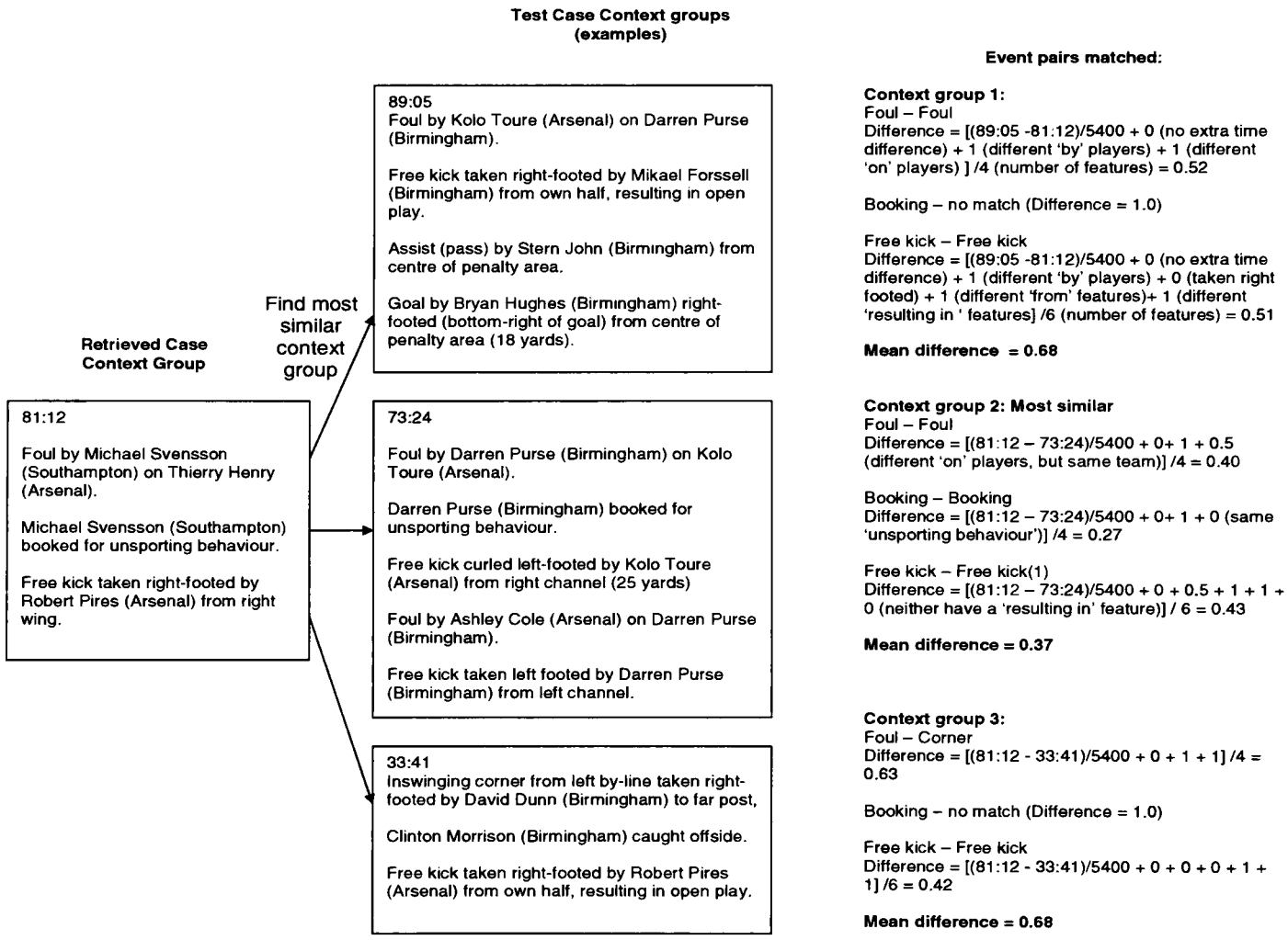


Figure 4.7: An example of the similarity calculations during adaptation for three context groups

solutions were generated in two ways, i.e. the naïve method taken directly from the highlighted events on the ticker-tape, as well as the more accurate method of event annotation from the television highlights, results are obtained for both sets of case solutions. The time required to annotate case solutions from the highlights broadcast on television is about fifteen minutes per case. Although semantic event recognition would render such a manual effort unnecessary, it did impact on the size of the case base we were able to construct, although it gave a more accurate representation of a summary than using the events highlighted in colour on the ticker-tape. We characterise our results using precision and recall as defined in equations 1.1 and 1.2. A 'true positive' is defined as an event which we included in our summary output, which was also highlighted on the ticker-tape, (or shown on the TV highlights, for the second method), while a 'false positive' is an event which we suggested should be shown, but was not. The true positives + false negatives are then the number of events in the 'ground truth summary', that were highlighted on the ticker-tape (or shown on the broadcast highlights).

We also calculate the number of events in our output highlights that were of the correct class. In other words, they are of the same class in our 20-event ontology as the event that should have been selected, even though they may not have been exactly the right event instance. Although an event was not picked by a particular sports editor for the highlights, it does not necessarily mean that it is completely unacceptable to a user to watch it. For example, if we suggest that a particular Shot on goal should be shown, although the expert chose to edit a different Shot, this may still be a reasonable choice, and the 'correct classes' measure reflects this. We also record the difference in total duration of the highlights output by our system compared with the length specified as part of the problem description of each test case, expressed as a percentage of required duration.

Table 4.1 shows mean precision and recall results along with duration error, to

compare which feature weighting method gives the best retrieval: non-weighted, Cross Category Feature weighting (which uses all event properties as features) or weighting using conditional probability of inclusion of event classes only. Since the conditional probability of event inclusion gives the best retrieval results, we use this weighting method for the remaining experiments.

Experiment	**Precision**	**Recall**	**Duration Error**
Non weighted retrieval	40%	47%	6.2%.
CCF weighted retrieval	42%	48%	7.2%.
Conditional probability of event inclusion weighted retrieval	46%	52%	5.6%

Table 4.1: Comparison of weighted and non-weighted retrieval methods: Mean precision, recall and duration results for annotated broadcast highlights

Table 4.2 shows mean precision and recall results, along with duration differences for four experiments: firstly, for random selection of events from each test problem. A uniform probability distribution is used, with no replacement of events once they have been chosen. Event selection is terminated once the required highlights duration has been exceeded. The rest of the experiments in table 4.2 use $k = 25$, (that is, 25 cases are retrieved for adaptation). The second experiment calculates adaptation similarity on an individual event basis, while the third experiment uses context-group based adaptation. The fourth result shows precision and recall for the context-group based CBR when the acceptance criteria are relaxed to include all events of the correct class.

In table 4.3 we present results (using the event class weighted case retrieval and context group based adaptation) where the value of k, the number of cases retrieved, is varied from 1 to 125, to investigate how the precision, recall and duration difference values vary. Experiments are repeated for both types of case solution.

Case solutions from ticker-tape			
Experiment	**Precision**	**Recall**	**Duration Error**
1. Random selection	9%	5%	7.9%
2. Single event based adaptation	53%	28%	10.7%
3. Context-group based adaptation	82%	88%	2.1%
4. Correct classes	92%	96%	2.1%
Case solutions from annotated broadcast highlights			
Experiment	**Precision**	**Recall**	**Duration Error**
1. Random selection	9%	5%	12.4%
2. Single event based adaptation	42%	23%	7.0%
3. Context-group based adaptation	46%	52%	5.6%
4. Correct classes	82%	88%	5.6%

Table 4.2: Mean precision, recall and duration results for event-driven CBR

4.4 Discussion

The experiments on case retrieval weightings (table 4.1) show that using conditional probability of each event class being included gives better precision and recall results than Cross Category Feature weighting, or unweighted retrieval. From experiment 1 in table 4.2 we can see that using any form of prior case knowledge is an improvement over random selection of events. The results for the experiments where case solutions are generated from highlighted events on the ticker-tape are much better than those where case solutions are annotated directly from the highlights broadcast on television. This is to be expected, as the ticker-tapes are written using a computer program that automatically presents certain notable events in colour, namely Goals, Penalties, Bookings etc. Thus the CBR system is easily able to reproduce these relatively simple rules. As it stands, however, performance is much poorer when trying to represent the more complex decision processes of multiple sports editors as represented by the case solutions of the annotated television highlights. Since even the experts do not

Case solutions from ticker-tape			
k	**Precision**	**Recall**	**Duration Error**
1	79%	82%	24.1%
10	83%	87%	7.0%
25	82%	88%	2.1%
50	82%	87%	1.7 %
75	82%	87%	1.4%
100	81%	87%	1.4%
125	80%	86%	1.0%
Case solutions from annotated broadcast highlights			
k	**Precision**	**Recall**	**Duration Error**
1	48%	49%	63.0%
10	44%	49%	10.7%
25	46%	52%	5.6%
50	45%	51%	3.3%
75	44%	49%	2.5%
100	47%	51%	1.9%
125	46%	51%	1.9%

Table 4.3: Mean precision, recall and duration results for k Nearest Neighbour case retrieval

always agree, it can be argued that using their summarised solutions for supervised learning would be unlikely ever to result in 100% accuracy.

A comparison between experiments 2 and 3 demonstrates that adaptation on a context-group basis is more successful, in terms of precision, recall and duration error, than calculating similarity of individual events separately: an increase of 29% precision, 60% recall and 8.6 % duration error on average, using the case solutions derived from the ticker-tape. This is to be expected, because we have included whole context groups of events in these artificially derived case solutions. What is more interesting however, is that there is also an improvement (of 4% precision and 29% recall), when context group-based adaptation is used in experiments using case solutions from the annotated broadcast highlights. This is because the soccer highlights broadcast on television include not only significant events such as goals, but also footage of the lead-up to those events, that is, why and how they happened. The

concept of context-groups enables us to represent these causally related events in our system, thus improving our results. Since the event causality does not, in practice, correspond exactly to the context groups of events as grouped together in paragraphs by the author of the web ticker-tape, the improvement gained from using context groups is smaller when using the case solutions from annotated broadcast highlights. It is interesting that while the recall increases substantially for both types of case solution, the increase in precision is minor for the case solutions annotated from broadcast highlights. This could be because the ticker tape paragraphs do not accurately reflect which events are causally related and ought to be clustered as a context group. Therefore, additional events are being included in the summary as part of a context group, which are inappropriate because they do not causally relate to the rest of the group.

Later in the thesis (section 5.2) we look at how to learn the causal relationships of events from the data in our cases, for example using the frequency of occurrence of certain event classes in pairs or larger groups. The 'correct classes' measure is a first attempt at evaluating the performance of the system beyond a binary "right or wrong", and experiment 4 in table 4.2 shows that our system's performance is reaching 82% precision and 88% recall on average, when the acceptance criteria are relaxed. From the results in the knowledge elicitation study (section 3.4) , we know that there are disagreements over whether the best events have been shown in the highlights, so using the broadcast events as a "ground truth" to evaluate our results against is quite a harsh measure.

The results in table 4.3 show that precision and recall rates are higher for small values of k. Algorithmic efficiency also increases as k decreases. However, these improvements must be traded off against the increase in duration error for small k. When only one case is retrieved from the case base (k=1), precision and recall are slightly down however, and the duration error is much higher, as the duration

parameter cannot be used to select from among the k cases. Conversely, when the entire case base is retrieved (k=125), the only means of discriminating the 'best' case is using the duration parameter, and not the retrieval similarity measure based on numbers of each event type, so again, precision and recall are lower.

4.5 Case base coverage analysis

Since we would like to improve on the results presented in section 4.3, at this stage it is useful to evaluate the quality of our case base. Do the cases offer sufficient coverage of the feature space? How sensitive are our results to the size of the case base and do the CBR regularity assumptions hold?

By adapting each test case to every other case in the case base, evaluating each result against the known solution of each test, and selecting the highest, (in terms of precision and recall percentage), we can effectively eliminate the case retrieval step. We always retrieve the "best" case in terms of precision and recall, and so the results depend only on adaptation success and case base coverage. Each case's frequency of use, as the optimal case to retrieve, is shown in figure 4.8. It can be seen from this bar chart that some cases are the best to retrieve for quite a few query problems, while other cases are not used at all. We then vary the size of the case base, from 1 to 125, using the cases with higher optimal retrieval frequency first. The mean precision, recall and duration errors are shown in table 4.4. Note that we are using k=1 in the k Nearest Neighbour retrieval step, to avoid the necessity of k varying with case base size, so that direct comparisons can be made.

We also use Wilson's problem-solution and problem-distribution regularity measures to compare these different-sized case bases. [Wilson (2001), described in section 4.1]. A value of 1.0 represents a perfectly regular case base, while a value of zero implies that the regularity assumption does not hold at all.

We calculate the problem-solution regularity of our case base, along with problem-

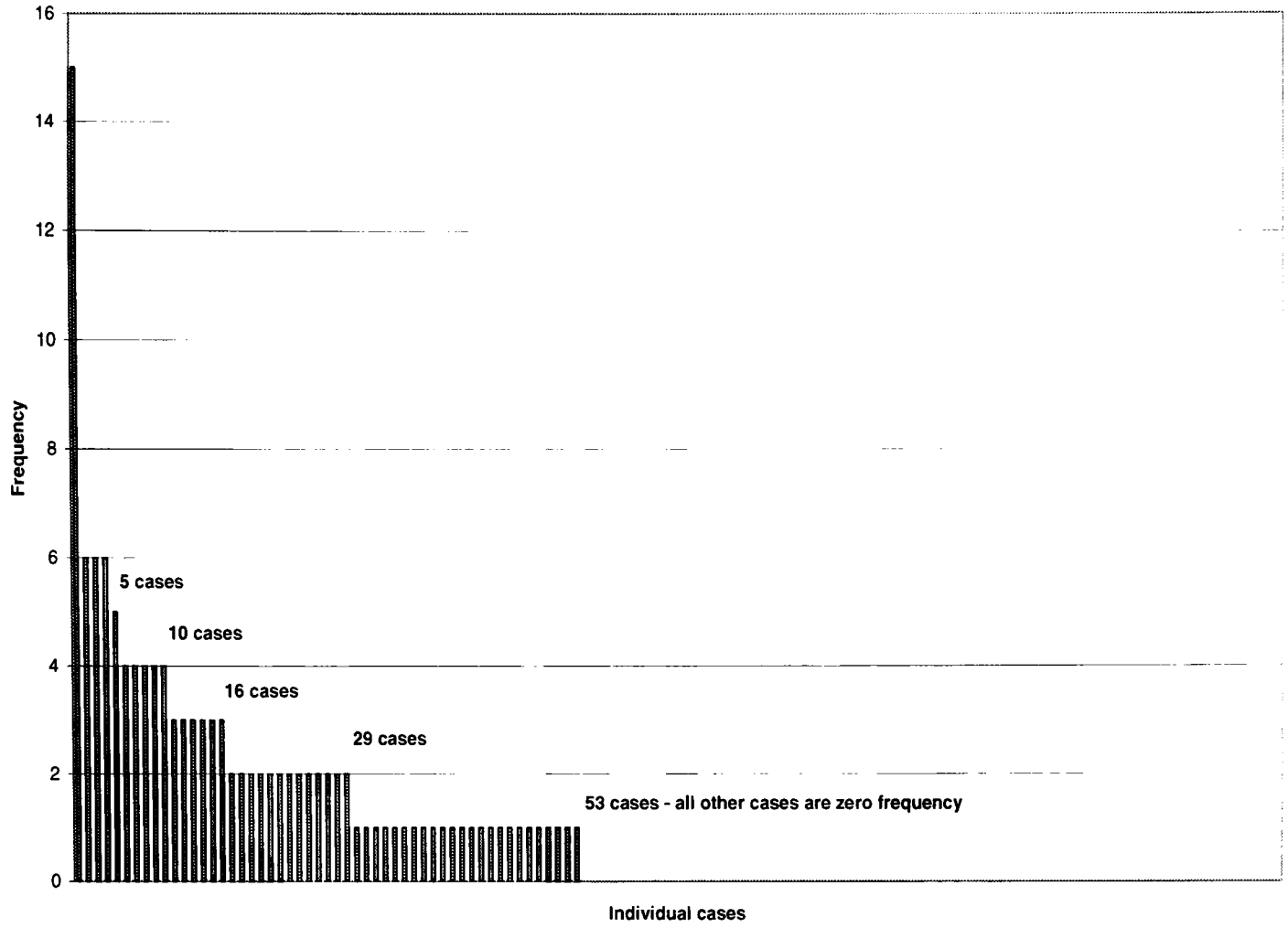


Figure 4.8: Frequency of optimal retrieval for each case.

distribution regularity using $\epsilon = 10$ (i.e. a neighbourhood of the cases which have an F_1 value within 10 of the optimal solution's value). *PDist* is our retrieval similarity metric (i.e. the sum of weighted differences between each event type in the case and the query) and we define *RDist* to be the difference between the F_1 rates using the retrieved case c and the optimal case $c^{`}$ for adaptation. The optimal case is obtained off-line, by adapting each test case to all other cases in the case base, evaluating each result against the known solution of each test, and selecting the highest (in terms of F_1 percentage). Assuming we could always choose to adapt the query to this optimal case, this would give a mean precision rate of 82% and recall of 70% with a duration error of 23% over the 126-case case base. This shows that either we need to improve adaptation or add more cases to the case base which could be more successfully adapted.

Case base size	Precision	Recall	Duration Error	Problem Solution Regularity	Problem Distri-bution Regularity
125	47%	52%	47%	0.02	0.98
53	55%	45%	44%	0.04	0.98
29	61%	42%	39%	0.21	0.93
16	66%	44%	40%	0.38	1.0
10	70%	45%	41%	0.60	1.0
5	71%	46%	41%	0.40	1.0

Table 4.4: Mean precision, recall and duration error for different case base sizes, along with their regularity measures

To find out whether the summary accuracy problem lies in the distribution of cases across the feature space, we measure the problem-distribution and problem-solution regularity of our case base. Wilson's original method calculates regularity as an average over a sequence of cases that are processed by the system, then added into the case base. This starts out with a very small case base, so the initial contributions to the problem-solution regularity are not really representative of our system. Instead, we modify Wilson's method to calculate regularity using leave-one-out testing over a fixed-size case base.

The results in table 4.4 show that some cases are much more useful than others, for example, if we don't use half of our set of cases, (the ones which are never the best case to retrieve for any of our queries) we actually have a gain in precision, with only a small decrease in recall. When the case base consists of only the five most useful cases, we only have a decrease of 5% in average recall, against a precision increase of 24%.

When the case base size[4] is 125, cases are well distributed across the problem space

[4]We have a total of 126 cases available, so when using the leave-one-out testing method, we have 125 cases in the case base at any one time. However, when reducing the size of the case base to 53 cases, we still test with all 126 cases as queries. If a particular test case is a member of the case base, it is removed, so the case base size is only 52 when processing that case. However, if the test case is not in the 53-case case base, the case base size remains at 53. This may be a small source of

(problem-distribution regularity = 0.98), but the problems that are retrieved using our similarity metric do not have solutions that allow successful test case adaptation (i.e. problem-solution regularity = 0.02: very low). This indicates that we need to improve the method of case retrieval, by making our similarity measure more representative of our solution evaluation criteria. As the case base size is reduced, we can trade off problem-distribution regularity for problem-solution regularity: by using fewer cases, problems are less well distributed across the problem space, but the solutions of the cases that are retrieved are more similar to the desired solutions of the test problems.

In conclusion, this section has highlighted problems in both the case retrieval and adaptation algorithms, while demonstrating that, for the full-sized case base, our problem-distribution across the feature space is sufficient.

4.6 CBR summary

This chapter began with an overview of CBR theory, then described our soccer ontology and reasoning system. In section 4.3 we found that it was relatively easy to summarise football games when their summaries were considered to be the shaded portions of the ticker-tape, because the program that assisted the web author used a limited number of rules to select which events would be shaded. For example, all goals and bookings were highlighted. This made it easier for our CBR system to learn. It was more difficult to generate good output highlights (in terms of mean precision and recall rates) when the solutions in the case base were derived from annotations of football highlights broadcast on television. This was because these case solutions represented more complex decisions made by many different sports producers when they edited the highlights packages together.

We found that case retrieval according to frequency of occurrence of each event

error, as the case base is not the same size across all tests.

type, weighted using the conditional probability of each event type being included in the highlights, given that it had occurred in the case, resulted in an increase of 6% precision and 5% recall over unweighted retrieval.

We then investigated the number of cases k to be retrieved and $k = 25$ was found to offer a good trade-off between precision and recall rates on the one hand, and complexity and duration error on the other. The concept of context groups of causally related events was introduced, and adaptation based on context groups rather than individual events improved results by 4% precision and 29% recall.

Section 4.5 analysed case base coverage, showing that over half the cases were redundant, and removing them had little effect on accuracy. We also found problems in our CBR system in the case retrieval step: problem-solution regularity for the full case base was very low, suggesting that our problem similarity measure did not correspond well to similar solutions. To a lesser extent, there were also weaknesses in the adaptation step: even when all cases are adapted, and the best results used, thus bypassing the retrieval stage, precision is only 82% and recall 70%. Since problem-distribution regularity was high, we were able to rule out lack of data as a problem in our system.

The main contributions of this chapter were twofold: firstly, a robust characterisation of case based reasoning for a summarisation application and secondly, the use of context groups to represent event causality in a higher level structuring of the sequence of events, which improves summarisation results. We have also shown that good summarisation results can be achieved (82% precision and 88% recall) when the acceptance criteria are limited to the correct event class. We have however not been able to address the issue raised in Chapter 3 of structuring the summary as a narrative. While we would argue that the context groups represent a local narrative episode, it is not obvious how to combine them with a macro-level narrative structure.

The next chapter looks at how we can improve on the results so far, by exploiting

additional information that the expert uses in their decision making, namely background knowledge about the players, clubs and so on. We also investigate the use of probabilistic techniques for summarisation as an alternative to case based reasoning.

Chapter 5

Probabilistic Approaches to Summarisation

The previous chapter explored the use of Case Based Reasoning for the generation of soccer highlights. Here, we investigate an alternative reasoning method to CBR: using a Bayesian approach, where probabilities derived from event frequency analysis are used to select events and context groups of events for inclusion in a summary. A probabilistic approach has the advantage over case based reasoning that it provides a natural framework to incorporate the probabilities of semantic event recognition from a video information source, if required. Research is also underway [Ding and Peng (2004)] to extend the OWL ontology language to allow probabilities to be marked up on instances in OWL files, which would allow, for example, the probability of occurrence of each soccer event class to be pre-encoded in its OWL file. It would then not be necessary to access the entire training set[1] of data each time a single soccer game was summarised.

We begin this chapter with a theoretical overview of Markov chains and Hidden Markov Models. As shown in figure 5.1, there are two probabilistic steps in our summarisation system. Firstly, we need to cluster events into context groups, and secondly, decide whether a context group should be included in the summary.

[1]Since we are no longer using case based reasoning, in this chapter our data will be referred to as a "training set" or "test set" as appropriate, rather than "case base".

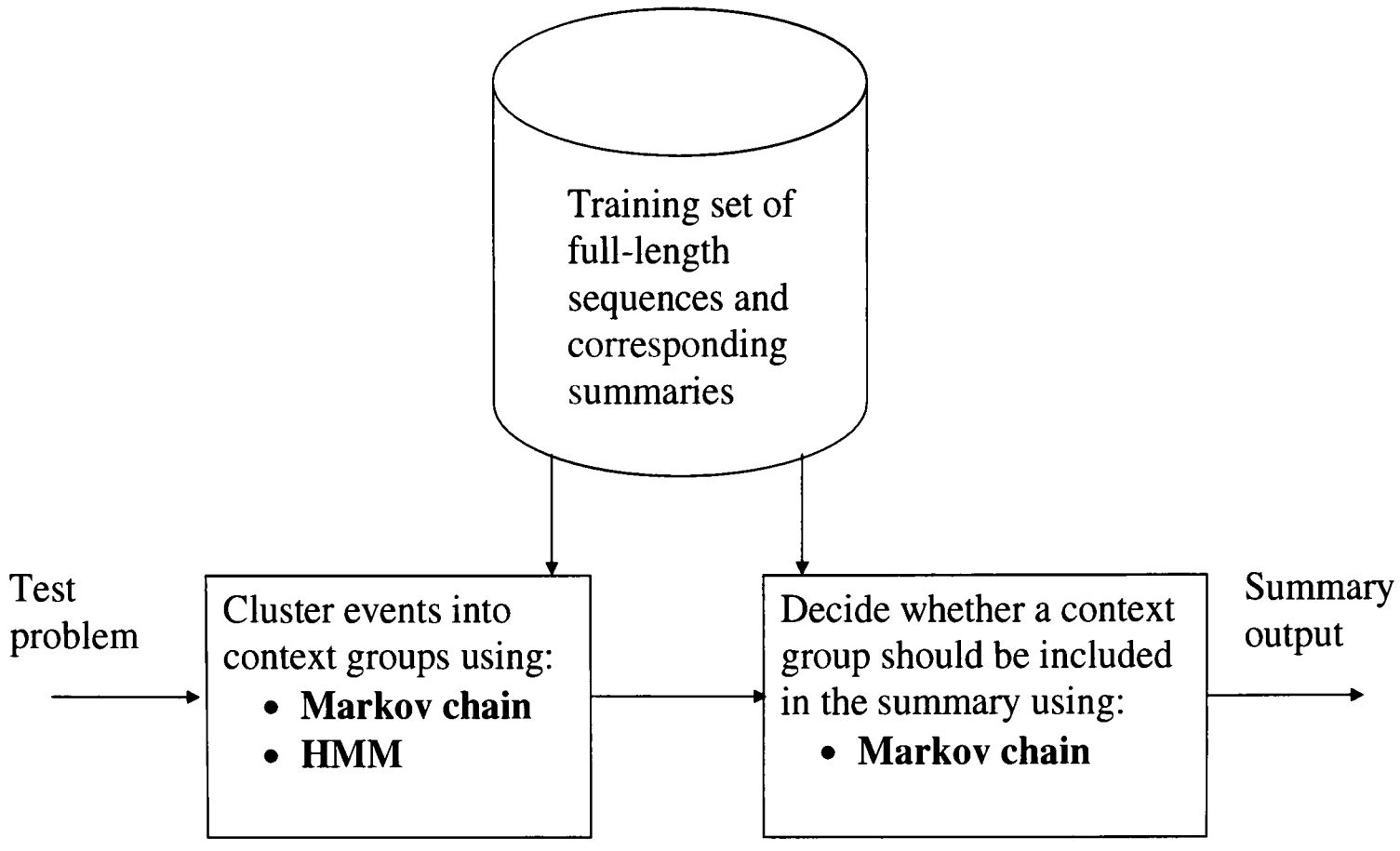


Figure 5.1: The two probabilistic stages in our summarisation system

For the first step, we begin by investigating the simplest approach using Markov chains, as they are well suited to capture the structure of temporal sequences. Markov chains are discussed further in section 5.1.1. We then look at the more powerful Hidden Markov Model, which is based on the assumption that some higher, hidden model exists that can explain the data observed at each time interval. Various forms of Hidden Markov Model: hierarchical, profile and semi-Markov, are explained in section 5.1.2. The second step of our summarisation algorithm, to assign a priority to each context group, is implemented using Markov chains only. The reasons for this, that is, the difficulties of modelling the priority allocation as an HMM, are discussed in section 5.1.3.

In section 5.2 we investigate the use of Markov chains and Hidden Markov Models for learning context groups from the event data in our training set. This is followed in section 5.3 by experiments to determine the relative importance of different context groups based on their probability of inclusion in the highlights. Section 5.4 discusses

the implications of the experimental results. Section 5.5 introduces our background knowledge ontology and investigates how additional event features can be included in the summarisation process, beyond event class alone, using this static knowledge base. In section 5.6 we look at event and context group-based summarisation using the K means algorithm. We finish the chapter with a summary in section 5.7.

5.1 Markov modelling

5.1.1 Markov chains

A Markov chain is a sequence of random variables $E_1, E_2, ...E_t$. At a certain time interval, the probability of the variable taking one of a finite number of values $q_i \in Q$, depends only on the value of the variable at the previous time interval (if a first order Markov process) or several previous intervals (if a higher order process). A Markov chain is specified by a state space, Q, an initial distribution $\pi = P(E_1)$ and a transition matrix A, where $a_{ij} = P(E_t = q_j | E_{t-1} = q_i)$. For example, in our soccer application, the set of states Q can represent the different event classes, *Assist*, *Block*, *Booking* and so on. The transition probability $P(E_t = q_j | E_{t-1} = q_i)$ is the conditional probability of the system entering a new state $E_t = q_j$, given the current state of the system $E_{t-1} = q_i$. As shown in figure 5.2, the transitions between states in a Markov chain can be represented as a directed graph, where edges represent transitions and vertices represent states. The Markov property allows the calculation of the joint probability of passing through any sequence of states in a system for which the first-order Markov assumption holds, using only the transition probabilities:

$$P(E_t, E_{t-1}, E_{t-2} \dots E_2, E_1) = P(E_t | E_{t-1}) \cdot P(E_{t-1} | E_{t-2}) \dots P(E_2 | E_1) \cdot P(E_1) \quad (5.1)$$

where $P(E_1)$ is the probability of being in an initial state and $P(E_t, E_{t-1})$ denotes the joint probability of E_{t-1} and E_t occurring in sequence. To estimate the transition matrix we can simply count of the number of co-occurring pairs of events and normalise appropriately.

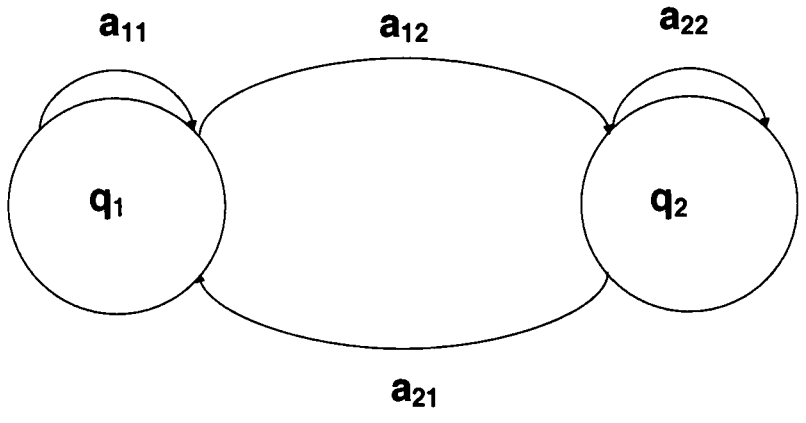


Figure 5.2: A two-state Markov chain transition diagram

5.1.2 Hidden Markov Models

Rabiner (1989) defines a Hidden Markov Model as "a doubly embedded stochastic process with an underlying stochastic process that is *not* observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observations." A Hidden Markov Model (HMM), differs from a Markov chain because each state no longer corresponds to an observable physical event, but instead, the observation is a probabilistic function of the state.

A Hidden Markov Model is defined by the parameters $\lambda = (A, B, \pi)$. As shown in figure 5.3, an HMM is made up of a finite number of states, $Q = q_1, q_2 ... q_N$. At each clock period t, the process moves into a new state q_j with a transitional probability distribution $a_{ij} = P(q_j \text{ at t} \mid q_i \text{ at t-1})$. This is the first-order Markov property, as the probability is dependent only on the previous state q_i. After each transition is made, an observation symbol O_t, which is a member of the set of possible symbols $V = v_1, v_2 ..., v_k, ... v_M$, is produced according to a probability distribution $b_j(k) = P(v_k \text{ at t} \mid q_j \text{ at t})$, dependent on the current state at time t. For example, in our soccer summary application, the observation symbol set could be the set of the 20 event classes (*Assist*, *Block*, *Booking* etc), and the hidden states could represent some unobservable editorial decisions of Inclusion or Exclusion from the summary.

When specifying a Hidden Markov Model, we have to choose N, the number of model states appropriate for our application, as well as to select state transition

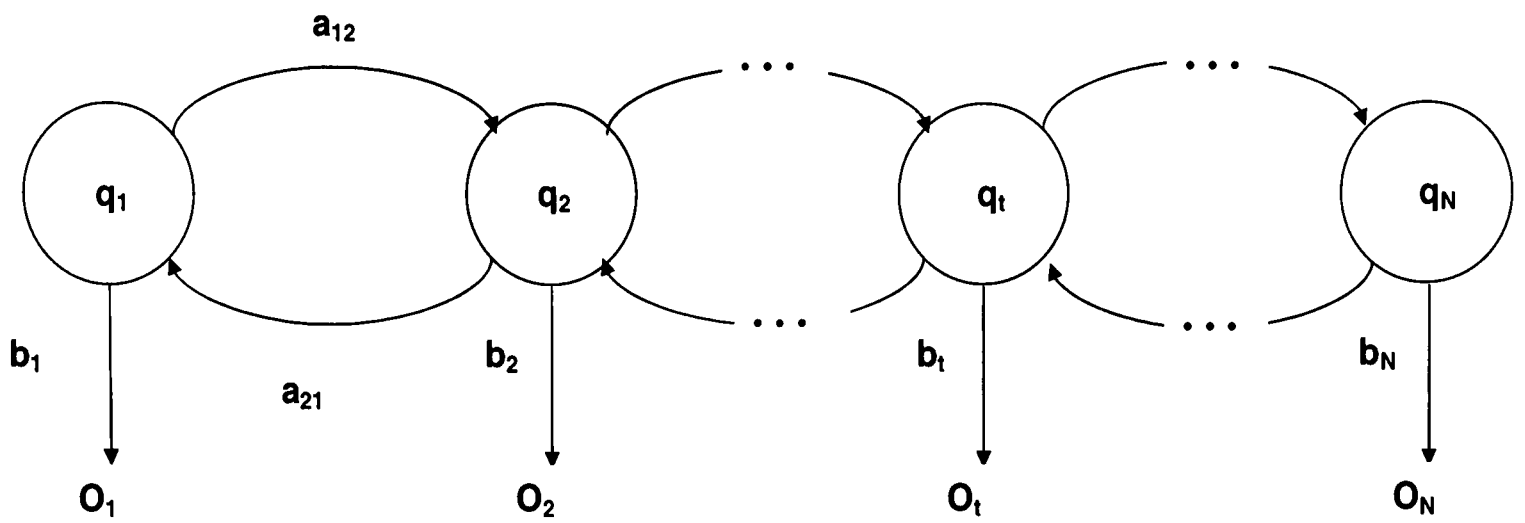


Figure 5.3: A Hidden Markov Model

probabilities, A, and observation probabilities, B, in each state, along with the initial state probability distribution, $\pi = P(q_i \text{ at t=1})$.

There are three problems associated with HMMs:

- **Evaluation (Likelihood Determination.)** Given the observation sequence of length T, $O = O_1, O_2, ...O_T$ and the model $\lambda = (A, B, \pi)$, compute $P(O|\lambda)$. the probability of the observation sequence, given the HMM, λ. This can be solved using the *forward-backward* algorithm [Rabiner and Juang (1986)], explained further in Appendix B.1.1.

- **Decoding (State Estimation.)** Given the observation sequence $O = O_1, O_2, ...O_T$, choose a state sequence $I = i_1, i_2, ...i_T$ that is optimal in some sense. As described in Rabiner and Juang (1986), this can be solved using the *Viterbi* algorithm, explained further in Appendix B.1.2.

- **Learning.** Given the observation sequence $O = O_1, O_2, ...O_T$, find the most likely HMM to have produced it. That is, maximise $P(O|\lambda)$ by selecting appropriate model parameters $\lambda = (A, B, \pi)$. When estimating the parameters in a Markov chain (i.e. the transition matrix a_{ij}) we need only count the co-occurrence of pairs of observation symbols. The parameter estimation process for a hidden Markov Model is substantially more complex, as we need to esti-

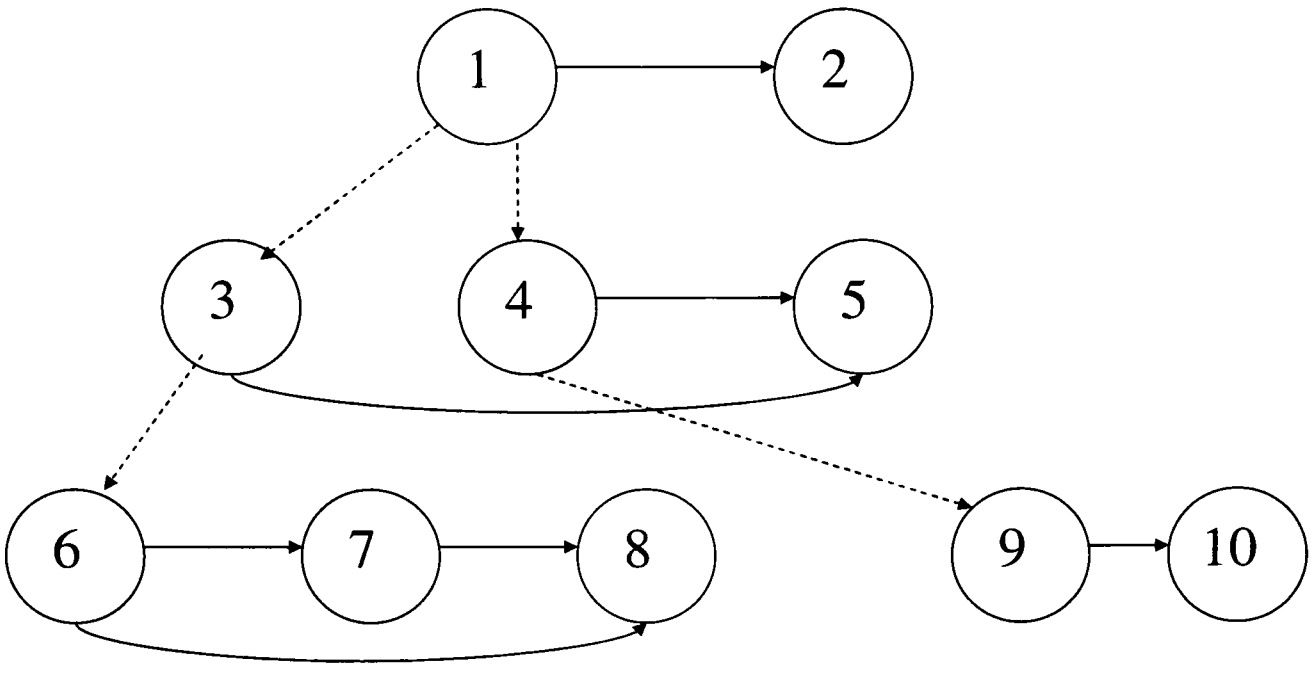
mate both the transition matrix A and the observation probability distributions B at the same time. This can be achieved using the *Baum-Welch reestimation formula*, see Appendix B.1.3 for details.

A mixture of Hidden Markov Models can be used to group (pre-segmented) sequences into K clusters [Smyth (1997)] for a known value of K, by identifying K different HMMs which are assumed to have generated the sequences from each cluster in the training set. The log probability of the test sequence, given each HMM, is used to identify which cluster the test sequence should fall under.

A Hidden Markov Model can be structured in several different ways, as discussed in Murphy (2002), such as hierarchical, semi-Markov or profile HMMs. Hierarchical Hidden Markov Models [Fine et al. (1998)] model domains with hierarchical structure at multiple length or time scales. A *production* state in an HMM emits single observations, while an *abstract* state emits a string of observations, and can itself be modelled as an HMM in the next layer of the hierarchy. An example of a hierarchical HMM is shown in figure 5.4.

Hidden semi-Markov Models are discussed in Murphy (2002). In these models, each state q_i emits a sequence of observations of length l_i, governed by a state duration distribution. The frequency of symbols in the emitted sequence is also governed by a state-specific distribution. This means that the process does not need to transition states into a new time period in order to observe a new symbol. An example is shown in figure 5.5.

Profile HMMs [Krogh et al. (1994)] originated in the field of bioinformatics and are used for sequence alignment. As shown in figure 5.6, for each state m_k where the sequence elements match, there is a delete state d_k that does not produce an observation, but is a dummy state used to skip m_k. There are also insert states either side of the match states, which generate observations in the same way as the match states, but will emit an extra symbol v with probability distribution $P(v|i_k)$.



Solid lines represent horizontal transitions between states in the same HMM. Dotted lines represent vertical transitions to sub-HMMs.

Figure 5.4: A Hierarchical HMM structure

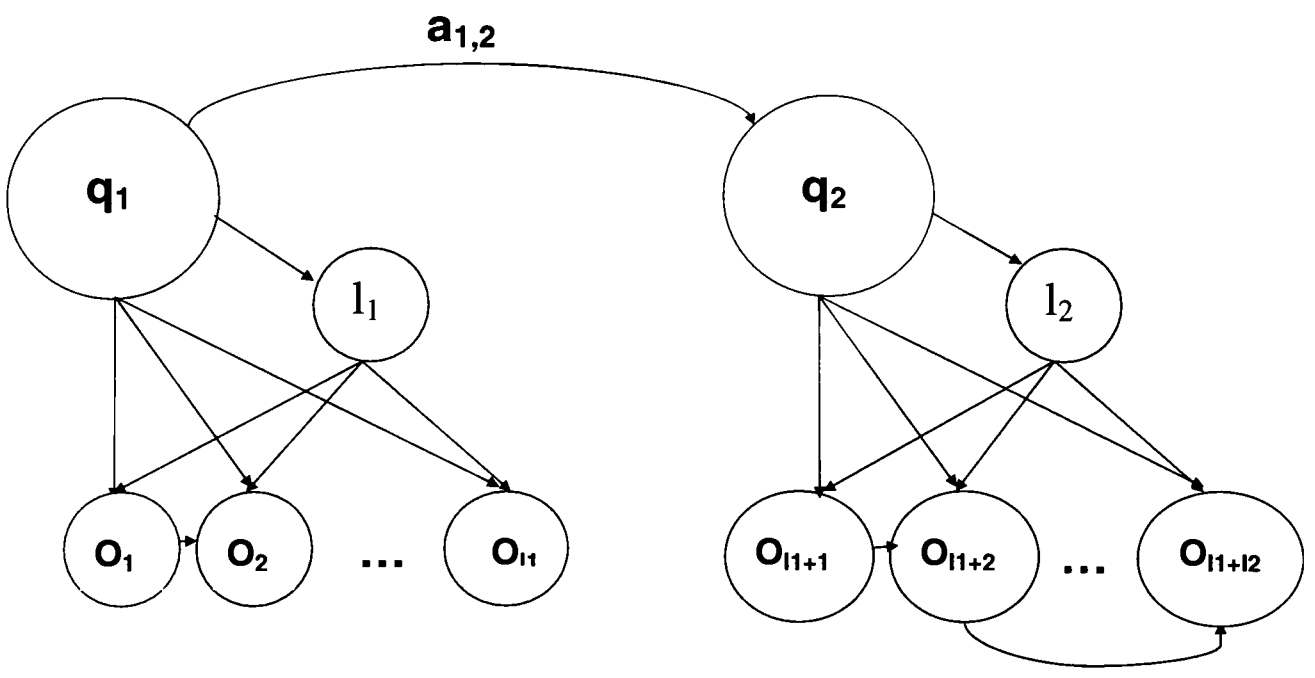


Figure 5.5: The Hidden semi-Markov Model structure from Murphy (2002), where each state q_i emits a sequence of observations $O_1 ... O_{li}$. The observation nodes within a segment need not be fully connected.

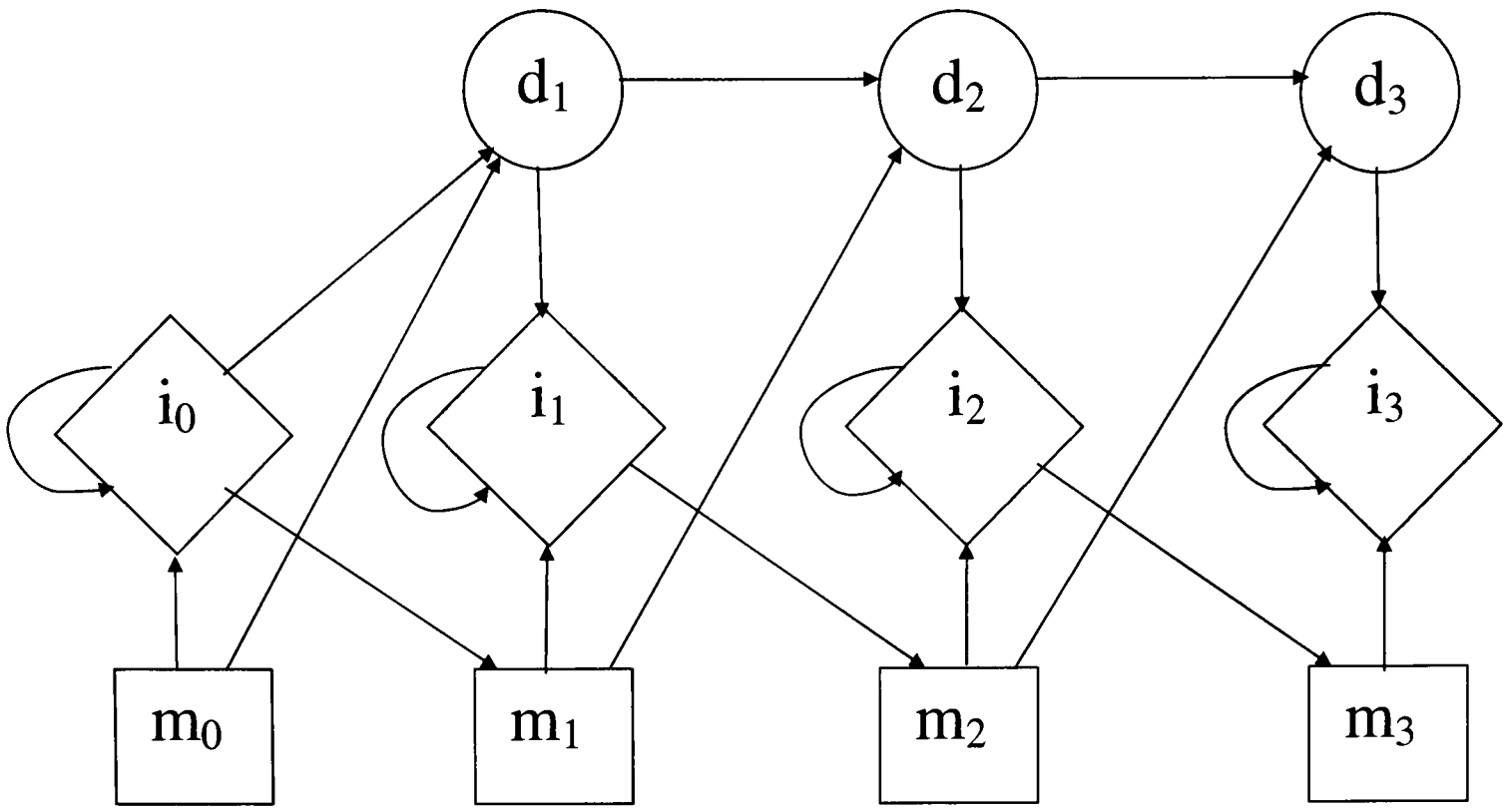


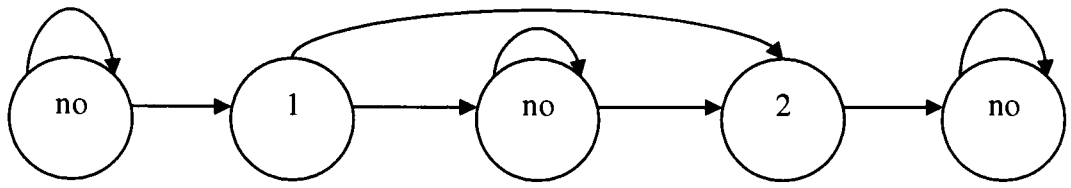
Figure 5.6: A Profile HMM structure, with delete states d_k, insert states i_k and match states m_k

The model also requires a dummy "BEGIN" and "END" state, which do not emit observations. That is, the sequences must again be pre-segmented.

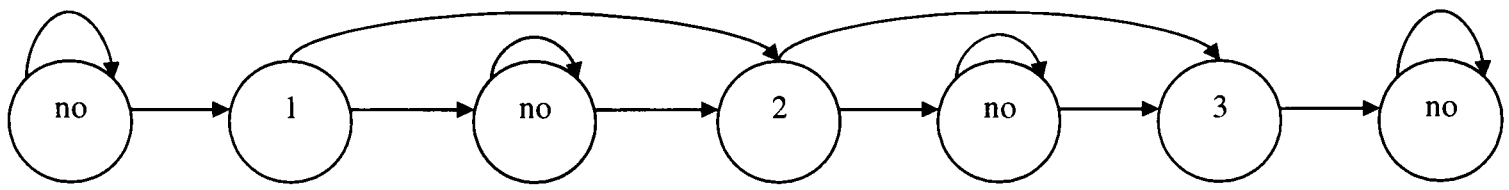
We next explain the rationale for the choice of Markov model structure in our summarisation system.

5.1.3 Using Markov chains and HMMs for summarisation

While Hidden Markov Models have frequently been used for video summarisation in the literature, as outlined in section 2.3, they have generally been employed in the extraction of semantic labels from the signal. The use of HMMs for semantic *summarisation* however is far less widespread. Conroy and O'Leary (2001) present the first example of the use of HMMs for text summarisation. The authors calculate the probability of each sentence being included in the summary, given a set of features for each observed sentence. Sentence features include position of the sentence in the document, number of words in the sentence, probability of those words occurring etc. That is, they do not represent any semantic meaning in the sentence. Conroy



Hidden Markov Model to summarise two sentences
'no' represents a state where the observation (i.e. sentence) is not included in the summary, and '1' and '2' represent the states whose observations are the first and second sentences to be included in the summary



Hidden Markov Model to summarise three sentences

Figure 5.7: Conroy and O'Leary (2001)'s Hidden Markov Model for text summarisation

and O'Leary's model has *2s+1* states, with *s* summary states and *s+1* non-summary states. The drawback of this is that the number of HMM states (and hence processing time) increases with the length of summary required; and a new model has to be developed each time the required summary length is increased (for each additional sentence in the summary, two extra states have to be added to the model). The structure of Conroy and O'Leary's HMM is shown in figure 5.7. The authors' best reported result is an F_1 score of 53.7%, mean average over 7 tests.[2]

As shown in figure 5.1, we have two probabilistic stages in our summarisation system: firstly to cluster events into context groups, and secondly to assign a priority to each context group based on their probability of being included in the summary. This essentially means that we are implementing a hierarchical model. We experiment with the use of Markov chains at both stages of the system, and attempt to use a Hidden Markov Model for context group clustering as an alternative to a Markov

[2]Recall from equation 1.3 that $F_1 = \frac{2*precision*recall}{precision+recall}$

chain. Markov chains are less complex than HMMs, which saves on processing power, but an HMM is a more powerful model that should be able to identify underlying structure in our data more easily.

Formulating our summarisation problem as an HMM does present significant difficulties however. We cannot take advantage of Smyth (1997)'s clustering method, as our sequences are not pre-segmented: in fact, the first stage of summarisation is to find the beginning and end points of the sequences of events that form the context groups. Consider the analogy of speech recognition, an application to which HMMs are frequently applied [Rabiner (1989)] in the identification of words from groups of phonemes. We cannot apply the same principle to identifying context groups within sequences of events because speech recognisers rely on the detection of a "silence" state to distinguish when the word has come to an end. In other words, the length of the group of phonemes is already known, before the model is applied. Our problem is the opposite: we want to apply the model to segment the events into groups. In section 5.2.2 we experiment with specifying such a "silence" or "end of group" state, using the ticker-tape paragraph groupings.

The second stage of summarisation, the decision of whether a context group is Included or Excluded from the summary, suggests a two-state HMM. However, from the definition of the first-order Markov property, a state must encapsulate all knowledge of past events, which a two-state model does not. Knowing that the current context group is Included or Excluded is not enough to tell us whether the next group to be observed will be Included or Excluded. The state actually needs to record information about *all* past context groups that have been included in the summary. It was for this reason that Conroy and O'Leary developed their HMMs with a transition to a new state each time a sentence was included in the summary, but as we have just discussed, this design requires an additional two states for each extra sentence in the summary.

The same difficulty would arise if we used a Hidden semi-Markov Model to combine the two summarisation stages together. As we have seen, Profile HMMs were developed for sequence alignment, which may be useful for personalisation feedback, so that the system can calculate how similar the current summary sequence is to sequences that the user has previously enjoyed. Although we do not implement the profile model in this thesis, the possibilities are interesting and discussed further in section 8.2 on future work.

5.2 Event clustering using Markov chains

Before applying Markov chains to our soccer summarisation problem, we first investigate whether they can be used in the discovery of causal relationships between events. In Chapter 4 we saw that representing event causality by clustering events together in context groups improved case-based adaptation results, but so far, we have relied on these context groups being manually created by the author of the web ticker-tape grouping events into paragraphs. The questions are then, whether these ticker-tape groupings are a good representation of the causality between events, and whether this causality can be learnt from the available data?

5.2.1 Unsupervised context group learning

In this section, we make the assumption that the ticker-tape paragraphs are not exact representations of the causality between events. Instead, we assume that events commonly occurring in sequence in the soccer match are causally related, and this will therefore be a better way of constructing context groups than relying on a web ticker-tape author.

We calculate the conditional probability of an event at some time t of class c ($E_t = c$) occurring in a case, given that the previous event was of class b ($E_{t-1} = b$).

as:

$$P(E_t = c|E_{t-1} = b) = \frac{P(E_t = c, E_{t-1} = b)}{P(E_{t-1} = b)} \tag{5.2}$$

$$P(E_t = c, E_{t-1} = b) = \frac{frequency(\text{event class pairs (b,c)})}{frequency(\text{all event pairs in the cases})} \tag{5.3}$$

$$P(E_{t-1} = b) = \frac{frequency(\text{event class b})}{frequency(\text{all events in the cases})} \tag{5.4}$$

The conditional probability matrix, shown in figure B.1 in appendix B. is calculated by counting the number of co-occurring pairs of events and the number of individual events of different classes. Our data set is quite small, and divided into separate games. such that it does not make sense to assume any causality between the last event in one soccer game and the first event in the next game. so these pairs are not counted. This means that the number of event pairs in the training set is not nearly equal to the total number of single events, as it would be if we had just one. infinite, sequence of events. Therefore we do not include the last event in each game in the frequency calculations for the marginal probability. and divide the numerator of equation 5.3 by the numerator of 5.4 (i.e. assume their denominators are equal.) Once we have made these adjustments. the columns of the conditional probability matrix sum to 1. as required. At this stage. for simplicity, we are only using one feature, the event class, to describe each event. Later, in section 5.5, we will add further features to the calculations. If a certain event pair combination does not occur in the training set, we do not set the corresponding element in the transition matrix to zero. as this could be due to the finite size of our training set, rather than the impossibility of that combination. Instead, a uniform prior probability, α is assigned to any element in the transition matrix a_{ij} that would otherwise be zero, and then the whole column is normalised to allow for this:

$$a_{ij} = \frac{\xi_{ij} + \alpha}{z} \tag{5.5}$$

where ξ is the original matrix element, the uniform prior $\alpha = 0.001$ and z is the normalisation factor (the sum of all the elements in the matrix column).

The drawback of assigning this uniform prior is that it always tries to fit the model to the sequence, even when it is inappropriate (for example, an *Assist* event is very unlikely to follow a *Goal* event, so the probability of this transition may actually be zero). Another design option at this point would have been to include prior knowledge in the system by assigning probability estimates of known relationships related to the rules of soccer. For example, a *Free kick* must follow a *Goal* when the game restarts.

However such rules are domain specific, so we stick to using the uniform prior, a technique equally applicable to any domain. This is a very simple form of *parameter tying*, a technique often used in speech processing to reduce the transition matrix size and simplify the model. Instead of applying a uniform prior only to zero entries in the transition matrix, parameter tying raises the threshold at which this takes place. If there are only a few instances of a particular event combination in the training set, the probability estimate it is assigned using the frequentist approach may be very inaccurate. For example, for a transition matrix element $a_{ij} = \xi(ij)$, if the value of the a_{ji} entry is below a certain threshold, it is set to $1 - \xi(ij)$, then normalised. However, we will keep the threshold at zero, and simply apply the uniform prior.

To cluster a game's events into context groups, we assume that the Markov property holds within a context group. That is, all knowledge about past events $E_{t-2}, ...E_2, E_1$ is captured in the previous state E_{t-1}, namely:

$$P(E_t|E_{t-1}, E_{t-2}, ...E_2, E_1) = P(E_t|E_{t-1}) \tag{5.6}$$

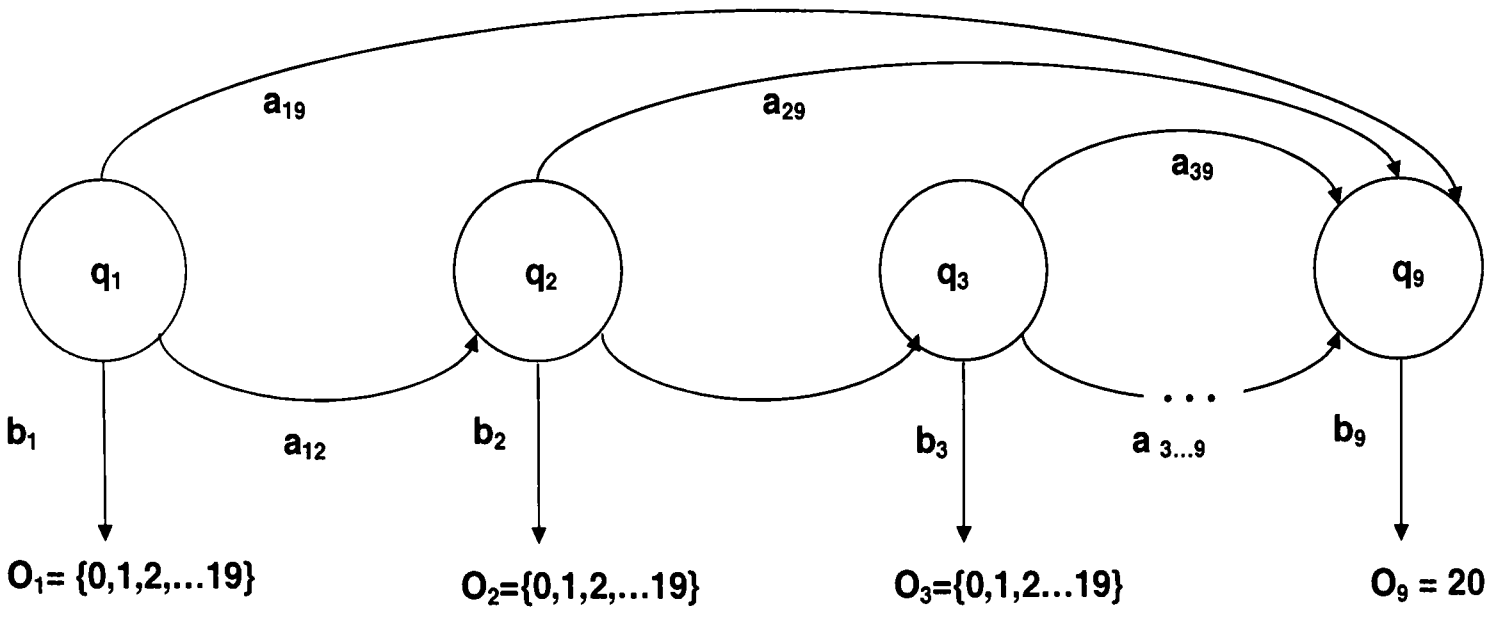
We construct a Markov chain beginning with the first event E_1, with subsequent events $E_2...E_t$, to calculate the joint probability of the group of events $E_t...E_2, E_1$ occurring in sequence:

$$P(E_t, E_{t-1}, ... E_1) = P(E_t|E_{t-1}) \cdot P(E_{t-1}|E_{t-2})...P(E_2|E_1) \cdot P(E_1) \tag{5.7}$$

When the joint probability of the context group falls below a certain threshold (we set the threshold to be 0.001) the context group is considered complete. Note that we do not normalise for long chains of events (by taking the t^{th} root) since we want to bias *against* very long sequences of events, which are less likely to be causally related. An example of context groups clustered using this method is shown in table B.1 in appendix B. Events that are clustered into intuitively reasonable groups in this example are: *Foul* $\rightarrow$ *Free kick* (joint probability 0.023) and *Shot* $\rightarrow$ *Save*(joint probability 0.005), while clusters that seem strange are *Throw in* $\rightarrow$ *Throw in* (due to its joint probability 0.077) and *Goal kick* $\rightarrow$ *Throw in* (joint probability 0.013). These latter two groupings are caused by the pre-filtering that the web editor has already carried out. Since nothing of interest happened between the two throw ins, or the goal kick and throw in, no other events are described between them, and so the two classes frequently appear in succession in the web ticker-tape. The only useful way of assessing the quality of the event causality analysis method, apart from subjective evaluation, is to use the automatically clustered context groups in the summarisation process. This we do in section 5.3. However, if we compare the context groups generated by the Markov chain method with the ticker-tape paragraphs grouped by the web editor, we find, over the 126 tests, that 24% are identical.

5.2.2 Supervised context group learning

An alternative to the unsupervised method, is to use the ticker-tape paragraphs as a "ground truth" from which to learn the context groups. At the end of each ticker context group, an additional "End of Group" symbol is inserted; with the twenty event classes, this increases the number of possible observations to 21. An HMM is then trained on the sequences of context groups, using a 9-state model. We need 9 states to model any sequence of events in a context group, since there can be up to eight events in a group (as found empirically in our training set) and we need a ninth, terminal state to indicate the end of the context group. Since the Markov



Our model has 9 states and observations in states 1 to 8 can take any one of 20 different values (the soccer event classes). The observation in the last state takes the dummy End Symbol (value 20). The process can only move forward to a new state, with no self-transitions, but can skip states if there are fewer than 8 event observations in the context group.

Figure 5.8: The Hidden Markov Model used in context group clustering

property implies that the current state only depends on the value of the variable in the previous state, in order to model a sequence of events, and maintain a record of what all of them were, we have to model each event in the context group as being emitted from a different state. A diagram of the model used is shown in figure 5.8.

The state transition and event emission probabilities of the HMM are learnt using the Baum-Welch algorithm, and the initial conditions, π, are set as the probability of occurrence of each event in state q1, and to zero for all other states, i.e. the sequence always begins in the first state. To cluster the events in the test problem, we insert an end symbol after the first event and calculate, using the Viterbi algorithm, the most likely state sequence that would have emitted that observation sequence. A second observation sequence is created using the first two events in the test, plus an end symbol. The probability of the sequence *First Event, End Symbol* fitting the model is compared with the probability of the second sequence *First Event, Second Event, End Symbol* fitting the model. This procedure is repeated by inserting an end symbol at different points in the test sequence, and choosing the point which gives the best

fit sequence (highest probability) to the HMM. Once the first context group has been established, the process is repeated to search for the end of the second context group, and so on.

As a mean average over the whole test set, 24% of events were grouped into the same context groups as the ticker-tape, using this HMM method. We might have expected a better result for the HMM than the Markov chain clustering, as it is a more powerful model. There are several possible reasons why this didn't happen. Either we do not have enough training data to estimate parameters for a 9 state model, or we should have assumed there were different types of context groups, generated by different HMMs, and tried to cluster accordingly. Another source of error may be the greedy search algorithm that assigns End Symbols one after the other. If one End Symbol is put in the wrong place, the error would propagate throughout the segmentation of the sequence into context groups.

5.3 Markov chain summarisation

We then want to find the probability of a context group being included, given that it has occurred:

$$P(CG(Included)|CG(Occurred)) = \frac{P(CG(Included), CG(Occurred)}{P(CG(Occurred))} \tag{5.8}$$

where a context group consists of a group of sequential events: $CG = \{E_1, E_2, ...E_{t-1}, E_t\}$.

Due to the nature of our problem, an event can only be included in the highlights if it has occurred in the case. However, perhaps counter-intuitively, this is *not* so with a pair or group. Even if the pair or group of events never actually took place one after the other, it may be included in the highlights as such, because of the editing that has taken place during summarisation. This means that while $P(E_1(Included), E_1(Occurred)) = P(E_1(Included))$ for a single event E_1, the joint probability: $P(pair(E_1(Included), E_2(Included)), pair(E_1(Occurred), E_2(Occurred))) \neq P(E_1(Included), E_2(Included))$.

The denominator of equation 5.8 is calculated as a Markov chain using probabilities of events occurring in the training set, as in equation 5.7. The numerator is also computed as a Markov chain:

$$P(CG(Included), CG(Occurred)) = P(E_{t,IO}|E_{t-1,IO}) \cdot P(E_{t-1,IO}|E_{t-2,IO}) \cdots \cdot P(E_{2,IO}|E_{1,IO}) \cdot P(E_{1,IO}) \quad (5.9)$$

where $E_{t,IO}$ denotes the event at time t being *Included* and having *Occurred*.

The conditional probability of an event E_2 being included and occurring, given that an event E_1 has been included and occurred immediately prior to E_2, is calculated as:

$$P(E_{2,IO}|E_{1,IO}) = \frac{P(E_{2,IO}, E_{1,IO})}{P(E_{1,I})} \quad (5.10)$$
$$= \frac{frequency(\text{event pair } (E_1, E_2) \text{ in both summaries and case problems})}{frequency(\text{event } E_1 \text{ in summaries})}$$

The conditional probability matrix for all the event classes is shown in figure B.1 in appendix B.

We also experiment with including event pair combinations that occur in the summaries, but not in the training problems, to see if representing the Kuleshov effect of event juxtaposition improves the results. This modifies equation 5.10 slightly:

$$P(E_{2,I}|E_{1,I}) = \frac{P(E_{2,I}, E_{1,I})}{P(E_{1,I})} \quad (5.11)$$
$$= \frac{frequency(\text{event pair } (E_1, E_2) \text{ in the summaries})}{frequency(\text{event } E_1 \text{ in summaries})}$$

These calculations make the assumption that the first order Markov property holds, not only between events in the problem description, but also over events that are included in the summary. Using this assumption, we can overcome the problem of not having enough data to calculate probabilities of combinations of events that occur rarely or not at all (the so-called "one-shot learning" problem). For example, if we wanted to calculate the probabilities of all 5-event context groups, we would need a

Precision	Recall	Duration Error
72%	28%	6%.
Correct classes precision	**Correct classes recall**	F_1
92%	36%	39%

Table 5.1: Mean summarisation results for single events selected according to their probability of inclusion in the highlights

minimum of $20^5 = 3200000$ events to get one occurrence of each possible combination of 5 events. Since our cases have on average 200 events, we only have about 25200 events available: far short of what we would need to apply a frequentist approach to calculation of the joint probability.

For each context group in the test problem, the probability of it occurring in the highlights, given that it has occurred in the original, is calculated. The groups with the highest probability are included in the highlights output, until the required time duration is reached. The results are shown in tables 5.1 to 5.5 and include the F_1 measure to allow us to make a direct comparison with Conroy and O'Leary (2001)'s text summarisation results using HMMs. Their highest reported result, taken as a mean average over the 7 tests they carried out, was $F_1 = 53.7\%$: table 5.5 demonstrates that our method has an F_1 value 7.3% higher. Admittedly, Conroy and O'Leary were summarising text sentences, rather than our ticker-tape representations of video events, however, this is the only result reported in the literature with which we can make a direct comparison, and our results are encouraging. Furthermore, the number of states in our Markov chain, unlike Conroy and O'Leary's model, does not change with summary length.

Precision	Recall	Duration Error
58%	62%	7%.
Correct classes precision	**Correct classes recall**	F_1
92%	93%	59%

Table 5.2: Mean summarisation results with context groups, taken from ticker paragraphs, selected according to their probability of inclusion in the highlights

Precision	Recall	Duration Error
62%	35%	8%.
Correct classes precision	**Correct classes recall**	F_1
92%	56%	43%

Table 5.3: Mean summarisation results with context groups generated from unsupervised clustering using Markov chains, selected according to their probability of inclusion in the highlights

Precision	Recall	Duration Error
44%	38 %	11%.
Correct classes precision	**Correct classes recall**	F_1
84%	72%	40%

Table 5.4: Mean summarisation results with context groups generated from supervised clustering using an HMM, with an End symbol, selected according to their probability of inclusion in the highlights

Precision	Recall	Duration Error
59%	65%	7%.
Correct classes precision	**Correct classes recall**	F_1
91%	94%	61%

Table 5.5: Mean summarisation results with ticker context groups, using probability of inclusion only (based on equation 5.11)

5.4 Discussion of event clustering and summarisation results

Using unsupervised or supervised event grouping with Markov chain or Hidden Markov modelling respectively, we can learn the ticker paragraphs with a 24% accuracy in both cases. For the unsupervised event grouping, a threshold of 0.05 or greater results in no groupings (i.e. context groups consist of only one event, which matches 48% of the ticker groups), and therefore the results are equivalent to the single-event based summarisation in table 5.1. Comparing tables 5.1 and 5.3, it seems that although the context groups learnt using the Markov chain method are not the same as those in the ticker-tapes, it is better to have some event groupings than to rely on summarising using single events. Comparing tables 5.2, 5.3 and 5.4, we can see that using the original ticker-tape paragraphs gives the best results, so we can conclude that the paragraphs are a good representation of event causality. Clustering events into context groups gives slightly better results when using the Markov chain method than the Hidden Markov Model (3% higher F_1 score): this may be because we need more than one HMM to model different types of context group. Another possible explanation of the better performance of Markov chains over HMMs could be that we do not have enough data to accurately estimate the probabilities of the larger number of variables in the HMM.

Table 5.5, which presents results using conditional probabilities of all event pairs that occur in the summaries, including combinations that don't occur together in the full length soccer games, shows a small improvement over the results in table 5.2. This is an interesting result, because it tells us that representing editing effects contributes to summarisation quality. We also retain reliability in our summary, that is, we do not introduce erroneous causal effects by juxtaposing unrelated events, as we learn causality from experts' summaries, which we can assume do not contain such errors. It's a small improvement, as the number of edit points in the summary

is small relative to the total number of pairs in the summary. However, the more edit points a summary has, which is often related to its length, the more significant this finding becomes.

The Markov chain method proposed so far only uses the class information of each event, without considering any additional event properties (such as start time, duration, player involved etc.) This is why the "correct classes" precision and recall measures are so high. It is difficult to extend this method to use all the available metadata for each event, due to the one-shot learning problem: each event that occurs is unique in its combination of attributes. The next section looks at how to overcome this problem.

5.5 Using background knowledge and event metadata

Background knowledge, defined by Zelikovitz and Hirsch (2002) as "readily available information...text, databases or other sources of knowledge related to a text classification problem", has been used to improve performance in text classification. Even if a test and training example do not have a high similarity match, if they have background knowledge items that are similar, their similarity score will be increased. In human terms, we also consider background knowledge to be information that an expert uses to make the decisions about what to summarise, as oppose to the actual information that is summarised. While knowledge has long been separated from the algorithmic process in A.I., the distinction we are making here is not between knowledge and algorithm, but between different types of knowledge: that which is time dependent, the soccer events, to be summarised; and the background knowledge, which is independent of time and only indirectly involved in the summarisation decisions. While previous work [Zelikovitz and Hirsch (2002)] has used background knowledge in text summarisation, our contribution is to make a time dependent split between back-

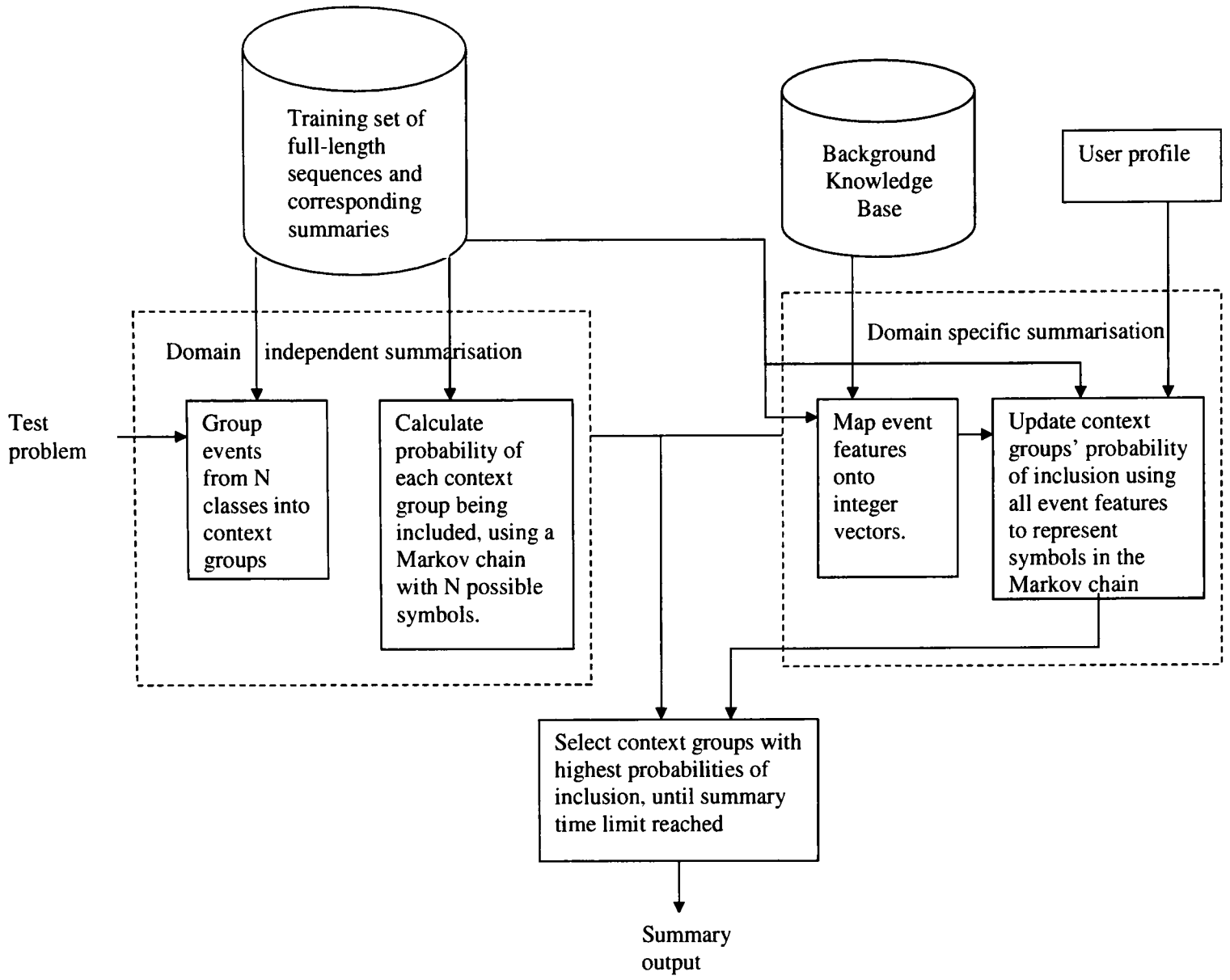


Figure 5.9: A Block diagram of our Markov chain summarisation system

ground and current knowledge. This section outlines the second, "domain specific summarisation" stage of the system in figure 5.9.

We define a background knowledge ontology for the soccer domain shown in figure 5.10 in OWL [McGuinness and van Harmelen (2004)] using Protégé 2000 [Protégé 2000]. Instances of the ontology are then read in from various sports web pages, which give information on players, teams and league tables. using a template mining approach as before.

Using this background knowledge, we turn each event instance into a vector representation. This requires a 16 dimensional vector of integer elements, as shown in table B.2 in appendix B. Strings are mapped on to integer codes. and times are grouped

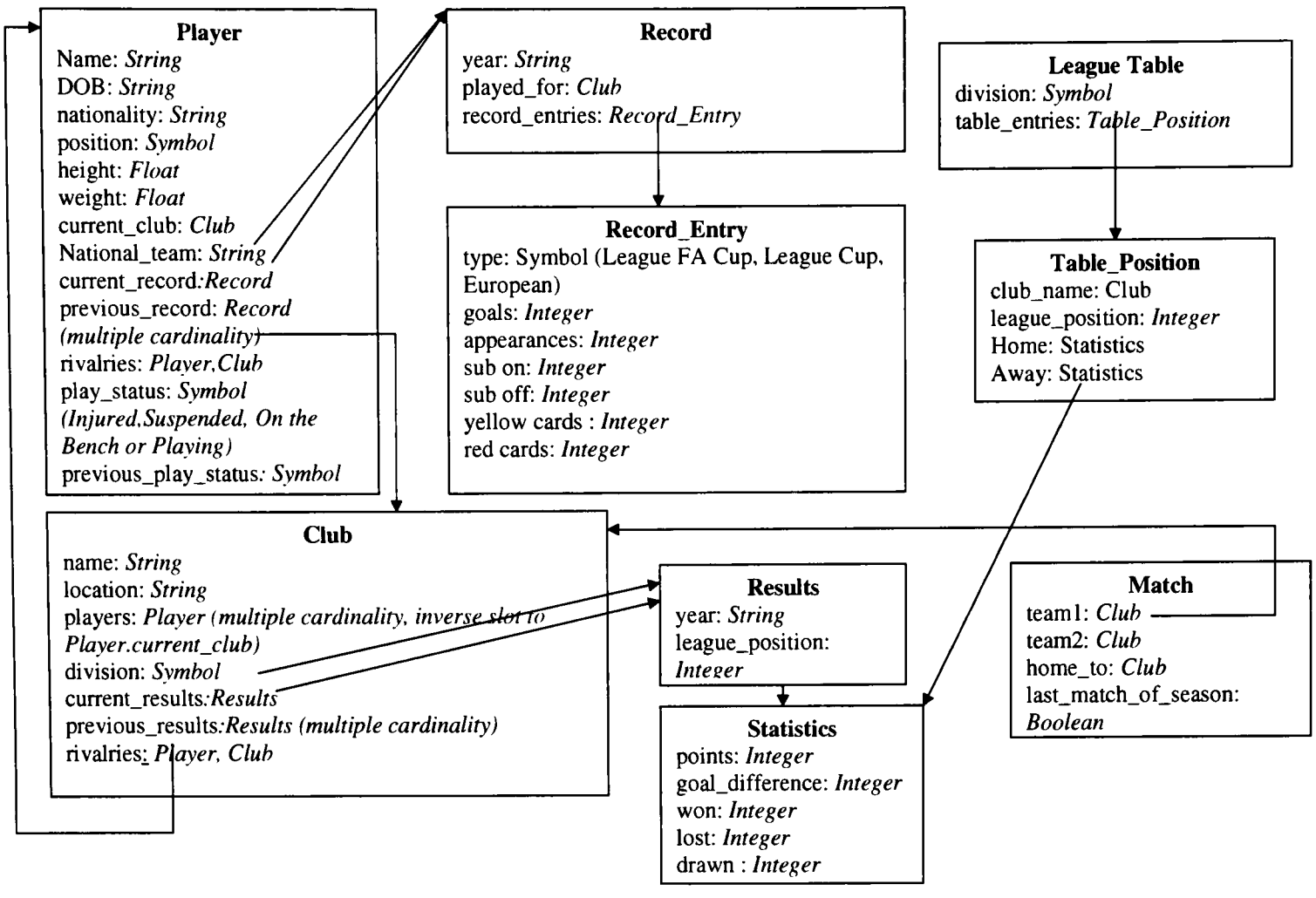


Figure 5.10: The soccer background knowledge ontology

into categories. The teams and players are given importance weightings, using the background knowledge. The background knowledge is also used in Chapter 6, for personalising summaries by learning from different types of soccer games.

We then create a conditional probability matrix $P(E_t = b|E_{t-1} = a)$ where a and b are 16 dimensional event vectors. Since the event space is of such a high dimension, the conditional probability matrix is sparsely populated and combinations of event values that do not occur in the training set are assigned a small uniform prior as before. Table 5.6 shows the precision and recall values when all features are taken into account, and table 5.7 lists the F_1 scores (for ease of comparison) when each individual feature is used to generate the summary. Context groups are generated from the ticker-tape, so variations in results due to the summarisation stage alone can be seen.

Precision	Recall	Duration Error
9%	8%	3%.
Correct classes precision	**Correct classes recall**	F_1
64%	53%	8%

Table 5.6: Mean summarisation results with ticker-tape context groups and all event features used in the summarisation process.

Feature Used	F_1
Event Class	63%
by	16%
duration	5%
start time	5%
extra time	8%
from	8%
taken	8%
to	8%
resulting in	3%
type	8%
on	3%
booked for	8%
dismissed for	8%
off	12%
team	7%
reason	8%

Table 5.7: Mean F_1 summarisation results with ticker-tape context groups and each individual event feature used in the summarisation process.

Table 5.6 shows that using all event features, equally weighted, gives very poor results, and table 5.7 indicates that some features (e.g. *event class*, and to a lesser extent *by*) are more useful than others when used on their own to determine summarisation priority of events. The final experiment of this section therefore looks at weighting the event features according to their relevance, using the F_1 scores of table 5.7 as a measure of relevance. The results, in table 5.8, show a substantial improvement on unweighted features, but are less successful than using the event class on its own. This may be because our training set is too small to populate such a

Precision	Recall	Duration Error
60%	61%	0%.
Correct classes precision	**Correct classes recall**	F_1
84%	88%	59%

Table 5.8: Mean summarisation results with ticker-tape context groups and weighted event features used in the summarisation process.

high dimensionality feature space with instances of every feature value, and hence the probability estimates are inaccurate.

5.6 K means summarisation

In this section we take a different approach to summarisation, for comparison against the Markov chain approach, using the well known K means algorithm. The elements of each event vector in the case base (excluding the current test) are normalised in the range 0 to 1, and assigned at random to one of K classes (in our case, K = 2). The centroids of the classes are calculated. and the events are reassigned to their new closest centroid. The procedure iterates until the centroids no longer move.

Now, we designate the centroid of the group containing the highest number of Included events to be the Included centroid, and the other centroid is designated the Excluded centroid. Each event in the test problem is then assigned to the Included or Excluded class. depending on which centroid is closer, and assigned a priority of inclusion dependent on the difference between its distances from the Included and Excluded centroids. Events are then added into the summary, smallest distance first. until the summary reaches the required time duration.

The experiment is repeated using vectors representing whole context groups, by concatenating the events together, rather than single events. Shorter context groups are padded with -1s, so that all vectors are the same length. The results are shown in tables 5.9, 5.10 and 5.11.

Precision	Recall	Duration Error
6.6%	8.0%	2.7%.
Correct classes precision	**Correct classes recall**	F_1
22.7%	26.8%	7.0%

Table 5.9: Mean summarisation results using the K means algorithm on an event-by-event basis.

Precision	Recall	Duration Error
38.4%	44.9%	4.9%.
Correct classes precision	**Correct classes recall**	F_1
57.6%	67.9%	40.6%

Table 5.10: Mean summarisation results with ticker-tape context groups using the K means algorithm on a context-group basis.

Precision	Recall	Duration Error
36.3%	40.7%	1.97%.
Correct classes precision	**Correct classes recall**	F_1
40.9%	45.6%	37.8%

Table 5.11: Mean summarisation results with Markov chain learnt context groups using the K means algorithm on a context-group basis.

As expected, the results for context-group based summarisation are better than for single event-based summarisation. Although, except for the event class, all feature elements are normalised in the range 0 to 1, it doesn't always make sense to reduce the features down to numbers as the distance isn't always meaningful. For example, for the *from* feature the distance between "right by-line" (coded as 1) and "right wing" (coded as 3) should not necessarily be less than the distance between "right by-line" and "right channel" (coded 5). However, without imposing these rather heuristic codings on the semantics, it is difficult to manipulate the features.

Compared to a standard K means algorithm, when using all features our Markov chain approach to summarisation gives an F_1 score 18.4% higher. However, neither of the two methods' results are as good as the results obtained when only the *event class* features is used. This is because the 16 dimensional feature vector representing each event means that we have created a very high dimensionality feature space, which

our training set can only sparsely populate. The K means method does not cope well with the semantic richness offered by our ontology, as the similarity between the "right wing" and "right by line" or the "right wing" and "left wing" in the *from* feature for example, cannot easily be distilled into a simple Euclidean distance, without specifying complex similarity measures for each particular feature.

Although the Markov chain mechanism sets a limit on the complexity of the semantics or the minimum training set size, it has the advantage that, when only the *event class* features is used, the same algorithm can be applied to any domain, using only the object class concept specified in any ontology.

5.7 Summary of probabilistic approaches to summarisation

The main contribution of this chapter is the development of a probabilistic alternative to CBR summarisation using Markov chains, which, at 59% precision and 65% recall, is a 13% improvement on the CBR method for both measures. This may be because, while the CBR method takes a global approach, comparing one whole football game to another, the Markov chain method is better able to model the local causality between events in the summary. As with CBR, the context-group based results from the Markov chain summaries are better than the single-event based summaries. This demonstrates that the episodic nature of the context group can, to some extent, introduce a more narrative style to the summary. According to the expert opinion elicited in Chapter 3, this delivers a better quality summary.

In section 5.2, causally related events were clustered into context groups using two alternative models: the Markov chain and HMM methods. Both only found the same groupings as the ticker-tape paragraphs for 24% of the groups, on average, but there were marginally better summarisation results using the Markov chain method, perhaps because it required fewer parameters than the HMM method.

In section 5.3 we developed a new method for summarising events using a Markov chain that represents causality across editing points without introducing erroneous causal effects. The average F_1 results over the test set were 7.3% higher than those reported in the literature [Conroy and O'Leary (2001)], as well as using a more flexible model that could be used to generate a summary of any length. Since only the *event class* feature was used to describe each event in the Markov chain, another advantage of our approach is that it is easily applicable for event descriptions in other domains.

The results of section 5.5 showed that the Markov chain mechanism sets a limit on the complexity of the semantics, so we were unable to exploit our background knowledge base as we would have liked. Although we were unable to take advantage of the influence of the static background knowledge base on the event features in this chapter, due to the "curse of dimensionality", in the next chapter we investigate the use of our background knowledge to inform the personalisation of a summary.

Chapter 6

Personalisation

In the previous chapters we have looked at alternative methods of automatically generating a neutral summary using case based reasoning and Markov models; now we investigate the personalisation of a soccer highlights package according to different users' preferences. The purpose of personalisation is to "give the customer a high quality product they really need and can use, at the ... lowest price" [Riecken (2000)]. Although the possibilities for personalisation are endless, taking a user-centred design approach, Kramer et al. (2000) make the point that it is only worth personalising features and tools that are of value to the end user. Our efforts are motivated by the results of Evans (2003), discussed in Chapter 3, which show that personalisation is sought after by football fans and can be a strong selling point for any content management system. Under the heading of 'personalised' summaries, our work considers both *biased* and *user-focused* summaries. As explained in section 2.1, biased summaries may concentrate on a particular narrative or plotline, rather than following the usual journalistic standard of objectivity, while user-focused summaries contain material of interest to a particular user, which may be specified via a search query or a user profile. Obviously, the two are related, in that a user may request summaries containing a certain narrative or episodic grouping of events.

We begin this chapter with an overview of previous work on user-focused and biased summarisation, along with user profile elicitation, summary coherence mea-

sures and personalisation of multimedia content. In section 6.2 we investigate a novel approach to the biasing of a summary, by using various subsets of our training data to produce summaries leaning towards a particular subplot. Section 6.3 describes the user profile ontology we have designed to capture soccer fans' summary preferences, along with our experimental method to generate personalised summaries according to different users' profiles, comparing a traditional weighting method to our biased training set approach. In section 6.4 we evaluate the accuracy of our summariser in producing soccer highlights of the preferred length, and section 6.5 reports on results to generate personalised summaries for two different user profiles. In section 6.6 we introduce a measure of utility to estimate how well the summary content meets user requirements, and plot a graph of utility against summary duration. The trade-off between coherence and personalisation is investigated in section 6.7 and we conclude with a summary of the chapter in section 6.8.

6.1 Summary personalisation literature

The literature we review in this section covers methods for text-based user-focused summary generation from the natural language processing community, along with techniques for explicit and implicit user profile knowledge elicitation. We address the need to balance summary coherence against personalisation requirements, and discuss some methods from the literature for measuring coherence. We then look at research on personalising multimedia summaries and consider what we can learn from this for our own system.

6.1.1 User focused summary generation

The approach to user-focused or personalised summaries from the natural language processing community has mainly concentrated on extending generic text summarisation techniques to include user needs as additional features in the selection process.

Mani and Bloedorn (1998) include features derived from user needs in their salience function to determine what information in the source text should be included in the user-focused summary. Their training set consists of a number of positive and negative examples of sentences, i.e. ones included in a summary, and ones that are not sufficiently interesting to be included. The authors resolve the difficulty of acquiring enough personalised summaries to act as training data by generating a summary automatically from a specification of a user's information needs. Their method involves a user picking a sample of ten documents from a corpus to match their interests. The top content words in terms of the G^2 score[1] are then extracted from each document. The mean G^2 score over these ten documents is calculated and all words more than 2.5 standard deviations greater than this mean are considered to represent the user's interests. For each training document in the corpus, each sentence is weighted based on the number of user-interest words it contains. The top c% (where c is the compression ratio) of sentences in each training document are then designated as its summary. In other words, this is an artificial way of generating a large user-focused training set, without requiring much manual intervention. However, since it is word-based, the selected sentences may not represent the user's semantic interests well, and the system could produce artificially high results, because the features used to produce the ground truth summary are already known. New sentences that are presented to the system are converted to vector representation using location, proper names, a feature based on the tf.idf metric[2], and two cohesion features using synonymy and co-occurrence, along with user-focused features which are based on the number of user-interest words in the sentence. The authors then experiment with various standard machine learning algorithms to discriminate between summary and

[1]The G^2 statistic is the log-likelihood ratio, indicating the probability that the frequency of a term in a particular document is greater than its frequency of occurrence in the corpus.

[2]The tf.idf metric [Maudlin (1991)] is the product of the term frequency in the sentence (tf) and the inverse document frequency (idf) where $idf = \frac{fc}{fs}$. fc is frequency of appearance in a standard corpus, and fs is the frequency of appearance in a scene or sentence. When the tf.idf metric is high, this implies that the term is important.

non-summary sentences.

The drawback of this method is that it requires a substantial amount of user intervention to specify preferences (i.e. reading through the whole corpus to decide their preferred documents), and users are assumed to have only one favourite topic. In the soccer highlights application, we think it better to allow the user to specify their summary time limit, and preferred semantic content, for example favourite player, or favourite team, rather than requiring them to watch many highlights packages in order to indicate their favourites. We now look in more detail at methods in the literature for user profile information elicitation.

6.1.2 User profile elicitation

Agnihotri et al. (2003) report on a survey of experts to establish user requirements that might be expected for personalised multimedia summaries. Although not eliciting information directly from users themselves, this review raises a number of interesting points. It is thought that summaries would primarily be watched in "spare" time, while commuting or between scheduled events, for example in a waiting room, on various devices including TVs, PCs, PDAs and phones. Therefore the summary content is not only dependent on users' interests, but also on their available time and task. For example, if the user is engaged in another primary activity, summary content must be less detailed than if the user is able to devote their complete attention to it. The authors believe that the personal profile should consist of an *implicit* and an *explicit* section. The implicit section is derived from observed user behaviour, such as previous search criteria, summary access patterns and past usage of the summaries, programmes watched and their genre, list of topics consumed and users' viewing habits. On the other hand, information for the explicit user profile is provided directly by users and should include gender, age, home and work location, profession, time allocated for watching television, users' schedule, hobbies, preferred content medium (e.g. audio versus video), ratings of topics, preferred summary based

on genre or task, favourite and disliked people (e.g. celebrities or politicians), favourite sports teams and music, and parental guidance rating. This extensive profile places a high burden on the user to input a large amount of information. The authors also make no mention of how they expect the summary to change based on the user profile.

The PTV system [Smyth and Cotter (2000), O'Sullivan et al. (2004)] learns an implicit user profile using collaborative filtering and case based reasoning, avoiding the need for a user to input their own user profile data. It contains two types of information: domain preferences and programme preferences. The former include preferred viewing times, list of available TV channels, subject keywords and genre preferences. Programme preferences are represented by a list of liked and disliked programmes. Implicit user profiles such as this are particularly suited to domains in which a user may not know what they want, or what is available, for example when new television programmes are recommended to viewers. In the soccer highlights domain however, as we have seen in Chapter 3, there is a well recognised expectation of what will occur in a soccer game, so we assume that a user can easily specify their own preferences. This assumption is backed up by Evans (2003)'s study which was able to elicit user preferences for a soccer highlights application. Furthermore, it is beyond our means to derive implicit profiles from any large scale user testing, so we opt for the use of an explicit rather than implicit user profile. As discussed further in section 6.3, we limit the user profile to many fewer categories than suggested by Agnihotri et al. (2003). This makes it easier for us to test how each property in the user profile individually influences the personalisation of the summary.

Another well known method for eliciting and refining information in a user profile is to employ user feedback, such as in the Broadcast News Navigator [Maybury et al. (2004)]. This is a system for browsing multimedia news items, using a combination of text, graphic, image, video and audio media, according to users' preferences of delivery and content. Feedback from the user about the relevance of individual news

items allows the user profile to be updated and performance improved. The biased training set method which we discuss in sections 6.2 and 6.5 could allow such user feedback, by incorporating summaries the user is pleased with into the biased training set. Our personalisation is not at the programme level like PTV, or even at the topic level, like the Broadcast News Navigator, but is at a finer grain: at the event level. This means that we encounter a problem that does not arise in PTV or the Broadcast News Navigator: how to manage the trade-off between summary coherence and personalisation.

6.1.3 Coherence versus personalisation

Young (2000) notes that coherence comes from the selection of actions whose causal and temporal relationships highlight an underlying plot. User interaction, for example in automatic narrative generation tools, allowing the user to alter the state of the world at any given point in a story, can so radically alter the world that even the most accommodating plot lines cannot survive. This raises the question of how far we can personalise a summary before losing the sense of coherence. For example, including only events involving the soccer fan's favourite player may result in a meaningless sequence of disjoint events, providing no understanding to the viewer of what actually happened in the game. We need to make sure that personalisation only takes place within the framework of coherent summarisation, for example on a context-group basis, in order to avoid this problem.

The literature includes a number of methods for measuring coherence, such as Latent Semantic Analysis [Foltz et al. (1998)], tree-depth measurement in rhetorical structure theory trees [Mani et al. (1998)] and ontology-based semantic coherence scoring [Gurevych et al. (2003)]. None of these techniques are particularly suited to our requirements. Foltz et al. (1998)'s method compares the vectors for two adjoining segments of text in a high dimensional semantic space, to see how similar they are to each other. When considering the coherence of soccer highlights however, we are

more interested in the *causal* relationships between summary elements rather than their *similarity*. A highlights package consisting solely of very similar elements, for example, lots of Fouls, would not be particularly coherent. Mani et al. (1998) mark up the rhetorical structure of their small corpus (containing only 5 texts) by hand in order to implement their coherence measure. We would prefer to avoid this manual step, and instead develop a fully automatic coherence measure.

Gurevych et al. (2003) map each semantic instance in their ontology to a node of a directed graph, with the graph edges representing the relationships between the ontology classes. *is-a* relations are given a weight of 0 and all other relationships are weighted with a 1. The distance between two instances is then the minimum path score between their two nodes, and is used as the measure of coherence between them. This method requires explicit semantic relations (such as *has-agent* or *has-object*) to be specified between the instances, whereas we only have causal relationships between our soccer events, the strength of which is measured by their probability of occurrence in sequence. Therefore, we introduce a probabilistic coherence measure in section 6.7 based on our model of the causal relationships between events, and use this to evaluate the trade-off between coherence and personalisation.

6.1.4 Personalisation of multimedia summaries

We now look at some contributions from personalised multimedia summarisation research. The traditional method for personalising a multimedia summary (for example in Ferman et al. (2002)) is to assign a weight to each of the user's preferences, and use these weightings to vary the scores of the multimedia content entities, so that a filtering agent can then determine which content should go into the personalised summary. The WebInEssence personalised multi-document summarisation and recommendation system [Radev et al. (2001)] allows weighted retrieval based on the positions of certain words in the web page. For example, a user can specify in their profile that they would like to give a weight of 5 for a keyword appearing in the title,

4 for the anchor, and 2 for the body. The user profile also contains a field for the type of search the user wants to carry out, such as Boolean or Vector Space search. Summary length can be selected as a percentage size of the original document, and the user can also choose the ordering of sentences within the document, according to various schemes such as position, time sequence or relevance to the query. Although a user profile as detailed as this allows a summary to be tailored closely to users' preferences, such a user interface may be too complicated for a busy, non-technical user. It may not be clear to the user how these weighting values affect the personalisation of content presented to them qualitatively. Our user profile, since it is explicit, only contains fields that are easily understood by the user. However, we are still faced with choosing weighting values for our features (user profile fields) and it is for this reason in section 6.5 we compare the weighting method with our biased training set technique, which avoids the need to select specific weighting values.

Alternatives to the weighting approach are presented in Smeaton et al. (2003) and Jaimes et al. (2002). The Físchlár project [Smeaton et al. (2003)] links together similar news stories to provide a personalised digest of the original broadcast, using text dialogue, name identification, shot boundary detection, speaker segmentation and matching, advertisement detection and speech versus music discrimination. The Físchlár user profile is based on Smyth and Cotter (2000)'s PTV system, described in section 6.1.2 on page 134. While similarity between news stories is computed based on text-dialogue and proper name identification, and links from one news item to another are presented, our aim is somewhat different: a system that could create a single, coherent story from relevant parts of one item.

Jaimes et al. (2002) propose another system for personalising soccer highlights using manually-annotated MPEG-7 video. This system requires each user to watch a set of training videos and note all the events which they find interesting. A neural network is then used to produce digests of similar events from new soccer matches

for that user. Aside from the practical problems of getting each user to spend time training the system, the resulting highlights again consist of a number of separate, not necessarily semantically connected, events. Although our work can easily be extended to multiple media, since we are summarising semantic events rather than content in a particular medium, this is not the focus of our personalisation research.

Finally, a note about evaluation of personalised summaries. Clearly, subjective evaluation would be the most satisfactory method; unfortunately, this is very time-consuming, and since users' opinions vary, such approaches require a sizeable number of participants to be statistically meaningful. Due to time limitations, our results will only be evaluated through comparison against each other (for example how the summaries change when different user profiles and biases are brought in to play) and by manual inspection and by objective measurements we introduce to quantitatively evaluate coherence, duration error and fulfillment of user requirements.

6.2 Biased summaries

Our first experiment investigates the bias of a summary according to different plotlines or narratives. Much of the scope for personalisation lies in the counterfactual: that is, in events that plausibly could have taken place, but didn't. For example, comments such as, "If only that player didn't stray offside, we'd have won" or "How could the referee not have called that a foul?" form the basis of many fanzine reports. In the future, computer vision behaviour analysis may offer options for this depth of personalisation, but for now, we are limited to the ticker-tape descriptions of the soccer games. Although these have to a certain extent, already 'neutralised' the multitude of interpretations that could be put upon an event, especially a controversial one such as a Foul, we believe that the event descriptions are still sufficiently rich to allow us to apply different biases to a summary.

To generate two types of bias in the summary, the 'controversial' subplot and the

'skill' subplot, we analyse our training set and divide the soccer summary descriptions into those that have more events from the Controversial Incident classes[3] than Goal Incident events[4], and vice versa. The ten soccer summaries with the largest (positive) difference between the number of controversial events and goal incidents are used to train a conditional probability matrix for a Markov chain to determine which events to include in the summary. Note that the probability of occurrence (denominator of equation 5.8) is still calculated using the original conditional probability matrix derived from the full training set, so that the biased training set only influences the likelihood of an event's inclusion in the summary. Context groups are designated according to the ticker tapes, rather than being learnt automatically, so that we can examine the variation in narrative bias in isolation, avoiding differences due to context group learning.

To measure the bias towards controversy in the summaries, the number of controversial events in the annotated summaries of television broadcasts was compared with the corresponding numbers in our automatically generated neutral and biased summaries. On average, it was found that there are 2.72 controversial events in each summary broadcast on television, but only 2.15 in each of our neutral summaries. When we biased the summary using the 'controversial' training set, the average number of controversial events in a summary increased to 7.73. Figure 6.1 shows a graph comparing the number of controversial incidents in neutral and biased summaries, along with the number in the broadcast summaries as a 'ground-truth' benchmark. It demonstrates that we can increase the number of controversial incidents in a summary by modifying our training set in this way. It also shows that our original method for generating a neutral summary often included too few controversial incidents, compared to the summaries broadcast on television.

[3]From figure 4.4, recall that these are *Fouls*, *Bookings*, *Sendings off*, *Offsides*, *Handballs* and *Substitutions*.

[4]Note that "Goal Incidents" refers to the superclass of *Goals*, *Shots*, *Saves* and *Assists*, not just *Goals* alone.

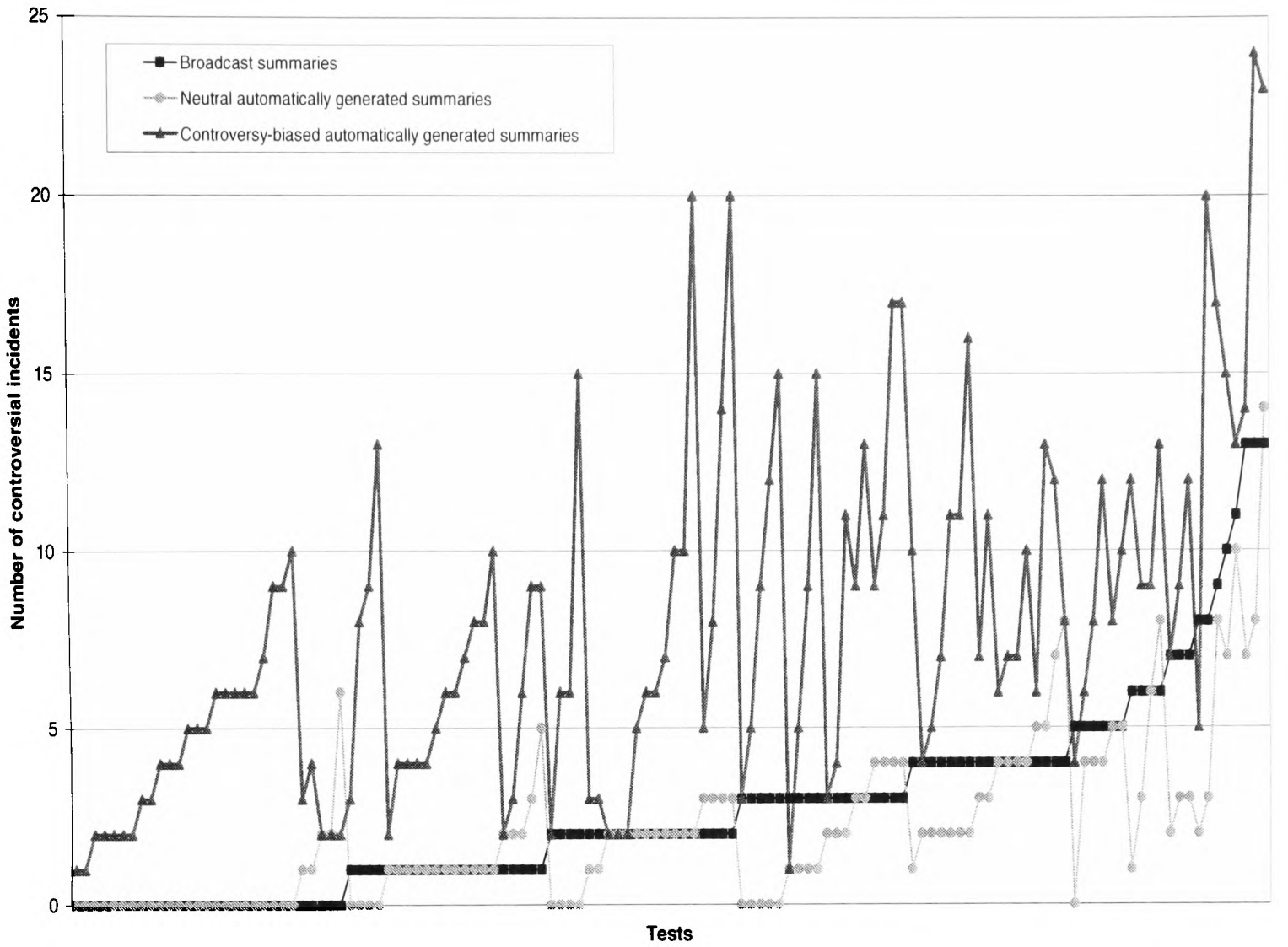


Figure 6.1: Frequency of Controversial Incidents in neutral and biased summaries, plotted for each individual test in our test set.

Repeating the experiment using a training set of the ten football games with the most goal incidents in their summaries, we find that we can also increases the average number of goal incidents in the summary (from a mean of 12.5 in a neutral summary, to 13.1 in a biased summary, compared with only 9.1 in the summaries broadcast on television). The increase for each test in our test set is also shown in figure 6.2. Again, the graph shows that we can increase the number of goal incidents in a summary, and we can see that our original method for generating a neutral summary actually included too many goal incidents, compared to the summaries broadcast on television.

It can be seen from the two graphs that in a few cases, the number of Controversial Incidents or Goal Incidents in the biased summary is actually smaller than in

Figure 6.2: Frequency of Goal Incidents in neutral and biased summaries, plotted for each individual test in our test set.

the neutral summary. This appears to be caused by our attempts to balance coherence against personalisation: the training set used to generate the 'biased' conditional probability matrix still contains examples of events other than Controversial Incidents or Goal Incidents respectively. This means that for the goal-driven bias, for example, a long context group containing one frequently occurring Goal Incident and several other events also frequently occurring in the biased training set can be given higher priority in the biased system than a shorter context group containing two Goal Incidents that occur less often in the biased training set. Owing to the summary time limit, this shorter context group may not make it into the summary, if the longer one is included first, whereas in a neutral summary, the longer context group may be assigned a lower priority, due to its length. (Recall that we do not normalise

for context group length, so the joint probability decreases for each additional event in the context group.) However, such a situation, where the biased summary has fewer Controversial or Goal Incidents than the neutral one, only rarely arises. and we consider it an acceptable trade-off against the added coherence advantages brought by the context-group concept. In the main, these results show that it is possible to bias the summary towards different narratives using the Markov chain summarisation method.

6.3 User profile design

Table 6.1 shows our user profile ontology, along with two instances, representing example users *Simon* and *Sarah*. The property values in these user profiles were chosen to reflect two users, one more interested in controversial events, and the other in skill and goal-related events. The properties themselves were selected partially in response to the user requirements elicited in the BUSMAN project [Evans (2003). discussed in section 3.2]. We also chose users with quite different needs in order that the differences in their personalised summaries would be large enough to measure objectively. Furthermore, some properties, such as levels of skill within the Shot event class. cannot be included in the user profile as they are outside the scope of our soccer ontology.

6.4 Personalised summary length

As seen in the previous section, one key property in the user profile is the user's required length of summary. In this section we look at how accurately we can generate a summary of the user's preferred duration (60 seconds for *Simon* and 5 minutes for *Sarah*). These times correspond to a compression ratio of 1.1% and 5.6% of original length respectively (assuming the minimum 90 minute game). Note that such compression ratios are larger than seen in most tests in the literature. Over

User profile property	Instance 1	Instance 2
Name	Simon	Sarah
Summary length	60 seconds	5 minutes
Favourite club	Manchester City	Everton
Secondary clubs	-	Arsenal
Favourite player(s)	David Seaman, Nicolas Anelka	Wayne Rooney, Thierry Henry
Favourite event	Goal	Goal
Second favourite event	Sending off	Penalty
Third favourite event	Foul	Shot
Fourth favourite event	Penalty	Save
Fifth favourite event	Booking	Assist

Table 6.1: Properties of the user profile class, along with values of two instances that will be used in personalisation experiments

the 126 tests. the mean average summary length error (using context-group based summarisation) is 30 seconds for *Simon* (50% error) and 23 seconds (7.7 % error) for *Sarah*. A graph of the percentage errors in duration over varying summary length is shown in figure 6.3.

The graph shows that single-event based summarisation is more accurate than context-group based summarisation, especially for shorter summaries, since an event's duration is of finer granularity than a context group's. The mean duration of a single event is 19.2 seconds, compared with 27.1 seconds for a context-group. However, beyond about five minutes (300 seconds) there is little advantage in using single-event based summaries, in terms of duration preference accuracy, and the advantage of context-group based summaries is in the additional coherence they provide to the overall summary. This coherence advantage is measured quantitatively in section 6.7. For very short summaries, the percentage error is very high, although this is only 20-30 seconds for each test, and is easily explained: given the length of a single event, the algorithm need only over-estimate by one event to have a substantial effect on the percentage error. The error decreases as the summary gets longer; this confirms research from the literature (such as Jing et al. (1998)) showing that the compression

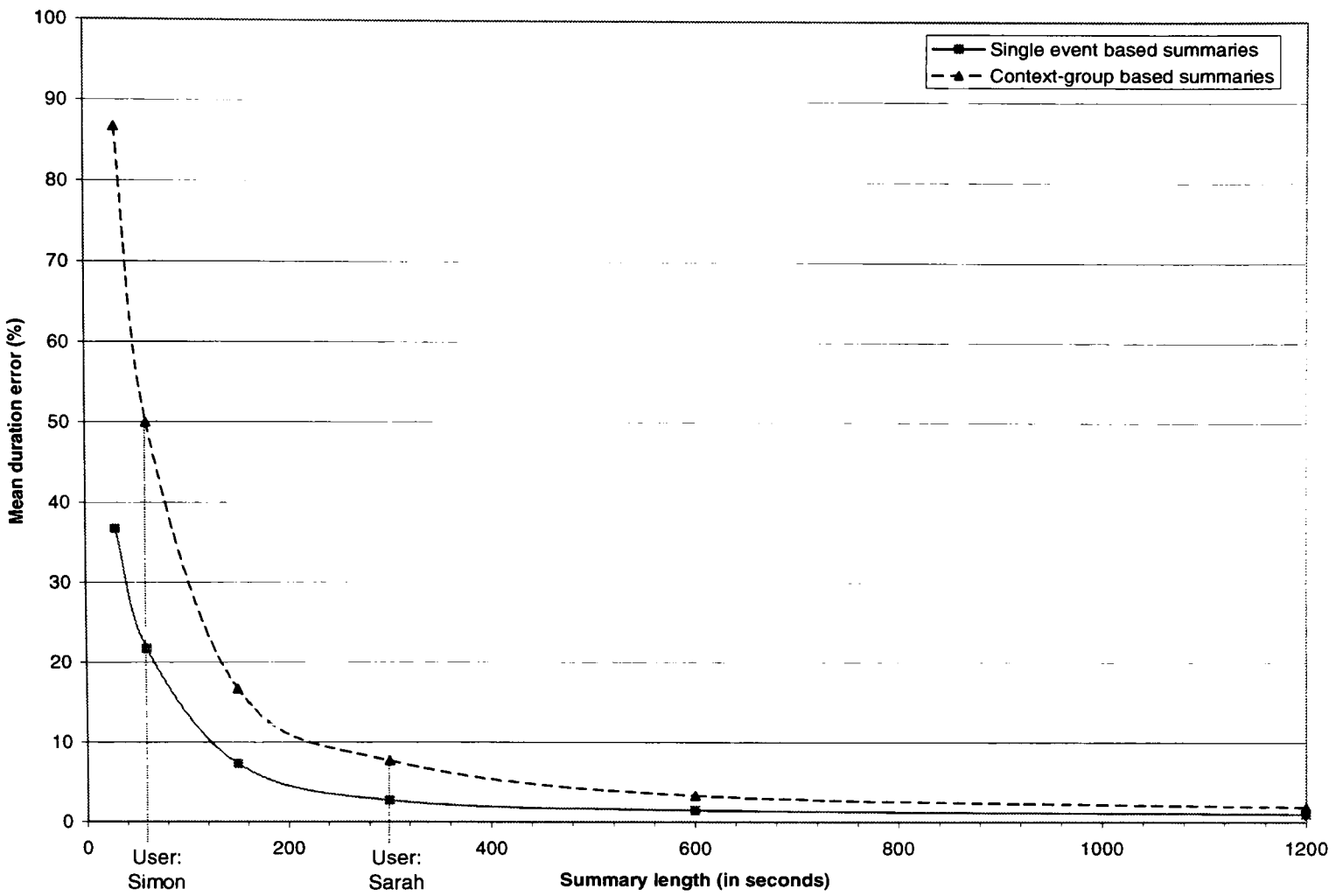


Figure 6.3: A graph of personalised summary duration error against summary length, with the two users *Simon* and *Sarah*'s preferred summary lengths marked.

ratio of a summary affects its quality: longer summaries tend to be of higher quality. The findings of this section lead to the recommendation that users should only be allowed to specify a preferred range of summary length, perhaps to the nearest minute, and charged accordingly, otherwise they may feel aggrieved that they are being charged for extra content that they didn't ask for.

6.5 Personalisation using the full user profile

We now take each of our two user profiles in turn, and generate a personalised summary for each of the tests in our test set. To personalise a summary, the priority of each context group in the full length game is first evaluated using the Markov chain method as before (conditional probability matrix trained using event class only, as in equation 5.8). Then, the score of events involving a favourite player are multiplied

with a weighting w_p, and the priority of event instances of the classes specified in user profile slots *favourite event* to *fifth favourite event* are multiplied by weightings w_1 to w_5 respectively. The weighting w_c is used to increase the score of 'positive' events involving the user's favourite club, and also of 'negative' events involving their opponents. In this way, the Schadenfreude effect is taken into account, so the user can enjoy seeing bad things happen to the opposing team. (Recall from the knowledge elicitation results in section 3.4.9 that Schadenfreude refers to the enjoyment by a fan when watching the opposing team do badly.) 'Positive events' are categorised to be Goals, Assists. Saves and Penalties and 'negative events' are events from the Controversial Incident superclass, except Penalties. A particular situation that may arise is when the user's favourite team is playing one of the 'secondary teams' they support. In this case, the secondary team is treated like any other opponent of the favourite team. The weightings we choose are as follows: $w_p = 5$; $w_c = 2$; $w_1 = 10000$, if the *favourite event* property value is "Goal", else $w_1 = 6$; $w_2 = 5$; $w_3 = 4$; $w_4 = 3$ and $w_5 = 2$. The personalised summaries are quite insensitive to variation in the weightings: varying them individually by + or - 10% has no effect on the results. We also compare this weighting method with the biased training set technique investigated in section 6.2.

We divide the tests into those containing the favourite clubs and player[5] (the 'Favourite Clubs' test set for a particular user) and those that do not (the 'Other Clubs' test set), since the number of events containing a favourite player or club only has meaning in the games in which they are involved. Then, the biased training set we used to summarise the Other Clubs test set consists of the ten games in our database containing the highest number of user-preferred events. To this we add all examples (except the test currently being summarised) of games involving that user's

[5]Since our test set consists of football matches from the 2003-04 season, the favourite players in the user profiles are always members of the corresponding user's favourite club, e.g. Wayne Rooney was playing for Everton in the 2003-04 season.

favourite clubs, to create a training set to summarise the Favourite Clubs test set.

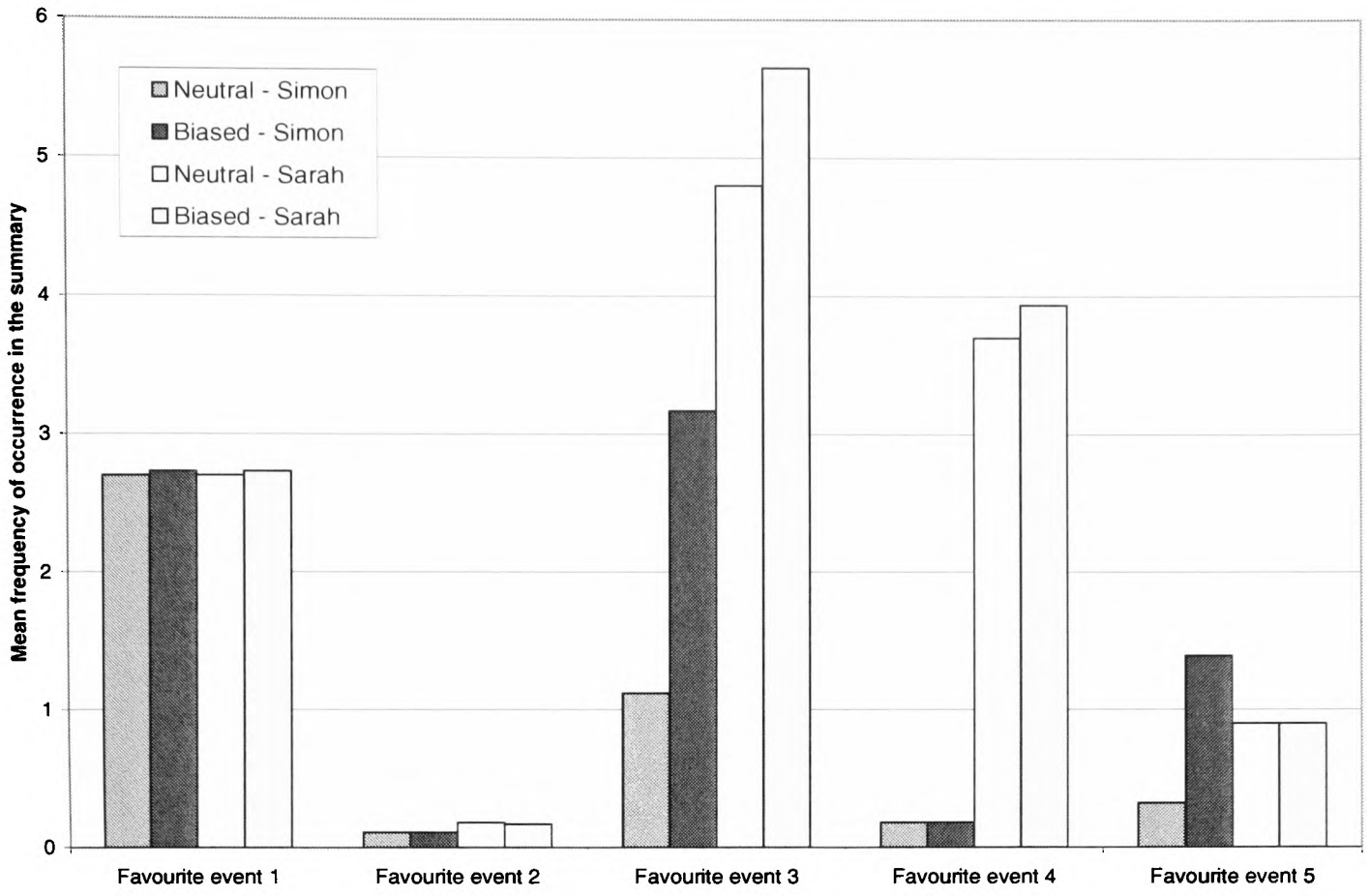
For both the weighting method and the biased training set method, we weight Goals very heavily, as people always want to see goals, according to both our knowledge elicitation study in Chapter 3 and Evans (2003). If we do not do this, we found that, due to the summary time limits, Goals which were normally included in the neutral summary were being left out in favour of events from the user's second or third favourite class.

To evaluate the personalised summaries we measure the mean number of events of the classes specified in each user profile. That is, mean number of favourite events, events involving a favourite player etc. The preferred summary length for each test in this experiment is set to be the same as that of the original broadcast highlights, so that a short summary duration does not affect the influence of personalisation. We found that a 60-second summary for *Simon* would contain only three or four events, making it quite difficult to evaluate what effect personalisation really had. This is as expected: too much personalisation of a very short summary would distort the game beyond recognition.

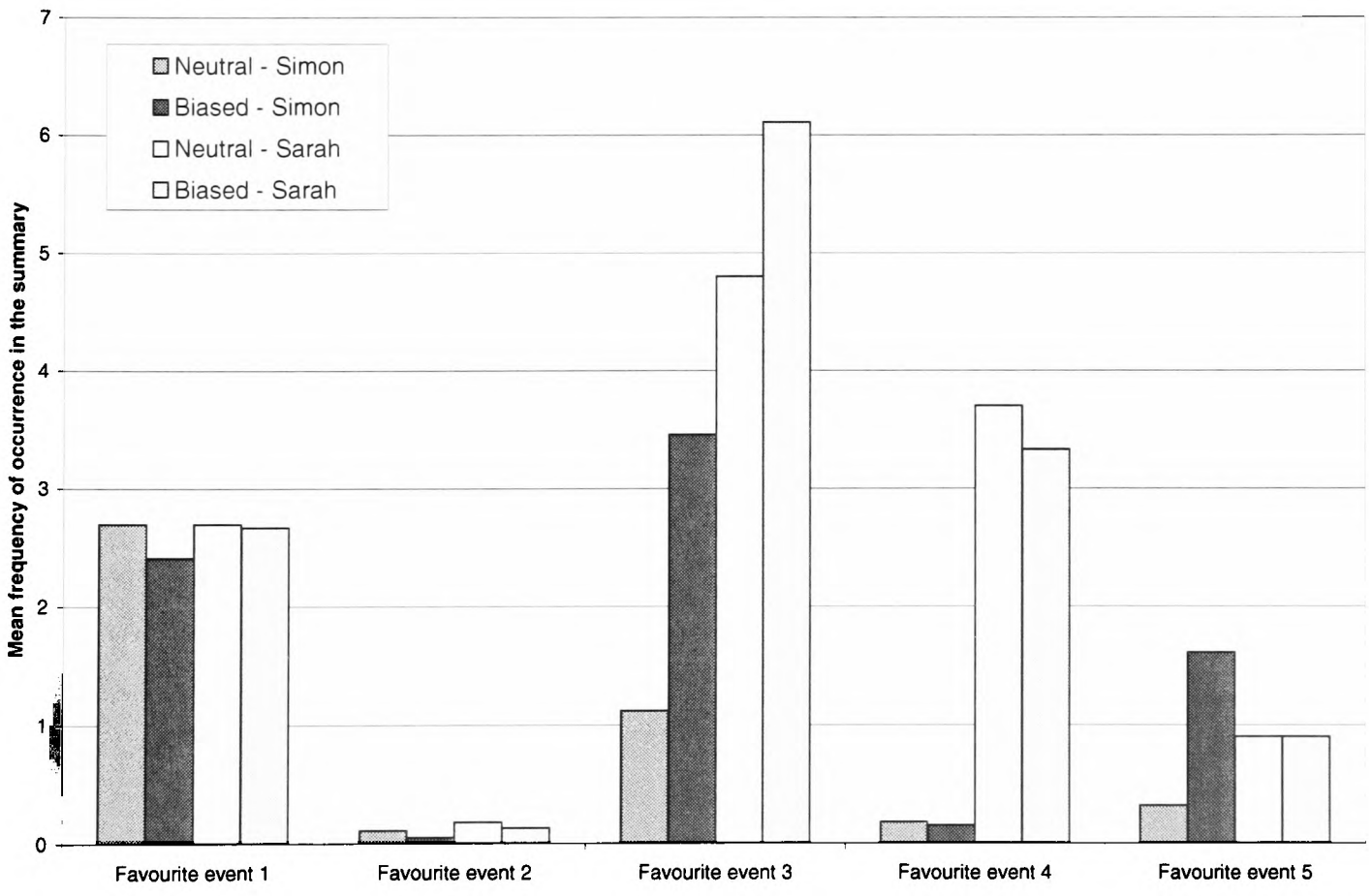
Results are presented in figures 6.4 and 6.5 which show the mean average number of events of the users' preferences in the summaries generated using different personalisation methods, compared with the neutral summaries.

To illustrate one concrete example, we now take an example test case, where the two users' favourite teams are playing each other, and evaluate how its summary changes when it is personalised for *Simon* and *Sarah*, compared to the neutral summary. The output of our summarisation system is shown in table 6.2.

All the summaries include Nicolas Anelka's goal. For *Simon*, this is the only context group that is included, due to his very short summary length preference. Anelka's goal is given priority over the other two goals that were scored in this game, as Anelka is listed as one of Simon's favourite players. The neutral summary and

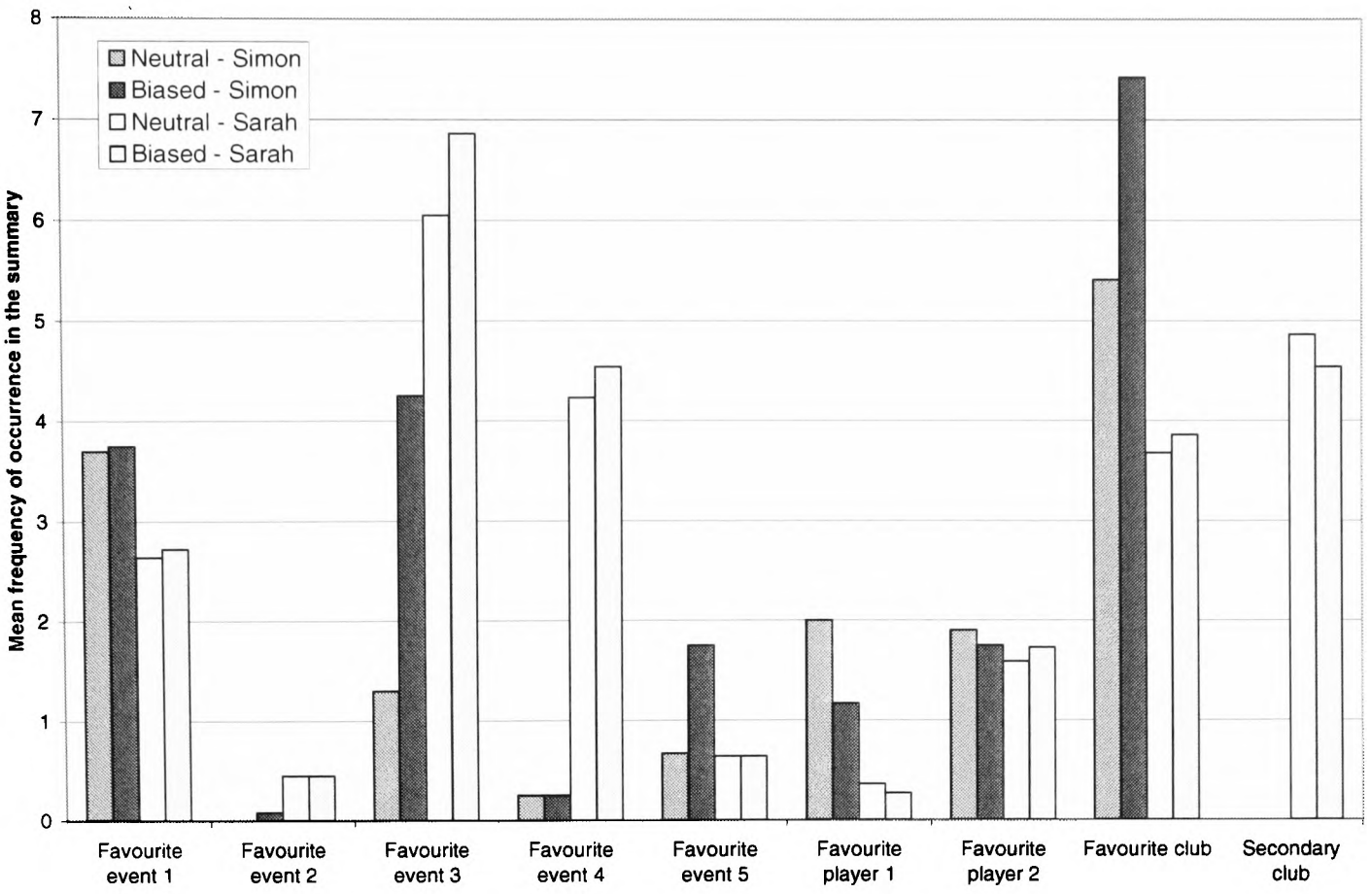


Weighting method, Other Clubs test set.

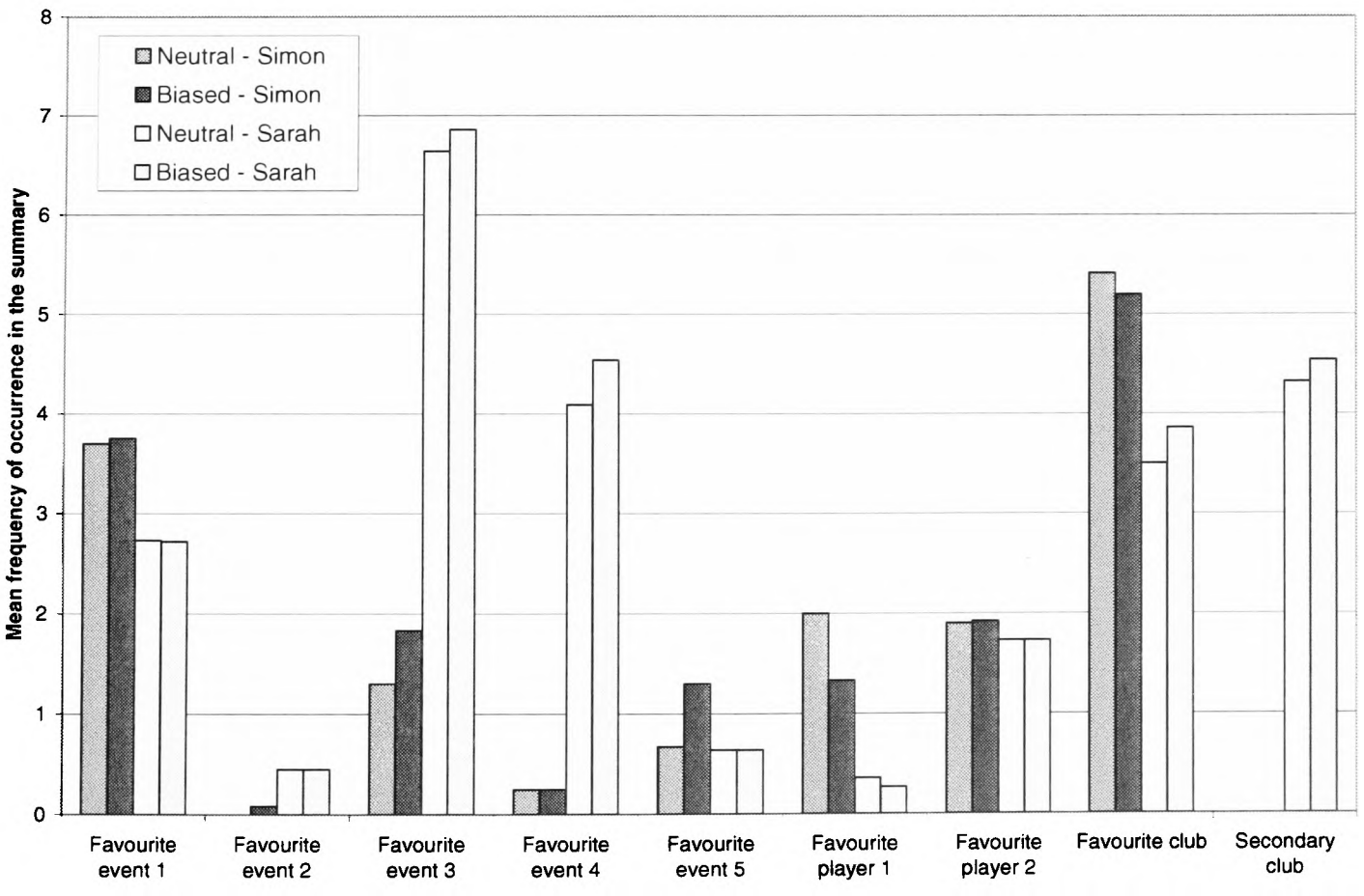


Training set bias method, Other Clubs test set.

Figure 6.4: Frequency of event occurrence in neutral and personalised summaries, comparing the weighting and training set bias methods of personalisation for the Other Clubs test set.



Weighting method, Favourite Clubs test set.



Training set bias method, Favourite Clubs test set.

Figure 6.5: Frequency of event occurrence in neutral and personalised summaries, comparing the weighting and training set bias methods of personalisation for the Favourite Clubs test set.

Context Group	Neutral summary	Summary personalised for *Sarah*	Summary personalised for *Simon*
36:27 Goal - Michael Tarnat. Arsenal 1-0 Man City. Cross by Thierry Henry (Arsenal), blocked by Michael Tarnat (Man City). Own goal by Michael Tarnat (Man City) left-footed (bottom-right of goal) from own half (18 yards). 136 sec.	✓	✓	X
41:29 Shot by Robert Pires (Arsenal) drilled right-footed from left channel (20 yards), save (caught) by David James (Man City) 40 sec.	✓	X	X
52:43 Shot by Sun Jihai (Man City) drilled left-footed from left channel (25 yards), save (caught) by Jens Lehmann (Arsenal). 64 sec.	✓	X	X
70:49 Shot by Robert Pires (Arsenal) drilled right-footed from centre of penalty area (18 yards), blocked by Richard Dunne (Man City). Shot by Edu (Arsenal) drilled left-footed from right channel (25 yards), missed left. Goal kick taken long by David James (Man City). 52 sec.	✓	X	X
73:17 Shot by Thierry Henry (Arsenal) drilled left-footed from left channel (25 yards), save (caught) by David James (Man City) 12 sec.	✓	X	X
78:30 Shot by Jose Antonio Reyes (Arsenal) right-footed from right channel (20 yards), save (parried) by David James (Man City) 27 sec.	✓	X	X
78:57 Shot by Thierry Henry (Arsenal) right-footed from centre of penalty area (6 yards), save (parried) by David James (Man City). 26 sec.	✓	X	X
79:53 Shot by Shaun Wright-Phillips (Man City) drilled left-footed from right channel (20 yards), save (tipped round post) by Jens Lehmann (Arsenal). Corner from right by-line taken short left-footed by Robbie Fowler (Man City). 24 sec.	✓	X	X

Table 6.2: A comparison between the neutral summary and summaries personalised for *Sarah* and *Simon* for an Arsenal versus Manchester City game. (*continued overleaf*)

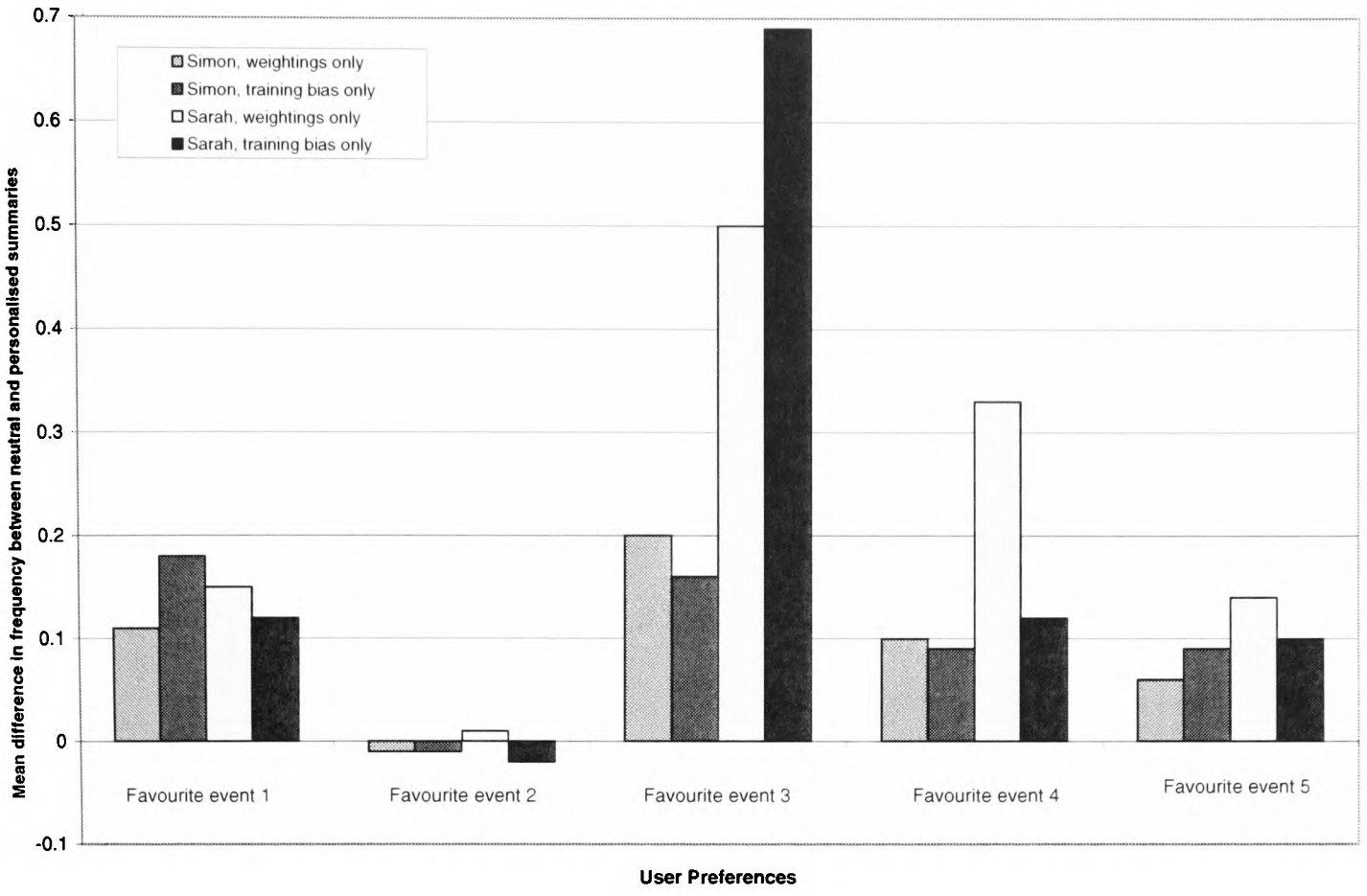
Context Group	Neutral summary	Summary personalised for *Sarah*	Summary personalised for *Simon*
80:17 Shot by Robbie Fowler (Man City) drilled right-footed from right side of penalty area (18 yards), save (caught) by Jens Lehmann (Arsenal). 28 sec.	✓	X	X
82:27 Goal - Thierry Henry.Arsenal 2-0 Man City. Goal by Thierry Henry (Arsenal) curled right-footed (top-right of goal) from left channel (20 yards). 59 sec.	✓	✓	X
88:06 Goal - Nicolas Anelka. Arsenal 2-1 Man City. Goal by Nicolas Anelka (Man City) left-footed (bottom-left of goal) from left side of penalty area (18 yards). Arsenal 2-1 Man City. Assist (pass) by Shaun Wright-Phillips (Man City) from left channel. 72 sec.	✓	✓	✓
89:19 Sent off. Nicolas Anelka (Man City) dismissed for violent conduct. Ashley Cole (Arsenal) booked for unsporting behaviour. 65 sec.	X	✓	X
Summary duration	**549 sec.**	**332 sec.**	**72 sec.**

Table 6.2: *(Continued)* A comparison between the neutral summary and summaries personalised for *Sarah* and *Simon* for an Arsenal versus Manchester City game.

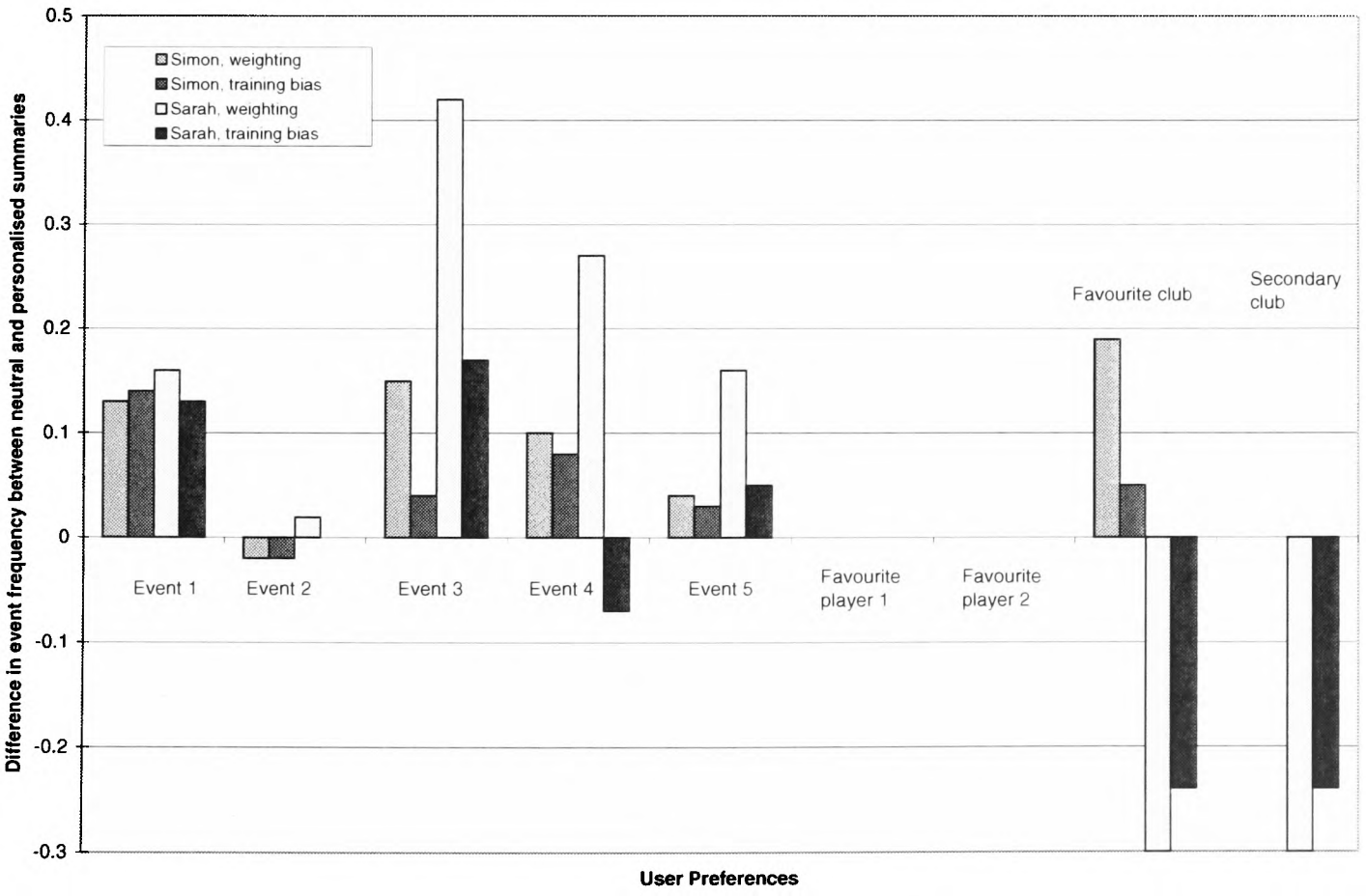
Sarah's both include all the goals, while the other context groups in the neutral summary do not appear in *Sarah*'s since her preferred length is only half that of the neutral summary. The sending off of a player who has just scored against her favourite team is the only context group appearing in *Sarah*'s summary and not in the neutral one: this is an example of the Schadenfreude effect.

The final set of results presented in this section are from experiments where the summary length requirement *is* taken into account (to give the complete picture of what each user's profile would deliver). Results are shown in figure 6.6.

Considering all the results presented in this section, we can draw several conclu-



Other Clubs test set.



Favourite Clubs test set.

Figure 6.6: Comparison of weighting and biased training set personalisation methods, evaluating the difference between the neutral and personalised summaries, where summary length is set to the user's preferred length, with the Other Clubs and Favourite Clubs test sets.

sions. Firstly, the effect of personalisation is quite dependent on the length of the summary. Comparing figures 6.4 and 6.5 with figure 6.6, it can be seen that for the shorter summaries in figure 6.6, personalisation does not always increase the number of preferred events, especially for the second favourite event. This is because, when the summary is very short, in order for an instance of a particular event class to be included in the personalised summary, an instance of another event class frequently occurring in the neutral summary must be left out. For example, the Sending Off and Penalty (the second favourite events for *Simon* and *Sarah* respectively) have a lower probability of occurrence, even when weighted, than the other favourite event classes in their profiles. There is a larger negative difference between the second favourite event frequency in the personalised and neutral summaries for the Favourite Clubs test set, as the weighting of events involving the favourite clubs means that these events also take up time in the summary. We have avoided a similar decrease in the most favourite event (Goal) frequency by weighting it heavily.

When the required summary lengths were set to those of the original broadcast summaries, generally ten minutes or longer, the positive effect of personalisation, as shown in figures 6.4 and 6.5, is much more pronounced. Both the weighting and biased training set methods show increases for most user profile properties, with the biased training set method showing larger changes, albeit some decreases. The advantage of the biased training set method is that it avoids the need to select weighting values, which might vary according to the application domain, as the weightings are implicitly learnt by calculation of the conditional probability matrix. Quantitatively, the differences between the two methods are only small, and we expect that some performance improvements could be achieved by tailoring the biased training set more carefully. The training set could be made up of more context groups containing events involving the favourite players and fewer groups involving other events. That is, the training set need not contain complete summaries (which themselves are unbiased)

and could be constructed from user feedback. This suggestion is discussed further in the future work section 8.2.

Overall, we have found that summary length has more of an influence on summary content than the user profile. This issue of the trade-off between providing a user's content preferences and fulfilling their summary length requirement is studied further in the next section.

6.6 Utility

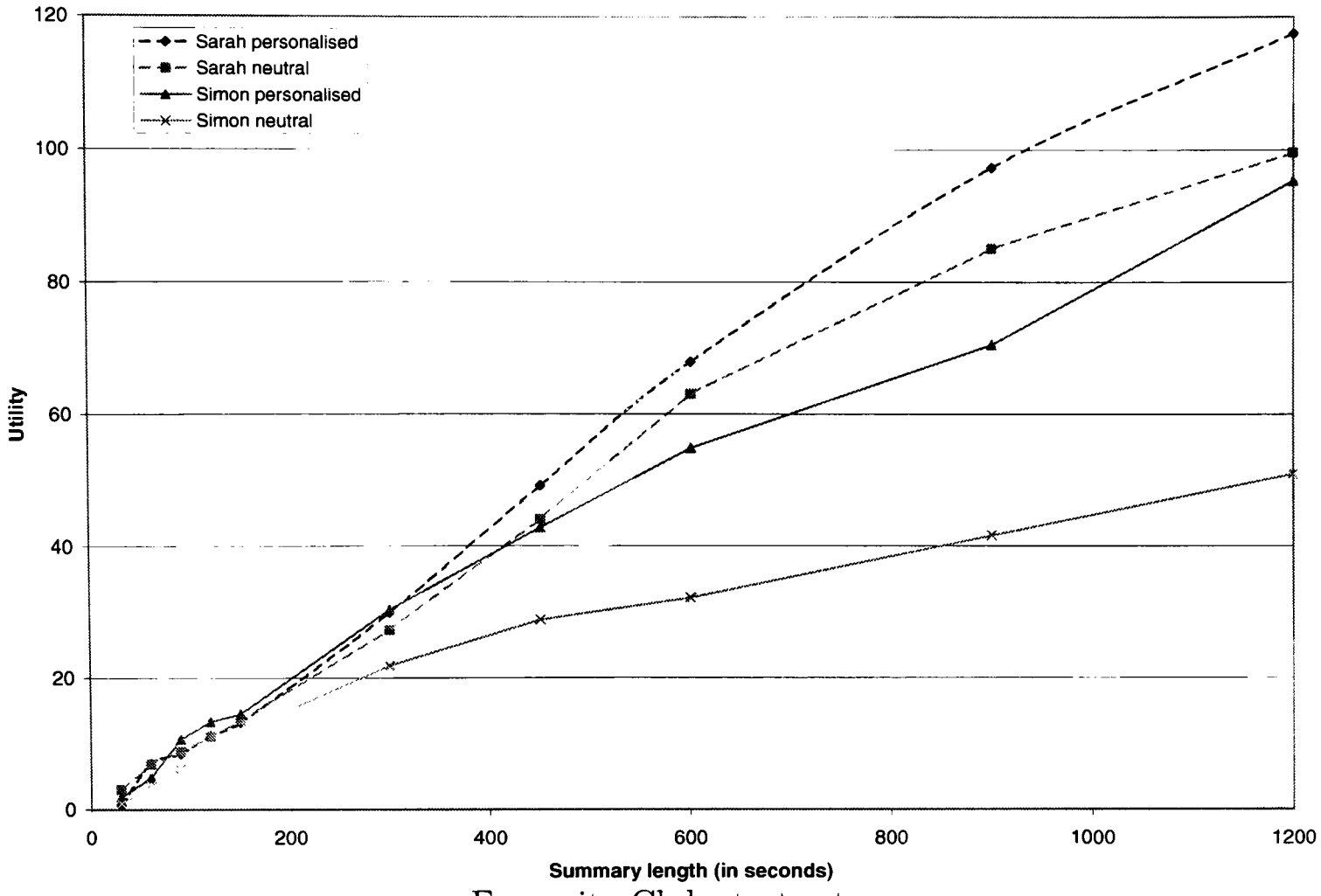
We would eventually like to be able to offer the user choices like, "If you pay for an extra minute of content, you can see that penalty everyone's talking about" or "We know you're an Everton supporter, would you pay for an extra five minutes of Everton's best moments in the game?" So in this section we investigate the trade-off between how well the content presented to the user fulfills their requirements, which we term *utility*, and the duration of the summary.

The major difficulty here is, of course, how to quantify utility appropriately, so that if the user does opt to extend their soccer highlights package, they are satisfied that it was money well spent. Since we are not carrying out subjective testing, our utility function makes several approximations, in order to evaluate summary quality based on the data that we have available. We define a utility function for a summary S and user profile U as:

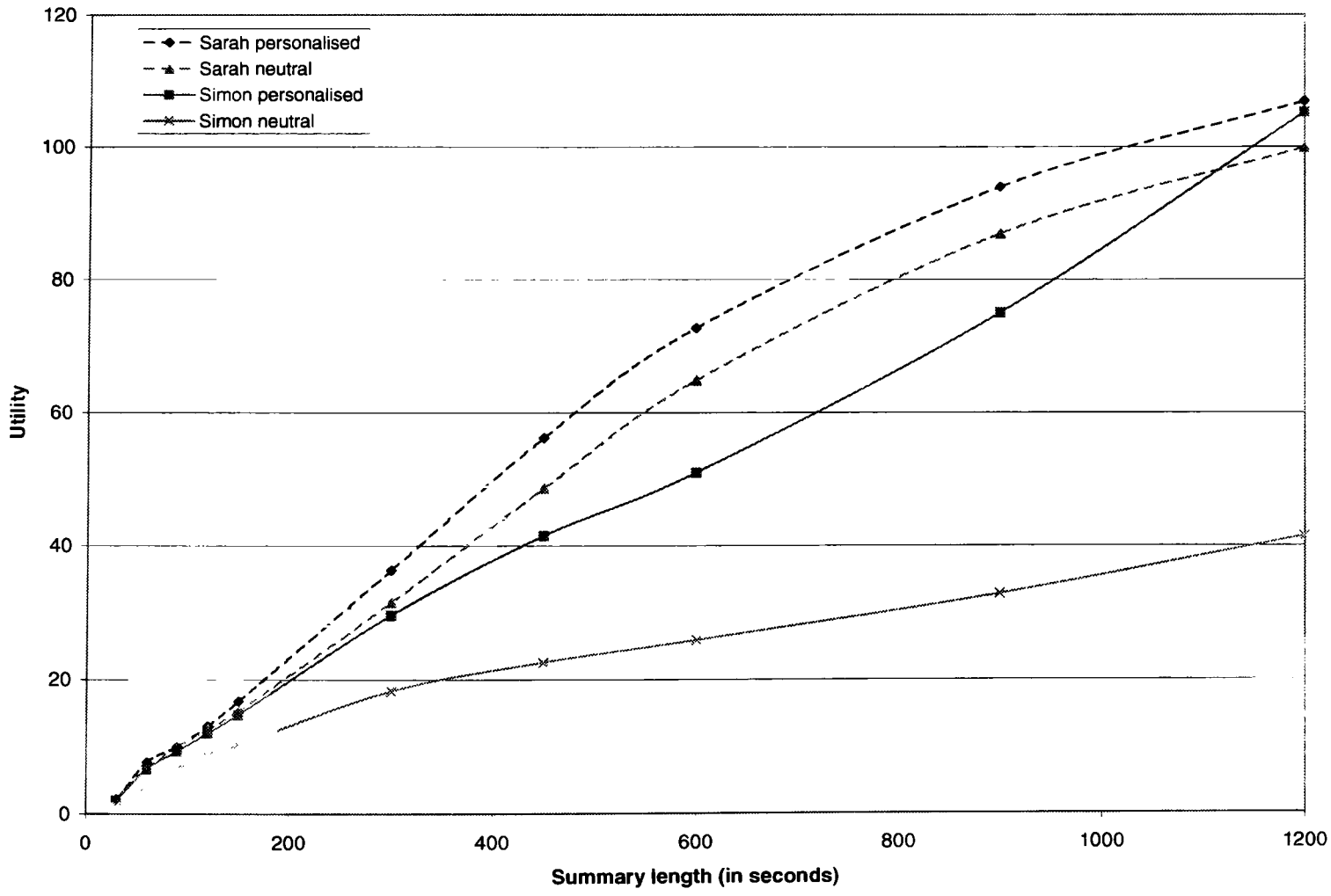
$$Utility(S,U) = \sum_{i=0}^{N} w_i * frequency(events\ of\ class\ U(i) \in S) \qquad (6.1)$$

where i is the index of user profile properties, and N is the number of properties in the user profile. The weightings w_i are the same defined in section 6.5 on page 146, except that w_1 is set to 6, not 10000.

Figure 6.7 shows how utility varies with summary length, for the Favourite Clubs and Other Clubs test sets, using the weighting personalisation method.



Favourite Clubs test set.



Other Clubs test set.

Figure 6.7: Personalised summary utility against summary length.

It can be seen that the Other Clubs test set has lower utility at all summary lengths than the Favourite Clubs test set, as the user is more satisfied when seeing their own team play. (Quantitatively, this is due to the additional w_c and w_p weightings.) Utility increases with summary length; for *Sarah* the rate of increase decreases with summary length, while for *Simon* it increases. *Simon* is a tougher customer to please than *Sarah*, although this difference is less noticeable at shorter summary lengths. As shown in section 6.2, *Sarah*'s favourite events, mainly Goal Incidents, are included more often in the neutral summaries than *Simon*'s favourites, mainly Controversial Incidents. This means that even the neutral summary is closer to *Sarah's* requirements than *Simon's*, also illustrated in figure 6.7. For example, in a neutral summary there are, on average 2.15 controversial events (Sending Off, Foul and Booking etc.) but 12.5 goal incidents (Goal, Shot, Save, Assist). *Sarah* requires a longer summary in order for her to notice the benefits of personalisation (beyond about 300 seconds), while *Simon*, would notice a difference in the utility of a neutral and personalised summary at about 100 seconds. For the same level of user satisfaction (utility = 30), Simon would need about 800 seconds of a neutral summary, compared with only 300 seconds in a personalised summary, for the Other Clubs test set. This means that personalisation allows him to achieve his viewing aims in less than half the time. An area for future work would be to investigate these trade-offs between cost of personalisation and time saving efficiencies for the user, as well as allowing personalisation to be targeted towards those users who would benefit most from the technique.

This section shows that we can objectively measure the fulfillment of individual user requirements, but, as discussed in the future work section 8.2, it would be useful to compare the performance of our measurement against the subjective opinions of real users.

6.7 Coherence measurements

The final experiment in this chapter looks at the trade-off between coherence and personalisation. To what extent is our suggestion valid that constraining personalisation to the context group level improves coherence? Our coherence measure for a summary is based on the causal relationships between its events, as calculated using the probability of occurrence of the entire sequence as a Markov chain. We are only interested in the relationship of one event to another, not whether the whole sequence is likely to occur, so we do not multiply by the marginal. That is, we actually want to calculate the conditional probability of the sequence, given the first event. Therefore, the coherence of a summary S, consisting of events $E_t, E_{t-1}, ...E_1$ is calculated as:

$$\begin{aligned} Coherence(S) &= \sqrt[t]{P(E_t, E_{t-1}, ...E_2|E_1)} \\ &= \sqrt[t]{\frac{P(E_t, E_{t-1}, ...E_2, E_1)}{P(E_1)}} \\ &= \sqrt[t]{P(E_t|E_{t-1}) \cdot P(E_{t-1}|P(E_{t-2}) ... \cdot P(E_2|E_1)} \end{aligned} \quad (6.2)$$

The coherence value is normalised over the number of events in the sequence by taking the t^{th} root. Since it is a probability, the coherence value can vary between 0 (completely incoherent) and 1 (perfect coherence).

Despite normalising, we found that coherence was higher for shorter summaries. For example, a 60 second summary personalised for *Simon* has a mean coherence of 0.61, compared with 0.11 for a 300 second one. This makes sense, because a shorter summary contains fewer edit points, and hence fewer jumps between non-causally related events. This may also be a reflection of the additional cognitive burden placed on the user to make sense of longer sequences. That is, a longer sequence may have a more coherent structure overall, but deeper contextual knowledge is required to understand it. It may be, using Mani et al. (1998)'s terminology, that we are in fact measuring cohesion here, the connectedness of the information, rather than

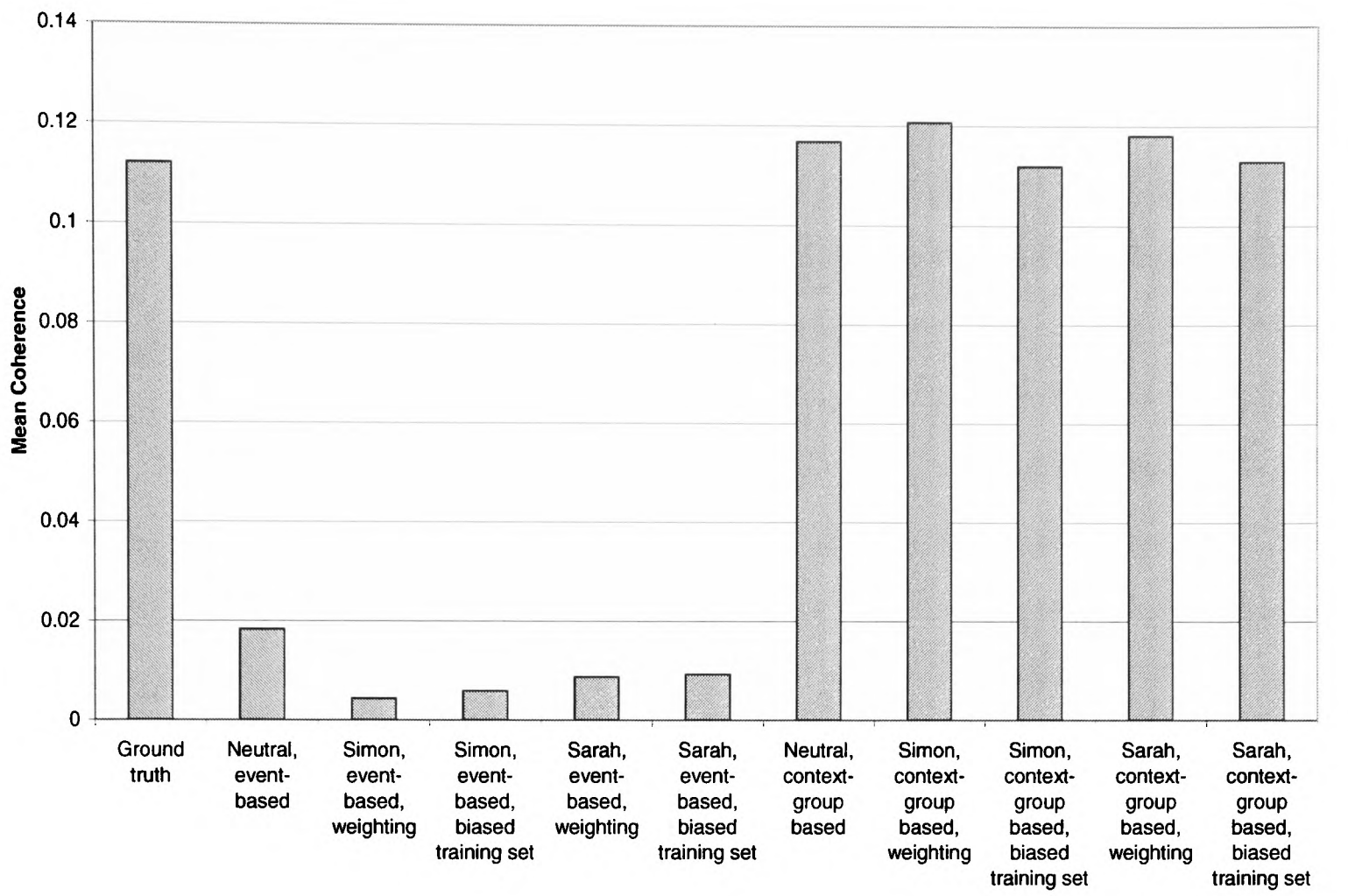


Figure 6.8: Mean coherence of various summaries; comparing ground-truth with neutral and personalised summaries.

coherence, the overall argumentative structure. Mani's delineation between the two is not clear, and perhaps this is an area for future work: to develop a measure of coherence at the higher level of narrative structure.

However for now, for the chart in figure 6.8, we generate summaries with the same required duration (that of the ground truth summary), so that a fair comparison can be made. The chart shows the mean coherence of the summaries broadcast on television, compared with our neutral summaries generated using both single-event based and context-group based summarisation; and summaries personalised for *Simon* and *Sarah*.

The results clearly show that coherence is much higher for context-group summarisation than when single events are included in the summaries. The small difference in *Simon* and *Sarah*'s results can be attributed to the small variations in

summary lengths. There is also no major difference between our two methods of personalisation (weighting and training-set biasing). It seems that the effect of context groups on improving coherence is much more significant than the influence of any particular method of personalisation. The surprising result is that the context-based weightings-personalised summaries are more coherent than the neutral ones. We expected that personalisation would reduce coherence in the summary, as suggested by Young (2000), and we do find this to be true for the event-based summaries. The increase in coherence due to personalisation for the context-group based summaries using the weightings method could be because we have chosen favourite event classes for our user profiles that frequently occur together and are causally related. For example, the group *Foul* → *Booking* → *Sending Off*, containing three of *Simon*'s favourite events, often occur together, and Sarah has *Shot* → *Save* or *Assist* → *Goal* causally related groups. By increasing the likelihood of these *groups* of events being included, through personalisation, we are actually improving the representation of causality in the summary, and hence a personalised summary is more coherent than a neutral one. Although, for context-group based summarisation, the biased training set method is less coherent than the neutral method, the decrease is much smaller than for the single-event based summarisation. This then is another advantage of context-groups: that they mitigate any reduction in coherence due to personalisation.

Figure 6.8 shows coherence measurements for the Markov chain method; for the CBR method developed in Chapter 4, coherence is 0.11 for the single-event based adaptation and 0.14 for the context group based adaptation. This is higher, in both cases, than the Markov chain method, significantly so for the single-event based adaptation. The single-event based adaptation result may almost match the coherence of the ground truth because the adaptation phase chooses the closest possible events in the test problem to the retrieved summary, choosing an event of the same class where possible. Since our coherence measure is based on event class relationships, the

coherence that was present in the summary retrieved from the case base is reflected in the events selected from the test case in the CBR method. To a lesser extent, the coherence between context groups in the retrieved summary is also preserved in the CBR adaptation, so the context-group based CBR method is also more coherent than the corresponding Markov chain method.

6.8 Personalisation summary

In this chapter we first reviewed some previous approaches to user-focused text and multimedia summarisation, user-profile elicitation and coherence measurement. In section 6.2 we introduced an alternative to the common weighting method for personalisation, by biasing a summary according to a 'controversial' plotline and a 'goal-driven' plotline, using different subsets of our training data. We found that for almost all tests, we could increase the number of Controversial or Goal Incidents in the summary by biasing the training set appropriately. In section 6.3 we presented the ontology for an explicit user profile, along with instances of two example users, influenced by the results of our knowledge elicitation study and work in the literature on user requirements for personalised soccer highlights.

Section 6.4 looked at how accurately we can generate a summary of the user's preferred duration: to our knowledge, the first study of its kind. We found that while single-event based summarisation is more accurate than context-group based summarisation, since an event's duration is of finer granularity than a context group's, this advantage decreased significantly with summary length. The mean percentage error between the actual and preferred summary length also decreases as the summary length increases. While most work in the literature limits compression to a lower bound of 10%, that is, the summary contains only 10% of the original material, our work investigates much heavier compression than this, down to 1.1%.

The weighting and biased training set methods were compared in section 6.5 and

the results showed that the effect of personalisation is quite dependent on the length of the summary, with longer summaries allowing the effect of personalisation to become more apparent. There was little difference between the weighting and biased training set methods: the weighting technique was slightly better, but the biased training set method has the advantage that it avoids the need to select the application domain. We believe that we could refine our selection of training set content, for example via user feedback, which could improve personalisation results.

The duration experiments have raised the question of how important it is to the user that the length of the summaries match the value indicated in their user profile exactly. We looked at a way of enticing a user to pay for extra content, or helping them save time, by plotting utility, a measure of fulfilment of user requirements, against summary duration in section 6.6. We found that utility increases with summary length, and, more interestingly, that *Sarah*'s utility, for both the neutral and personalised summaries, is always higher than *Simon*'s. This raises the issue of what personalisation *actually* changes and what the user *perceives* is being personalised for them. For example, generally all goals would be included in a summary anyway, so if a user specifies Goals as their favourite event, we don't actually *need* to personalise anything for them, and they would still be happy.

Finally, in section 6.7 we examined the trade-off between coherence and personalisation, as reported by Young (2000), using a new coherence measurement based on the causal relationships between events in the summary. We found that while this was an issue for single-event based summarisation, when context groups were used, the weighting method of personalisation actually increased coherence. While we have shown qualitatively throughout this thesis that the use of context groups improve coherence in summarisation, these coherence measurements now also demonstrate our argument quantitatively.

Chapter 7

Adaptability to other domains

So far in this thesis we have developed an information filtering system to generate personalised soccer highlights: in this chapter we examine how adaptable this is to other domains. As described in section 2.6.1, most ontology mapping techniques in the literature have been tested on very similar ontologies, for example two university course catalogues or two airline websites. This is because their objective is to reuse the ontology *vocabulary*, rather than reuse the *reasoning* carried out on instances of the ontology. In contrast, our aim is to reuse the summarisation process we have developed, by mapping to a substantially different ontology. The question we address in this chapter is therefore: to what extent can a system developed to summarise information in one domain, such as soccer, be reused to summarise information in another, significantly different domain, such as business meetings?

McKeown et al. (1999)'s experiments mapping a basketball summary to the stock market, described in section 2.4, are encouraging, since the authors have shown that summarisation rules developed in one domain can be at least partially reused in a markedly different domain. Entities in the soccer and meeting domains can be considered, at a high level of abstraction, to be semantically similar. For example, a soccer *Player* can map to a meeting *Participant*. However, we believe that the concept of context groups of causally-related events is also relevant to the business meeting domain. Often, the minutes of a meeting are too brief to be useful if somebody

needs to remember why a particular decision was taken, on what information it was based, or who supported or opposed the decision. We believe that a context group can be used to represent this type of "group memory" by modelling the sequence of events that lead up to a decision being taken. That is, our aim is not simply to reproduce the type of meeting summary that appears in the minutes, but provide a richer description of the debate that took place.

This chapter begins with a description of the business meeting domain and previous work on meeting analysis. Preliminary experiments to test out our hypothesis are described in section 7.2, along with a discussion of the results in section 7.3. Finally the chapter is briefly summarised in section 7.4.

7.1 The business meeting domain

We have chosen business meetings as an alternative domain to summarise for several reasons. Firstly, the interest of companies such as HP [Harville et al. (2003)] and Ricoh [Erol et al. (2003)] demonstrates its potential as a commercial application. Secondly, the meeting domain is sufficiently different from soccer that we cannot merely rely on superficial linguistic techniques like term similarity matching, which might be enough to map soccer to a closely related domain such as rugby. Instead, the similarity of the two domains is at a higher semantic level, and we are able to test the relevance of our context-group framework to the meeting summarisation domain. We might expect soccer and rugby football to be summarised in similar ways, since a sports editor may use similar techniques to edit highlights of any ball sport. By using the meeting domain for comparison however, we can test whether our Markov chain method is able to model the process of summary generation in a more general way.

7.1.1 Related work on meeting analysis and summarisation

As with sports' highlights generation, much of the work in the literature on meeting summarisation concentrates on low-level audio or video event recognition, as well as methods to compile key frames and video skims for browsing purposes. For example, Gatica-Perez et al. (2003) identify low-level motion such as standing up or speaking at a whiteboard, which are then clustered using HMMs to assign the higher level semantics of *note-taking, consensus, discussion, presentation, disagreement, monologue* and *white board.* Erol et al. (2003) generate key frames and skims of meetings using natural language processing of speech transcripts, audio analysis and video motion identification. For example, loud portions of the audio are assumed to correspond to important segments such as arguments within a meeting.

Although not specifically generating summaries, the CoAKTinG project [Bachler et al. (2003)] is of interest to us because of its meeting ontology, used to track processes and navigate resources before, during and after a meeting. Bachler et al have developed tools to transcribe argumentation and group memory from a meeting. Their ontology is quite small and contains the concepts *question, idea, pro, con, reference, note* and *decision.* The output of the system is a meeting trace: "a structured, collectively owned, searchable group memory that is generated in real time as a product of a meeting" which facilitates browsing. For example, a *decision* node can be retrieved, and then if the user is interested in further details they can access the video recording of the meeting at the point when the decision was made. We would like to do this automatically, that is, create a summary not only of the decisions, but the argumentation events that occurred prior to the decision, which would explain why certain judgements were made, and what opinions were held by whom.

The NIST Meeting Room Project [Garofolo et al. (2004)] uses work task descriptions as the basis for its ontology: *planning, brainstorming, negotiation, decision-making, competitive performance* (where a group within the meeting is competing

to win over the other group) and *dissemination of information.* However, the Multimodal Meeting Manager project [Marchand-Maillet (2003)] has developed a much richer meeting ontology "which is composed of meeting parts driven by an agenda ... Meeting participants take part in the meeting and can, at any moment, be in a given (physical) state performing a given activity." Meeting parts can be an instance of *Discussion*, *Vote*, *Presentation* (all with a *topic* property attached), *Silence* or *Break*, and they end with a meeting milestone such as a *Topic Change* or *Decision Point.* Each *Meeting Topic* instance also has an *importance* property attached. Our meeting ontology, described in the next section, is influenced by this work. However, we do not need an importance property to be specified a priori, as the event priority can be learnt by our Markov model. We also take into account some of the weaknesses of Marchand-Maillet's ontology when applied to summarisation. For example, our discussion items are split into several subclasses, to include what was said by different participants in support of or in opposition to the proposal, so that the content of a discussion item can be tracked, rather than only recording the decision outcome.

7.1.2 Meeting ontology and domain mapping design

Our meeting ontology is shown in figure 7.1. The classes are broad enough to describe meetings from many areas of business using the *topic* property in the top level *MeetingEvent* class. As with the soccer domain, a background knowledge ontology could also be developed, to included, for example, the roles of different meeting participants (CEO, Company Secretary etc.) However, since the purpose of this chapter is simply a concept demonstration, not a complete implementation, we are not using a background knowledge base in our experiments.

At this stage there are two alternatives open to us: either to retrain the Markov model using a full training set of meeting summary examples, or to use the original soccer training set and a mapping from the meeting to the soccer domain. The latter is the more challenging option, and we take it for the practical reason that we do not

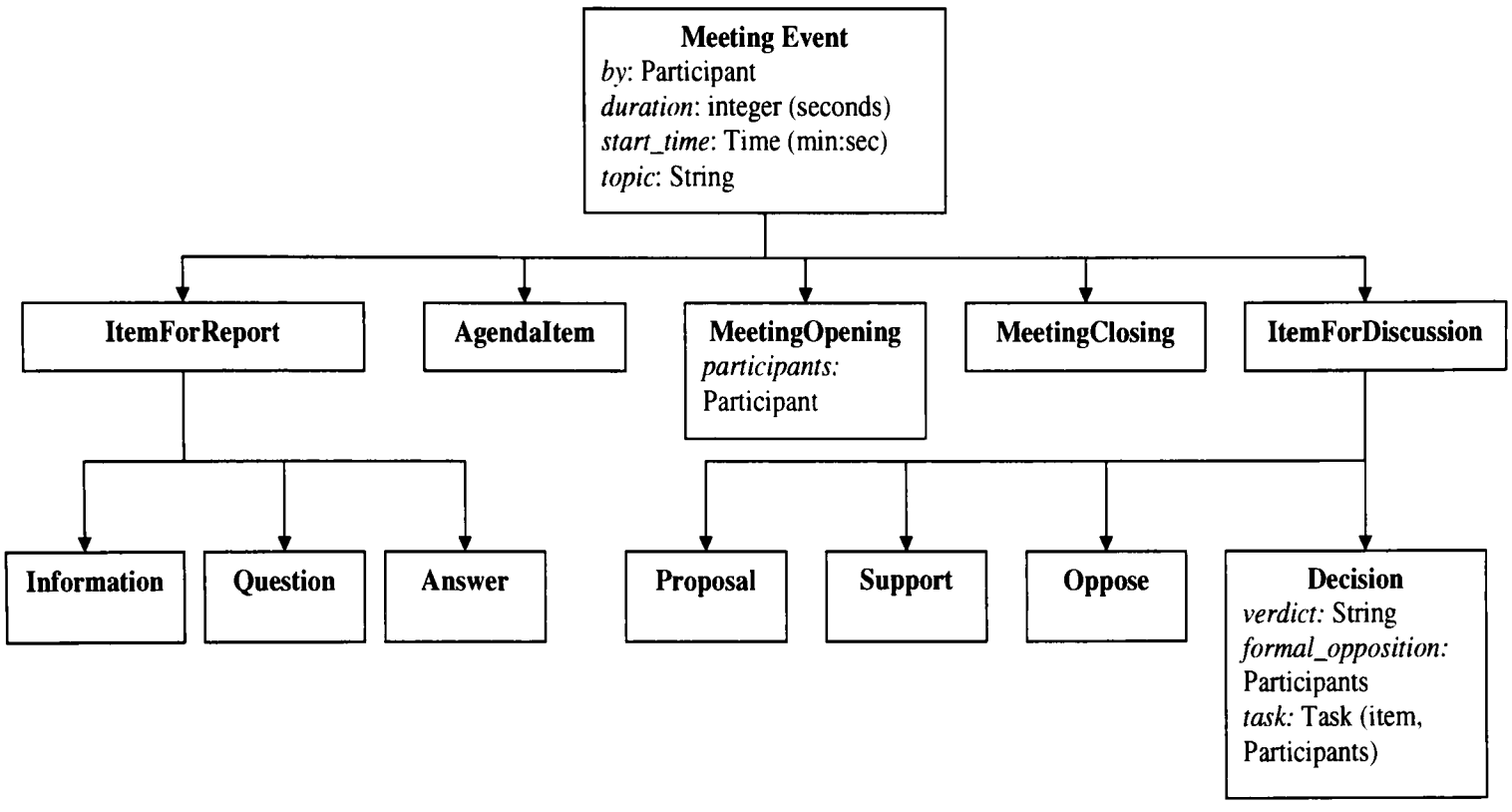


Figure 7.1: The business meeting ontology

have enough meeting summary data available to train the Markov model adequately.

The mapping between the two domains is based on high level semantic similarity, and is determined manually. It is not possible to map at this semantic level automatically using current technology, but such mapping information could be stored as metadata along with the ontology. So for the purposes of this very limited experiment, we will assume that the mappings, as shown in table 7.1, are available to us. As noted in the comments of table 7.1, the mappings are based not only on the semantic similarity between the event functions, but also, as far as possible, on the similarity of the causal relationships between events. Experimenting with mapping based on similarity of causal relationships has, as far as we are aware, not been tried before, and hence is a contribution made by this chapter.

7.2 Summarising business meetings

The input data was gathered by transcribing an audio recording of a meeting of the Board of Trustees of a charity. Although a charity was chosen because it was easier to

Meeting class	Soccer class	Comment
MeetingOpening	Goal	MeetingOpening might be better mapped to the soccer KickOff event, except it is not defined as a class in our soccer ontology.
AgendaItem	Goal	The rationale here is that reaching the next AgendaItem is an achievement, or goal of the meeting.
Information	Stoppage	An Information event can "restart" the flow of discourse in a meeting, in a similar way to a Stoppage event restarting soccer play.
Question Answer	Foul Booking	These two are less clear; however the causal relationships between Question-Answer and Foul-Booking are similar.
Decision	Goal	
Proposal	Shot	
Support	Assist	
Oppose	Save	
MeetingClosing	Block	MeetingClosing would more naturally map to a FinalWhistle soccer event, but it does not exist in our ontology. Block is chosen because it has very few causal relationships, like MeetingClosing.

Table 7.1: Mapping from Meeting classes to Soccer classes

obtain permission to record the meeting, it operates under company law, and hence the ontology we use is equally applicable to a board meeting for any company. The meeting transcription was anonymised and then converted to a "ticker-tape" format, as shown in appendix C. That is, the words used were limited to those available in our meeting ontology, so that it could be parsed into the system with the same template-mining approach used for the soccer ticker-tapes. Obviously this method suffers from the same weakness previously discussed for the soccer domain, namely that the ticker-tape has already been pre-filtered to some extent. Only the more interesting meeting occurrences will be listed in the meeting ticker-tape. However, the meeting ontology we use is sufficiently rich that there is still a substantial amount of filtering that needs to be done by our summarisation system (the examples shown

in tables 7.2 and 7.3 are compressed to 25% and 10% of original length respectively.)

The context groups were manually annotated, so that we limit the problem to the priority allocation step of summarising, not the context group clustering.

To obtain the marginal and conditional probabilities of occurrence and inclusion, $P(E_{t,O})$, $P(E_{t,O}|E_{t-1,O})$, $P(E_{t,IO})$ and $P(E_{t,IO}|E_{t-1,IO})$, we use the corresponding marginal and conditional matrices obtained for the soccer training set, as we have no meeting summary data available for training. To populate each element in the 10x1 marginal vectors and 10x10 conditional matrices, we use the corresponding value of the soccer classes which the meeting classes are mapped on to in table 7.1. For example, the conditional probability $P(E_t = Booking|E_{t-1} = Foul)$ is used as an estimate for $P(E_t = Answer|E_{t-1} = Question)$. The marginal and each column of the conditional probability matrix are then normalised so that each sums to 1.

We use the Markov chain summarisation method as described in section 5.3 to summarise the business meeting ticker-tape. The results, using single event and context-group based summarisation, with two different summary lengths, 25 minutes and the shorter 10 minutes, are shown in tables 7.2 and 7.3.

7.3 Discussion

The experiment in this chapter is very limited, and as such we are unable to carry out any subjective testing to evaluate our summary output in a more detailed way. However, from a visual inspection of the results in table 7.2 we can see that both the single-event and context-group based Markov chain methods produce reasonably successful summaries: the important information items, proposals, support, opposition and decisions are included, while irrelevant exchanges like "Why is this item on the agenda?" "Because it was last year" are excluded. This in itself is an important result: not only is our Markov chain method applicable to more than one domain, but we can even use training data from one domain to summarise data from a very

Meeting event description (event class, start time, speaker, duration, topic)	**Single event based summary**	**Context group based summary**
MeetingOpening 00:00 Dave 60 sec. "Present: Dave John Jan Alison Tony Rae Cathy Eddie Ben"	✓	✓
MeetingOpening 01:00 Jan 125 sec. "Apologies: Alastair"	✓	✓
AgendaItem 3:50 "Matters Arising"	✓	✓
Information 3:50 Jan 130 sec. "Interviewing for new staff member. Tony and I will interview 4 candidates on the 25th"	X	✓
Proposal 5:15 Jan 15 sec. "Who else can be there?"	X	✓
Support 5:30 Rae 20 sec. "I can"	✓	✓
Decision NemCon 5:50 5 sec. "Action: Tony Jan and Rae, interview 4 candidates on 25th."	✓	✓
Information 5:55 Jan 5 sec. "The Northampton charity had signed the naming agreement"	X	✓
Information 7:00 Jan 40 sec. "The Trustee Year Plan is done".	X	✓
Information 7:40 Tony 90 sec. "The site plan hasn't been done, now we're in the National Park it's harder to build. I've contacted Planning Aid".	X	✓
Information 9:10 Jan 60 sec. "Discussion of Birthday plans deferred as Alastair isn't here."	X	✓
Information 10:10 Jan 250 sec. "The date of Finance Sub Committee should be before 8th January. But in practice it's unlikely we can squeeze it in before the morning of the 8th"	X	✓
Proposal 14:20 Dave 20 sec. "Shall we wait until Chris has done the audit?"	X	✓
Decision 14:40 Nem Con. 10 sec. Choice of date for Finance Sub Committee deferred until Chris has done the audit	✓	✓
Proposal 14:50 Dave 10 sec. "Are the minutes accepted as correct?"	X	✓
Decision 15:00 Nem Con. 10 sec. Minutes accepted	✓	✓
AgendaItem 15:10 CEO Report	✓	✓
Information 24:15 Jan 140 sec. "The Charity Commission visit has changed to Tuesday 5th April. We do not have signed forms from the Trustees that they are not bankrupt, over 18 and have no convictions."	X	✓

Table 7.2: A 25 minute summary of a Board Meeting, using both single event and context-group based summarisation. (*continued overleaf*)

Meeting event description (event class, start time, speaker, duration, topic)	**Single event based summary**	**Context group based summary**
Proposal 26:35 Tony 10 sec. "I will circulate a full declaration to all for signature"	X	✓
Decision 26:45 Nem Con 5 sec. Action: Tony to circulate a declaration of eligibility of trusteeship to all trustees for signature	✓	✓
Information 26:50 Jan 30 sec. "Papers for January 8th meeting will be ready mid December, as I am on holiday from 20th December"	X	✓
Support 30:25 Alison 15 sec.	✓	X
Support 30:40 Tony 15 sec.	✓	X
Decision 32:30 Nem Con 210 sec. Action Jan to ask Nicky to look at alternative designs, with the likelihood that we will change the logo.	✓	X
AgendaItem 36:50 Process for the Auditors Report	✓	✓
Question Dave 38:30 15 sec. "Can it go via John so he can add comments?"	X	✓
Question 38:45 John 15 sec. "When will it be emailed?"	X	✓
Answer 39:00 Jan 30 sec. "December 12th. Then posted to trustees a few days later."	X	✓
Question 41:55 Dave 30 sec. "Can we get back to the point? Can we discuss this at the next meeting with Chris?"	X	✓
Proposal 42:25 Jan 15 sec. "Do we accept the process as it stands?"	X	✓
Support 42:40 Dave 50 sec. "The formal decision is still at the 8th January meeting, but we will look at it earlier. Agreed?"	✓	✓
Decision 43:30 Nem Con 50 sec. Action Jan to email auditor's final figures to John on December 12th. If he agrees. post to the trustees a few days later. Formal decision still at the 8th January meeting.	✓	✓
AgendaItem 44:20 October Finance Report	✓	✓
Information 44:20 John 100 sec. "The September report should have said 'we have a shortfall against budget of £X' rather than 'a deficit of £X' when we actually had a surplus that month"	X	✓

Table 7.2: *Continued...*A 25 minute summary of a Board Meeting, using both single event and context-group based summarisation. (*continued overleaf*)

Meeting event description (event class, start time, speaker, duration, topic)	**Single event based summary**	**Context group based summary**
Information 46:00 Jan 50 sec. "In the 2 months to October our income was £Y below budget while expenditure was £Z below budget. Leaving us with a net deficit £Y-Z adverse to budget"	X	✓
Information 50:45 Dave 10 sec. "We need to keep watching our income"	X	✓
Support 51:15 Ben 120 sec. "Can we get the Finance Sub Committee to look at the graphs, so that we can understand our income relative to budget?"	✓	X
Support 53:15 Dave 60 sec. "That can be an action item. Agreed?"	✓	X
Decision 54:15 Nem Con 70 sec. Action Finance Sub Committee to review financial graphs so that financial state relative to the budget is clear	✓	X
Decision 56:50 Nem Con 55 sec. Action Jan to follow up balance sheet discrepancy with Chris and Lydia. Action John to email Jan the details.	✓	X
AgendaItem 57:45 Group Kit Numbers Report	X	✓
AgendaItem 63:45 Ethical Policy		✓
Proposal 65:00 John 30 sec. "We should add 'for tools they purchase from the charity for their personal use'. "	X	✓
Support 65:30 Dave 30 sec. "OK, with that amendment, can we endorse this?"	✓	✓
Decision 66:00 Nem Con 5 sec. Ethical policy accepted, with the amendment to the last line 'for tools they purchase from the charity for their personal use'	X	✓
Question 69:0 Dave 40 sec "Can we move on?"	X	✓
AgendaItem 69:40 Update on Trustee Year Jobs	X	✓
Support 73:35 Dave 15 sec. "But the reality is of trustees' free time and getting things done."	✓	X
Support 75:20 Tony 10 sec. "Yes, you keep noticing that you haven't done it."	✓	X
Support 75:30 Ben 30 sec. "And when it becomes a priority."	✓	X
Support 76:00 Jan 75 sec. "So the year tasks for Employment Sub Committee, Finance Sub Committee and working groups get added to the agenda."	✓	X

Table 7.2: *Continued...*A 25 minute summary of a Board Meeting, using both single event and context-group based summarisation. (*continued overleaf*)

Meeting event description (event class, start time, speaker, duration, topic)	**Single event based summary**	**Context group based summary**
Support 77:15 Eddie 100 sec. "What about photocopying the Trustee Year Plan on the back of the agenda each meeting?"	✓	X
AgendaItem 79:10 Board of Trustees' Development	X	✓
Information 79:10 Dave 100 sec. "We're half an hour behind schedule, so we'll postpone Board of Trustees development to the next meeting"	X	✓
AgendaItem 80:50 Membership	X	✓
Proposal Eddie 80:50 45 sec. "2 new members have been proposed: Carol, Southampton and Philip from Emmer Green"	X	✓
Decision NemCon 81:35 10 sec. Accept as members	X	✓
AgendaItem 81:45 Organisational Strategy	X	✓
Support 84:40 Jan 245 sec. "You don't have to go to all 5 group weekend workshops. They collate the questionnaire data. The chair, it's more about getting other people to do it."	✓	X
Support 86:45 Dave 35 sec. "Did you say 'I'm the chair'? It's on the tape."	✓	X
Support 89:20 Ben 20 sec. "Ok, I'm prepared to do it."	✓	X
AgendaItem 89:50 Dates of Next Meetings	X	✓
Support 93:45 Dave 20 sec. "So we can have 45 minutes to 1 hour on our skills development. Then April 2nd can be a training workshop based on what we've identified at Ben's place."	✓	X
Support 94:50 Jan 65 sec. "OK. Email next time to remind me to put the heating on."	✓	X
AgendaItem 97:40 Any Other Business	✓	✓
MeetingClosing 98:50	X	✓

Table 7.2: *Continued...*A 25 minute summary of a Board Meeting, using both single event and context-group based summarisation.

Meeting event description (event class, start time, speaker, duration, topic)	**Single event based summary**	**Context group based summary**
MeetingOpening 00:00 Dave 60 sec. "Present: Dave John Jan Alison Tony Rae Cathy Eddie Ben"	X	✓
MeetingOpening 01:00 Jan 125 sec. "Apologies: Alastair"	X	✓
AgendaItem 3:50 "Matters Arising"	X	✓
Support 5:30 Rae 20 sec. "I can"	✓	X
Information 10:10 Jan 250 sec. "The date of Finance Sub Committee should be before 8th January. But in practice it's unlikely we can squeeze it in before the morning of the 8th"	X	✓
Proposal 14:20 Dave 20 sec. "Shall we wait until Chris has done the audit?"	X	✓
Decision 14:40 Nem Con. 10 sec. Choice of date for Finance Sub Committee deferred until Chris has done the audit	X	✓
Proposal 14:50 Dave 10 sec. "Are the minutes accepted as correct?"	X	✓
Decision 15:00 Nem Con. 10 sec. Minutes accepted	X	✓
AgendaItem 15:10 CEO Report	X	✓
Support 30:25 Alison 15 sec.	✓	X
Support 30:40 Tony 15 sec.	✓	X
AgendaItem 36:50 Process for the Auditors Report	X	✓
Support 42:40 Dave 50 sec. "The formal decision is still at the 8th January meeting, but we will look at it earlier. Agreed?"	✓	
AgendaItem 44:20 October Finance Report	X	✓
Support 51:15 Ben 120 sec. "Can we get the Finance Sub Committee to look at the graphs, so that we can understand our income relative to budget?"	✓	X
Support 53:15 Dave 60 sec. "That can be an action item. Agreed?"	✓	X
AgendaItem 57:45 Group Kit Numbers Report	X	✓
AgendaItem 63:45 Ethical Policy	X	✓
Proposal 65:00 John 30 sec. "We should add 'for tools they purchase from the charity for their personal use'. "	X	✓
Support 65:30 Dave 30 sec. "OK, with that amendment, can we endorse this?"	✓	✓

Table 7.3: A 10 minute summary of a Board Meeting, using both single event and context-group based summarisation *continued overleaf*)

Meeting event description (event class, start time, speaker, duration, topic)	**Single event based summary**	**Context group based summary**
Decision 66:00 Nem Con 5 sec. Ethical policy accepted, with the amendment to the last line 'for tools they purchase from the charity for their personal use'	X	✓
AgendaItem 69:40 Update on Trustee Year Jobs	X	✓
Support 73:35 Dave 15 sec. "But the reality is of trustees' free time and getting things done."	✓	X
Support 75:20 Tony 10 sec. "Yes, you keep noticing that you haven't done it."	✓	X
Support 75:30 Ben 30 sec. "And when it becomes a priority."	✓	X
Support 76:00 Jan 75 sec. "So the year tasks for Employment Sub Committee. Finance Sub Committee and working groups get added to the agenda."	✓	X
Support 77:15 Eddie 100 sec. "What about photocopying the Trustee Year Plan on the back of the agenda each meeting?"	✓	X
AgendaItem 79:10 Board of Trustees' Development	X	✓
AgendaItem 80:50 Membership	X	✓
Proposal Eddie 80:50 45 sec. "2 new members have been proposed: Carol, Southampton and Philip from Emmer Green"	X	✓
Decision NemCon 81:35 10 sec. Accept as members	X	✓
AgendaItem 81:45 Organisational Strategy	X	✓
AgendaItem 89:50 Dates of Next Meetings	X	✓
AgendaItem 97:40 Any Other Business	✓	✓

Table 7.3: *Continued...*A 10 minute summary of a Board Meeting, using both single event and context-group based summarisation

different domain.

Comparing the single-event based 25 minute summary with the context-group based one, seems to show that the context-group based method is better, as several important information items in the context-group based summary are not included in the single-event based summary. Furthermore, the debates that are included in the single-event based summary often only contain part of the exchange. For example, the debate over changing the charity's logo makes little sense as the *Proposal* isn't included. All the *Support* events are included, but not the *Oppose* events. This leaves the summary somewhat biased, as it looks as if there was no opposition to the decision. The ten minute summaries in table 7.3 also reveal weaknesses in the single-event based method, as several *Support* events are included, with no corresponding *Proposal* or *Decision* event, so the summary again doesn't make sense. The context-group method on the other hand, groups related events together in the summary, so that this disconnect does not happen.

We also found that the system output is very sensitive to the ontology mapping. For example, if we swap over the mappings of the *Proposal* and *Decision* classes (i.e. so *Proposal* maps to *Goal* and *Decision* maps to *Shot*) an additional 20 events that we judge to be irrelevant are included in the summary, out of a total of 50 summary events. Our system only outputs summaries that detail the cause and effect of decisions, which are useful for recording "group memories". Unfortunately, the priority of *Decision* events is not set high enough to enable meeting Minutes, containing only *AgendaItems*, *MeetingOpenings*, *MeetingClosing* and *Decisions*, to be generated. Further work needs to be carried out to ascertain the optimum mappings, so that not only "group memory" style summaries, but also shorter meeting Minutes can be generated. It may also be possible, rather than relying on changing the mappings, to use a personalised training set as in Chapter 6, to obtain Minutes-style summaries. Conversely, changing the mappings could also be a form of personalisation

or bias.

The summaries in the meeting domain need to be longer than those in the soccer domain to convey the gist of the meeting as a whole. This is not a problem if the summary is represented as text, but a 25 minute video summary is probably too long to be useful. Therefore further work is needed to identify the salient spoken sentence within each meeting event that has been annotated here. Table 7.1 detailing the mappings from the meeting classes to the soccer ontology classes might suggest that the soccer ontology ought to be extended to enable more relevant mappings (for example, including *Kick Off* and *Final Whistle* classes which *MeetingOpening* and *MeetingClosing* could map to.) The difficulty with this is that although the soccer and meeting domain ontologies might become more similar, it could then be harder to map from a third ontology to the soccer one. An alternative solution to increase the similarity between meeting classes and those in the training set might be to include examples from several different domains in the training set. Then the meeting ontology classes could be mapped to whichever class in any of the domains is the most similar.

7.4 Domain adaptability summary

Due to resource limitations, in this chapter we have only tried to demonstrate the feasibility of mapping our summarisation system to another domain. We began in section 7.1 with our reasons for choosing the business meeting domain and discussed related work in the literature on meeting analysis and summarisation. We also outlined our meeting ontology and mapping to the soccer domain. In section 7.2 we reported on an experiment to generate a meeting summary, and discussed the results in section 7.3. We found that the soccer summarisation system could successfully be mapped to the meeting domain, and that the context-group based method was better than the single-event one. More robust experiments are required to fully prove the

adaptability of the Markov chain method and context-group based summarisation to other domains, especially in the area of evaluation. It would be useful to have a comprehensive evaluation of our system's performance in the meeting domain, using more tests and subjectively rating summary contents.

We could also map to several other domains, to understand at what point the Markov chain method breaks down. (One suggestion is that it would not work in a domain whose summary output contains items that have no causal relationship to each other.) Other recommendations for future work are discussed in the next and final chapter.

Chapter 8

Conclusions

The things we thought would happen did not happen. The unexpected, God makes possible. **Euripides**

8.1 Thesis summary

Motivated by the information filtering problem, the primary contribution of this thesis has been the design and implementation of a system to summarise sequences of events by automatically modelling certain causal relationships between them. More specifically, we have shown that the coherence and accuracy of a summary can be increased by including "context groups" of causally related events in the summary, rather than one event at a time. The importance of these narrative episodes is supported by evidence from a knowledge elicitation study carried out with expert summarisers. We have also shown that the system can be personalised using either the traditional weighting method, or using by a technique we have developed which uses a biased subset of the training data. Our system has been robustly evaluated against summaries generated by experts in the soccer domain, and preliminary experiments show that our Markov chain method is also applicable in a substantially different domain, that of business meeting summary.

Chapter 2 discussed work in the literature to motivate our approach. In Chapter 3 we found that despite some difficulties in eliciting the knowledge from experts,

much useful information could be gleaned. The summary should contain the "talking points" of the soccer game such as goals, major fouls, controversial incidents or decisions, glaring misses and serious injuries, particularly those of significant players. The content and length of each event was chosen to maintain the flow of play and illustrate the story of the match. In the personalised data review, the theme of Schadenfreude was raised, whereby a fan enjoys seeing negative events involving the opposing team. The major finding of our knowledge elicitation study in Chapter 3 was that summarisation is best performed using a narrative structure. If the summary is being personalised, different subplots within a game can come to the fore or be demoted to change the bias of the story. We also found that several narratives were repeated consistently, e.g. the stronger team is expected to win; there is an unexpected change in behaviour of a team or player; or the strongest player is missing from the team. The episodic structure of the narrative allowed the flow of play to be represented and placed events in context. As well as showing the important events themselves (e.g. goals), we found that the events that cause them (e.g. the assisting passes) and events that they themselves cause should also be included in the highlights.

In Chapter 4 we developed a case based reasoning summarisation system for soccer and introduced the concept of the "context group" of causally related events. We found that case retrieval according to frequency of occurrence of each event type, weighted using the conditional probability of each event type being included in the highlights, given that it had occurred in the case, resulted in an increase of 6% precision and 5% recall over unweighted retrieval. Adaptation based on context groups rather than individual events also improved results by 4% precision and 29% recall, so that our system achieved an overall mean precision of 46% and recall of 52%. With relaxed acceptance criteria (correct event class only), we have shown that good summarisation results can be achieved: 82% precision and 88% recall. Our case base coverage analysis showed that over half the cases were redundant, and removing

them had little effect on accuracy. By measuring case base regularity, we found some problems in both case retrieval and to a lesser extent, adaptation. Problem-solution regularity for the full case base was very low, suggesting that our problem similarity measure did not correspond well to similar solutions. Weaknesses in the adaptation step were revealed by the result that even when all cases were adapted, thus bypassing the retrieval stage, precision was only 82% and recall 70%. For this reason, we adopted an alternative mechanism for summarisation in Chapter 5, using Markov models.

The main contribution of Chapter 5 was the development of a two-stage probabilistic alternative to CBR summarisation using Markov chains. At 59% precision and 65% recall, this was a 13% improvement on the CBR method for both measures. While the CBR method took a global approach, comparing one whole football game to another, it may have been that the Markov chain method was better able to model the local causality between events in the summary. As with CBR, the context-group based results from the Markov chain summaries were better than the single-event based summaries. This demonstrated that the episodic nature of the context group can, to some extent, introduce a more narrative style to the summary. Using both a Markov chain and Hidden Markov Model, we clustered causally related events into context groups, with similar performance for both methods. The second stage of the process was to use another Markov chain to assign a priority to each context group based on its probability of inclusion in the summary. The average F_1 results over the test set, at 61%, were 7.3% higher than those reported in the literature, as well as ours being a more flexible model which could generate a summary of any length. Since only the *event class* feature was used to describe each event in the Markov chain, another advantage of our approach was its ease of use in other domains. When investigating the use of background knowledge in the summarisation system, due to the 'curse of dimensionality' we found that the Markov chain mechanism set a limit

on the complexity of the semantics, and hence better results were achieved using the event class feature only.

In Chapter 6 we introduced an alternative to the traditional weighting method for personalisation, by biasing a summary according to a 'controversial' plotline and a 'goal-driven' plotline, using different subsets of our training data. We found that for almost all tests, we could increase the number of Controversial or Goal Incidents in the summary by biasing the training set appropriately. The effect of personalisation was found to be quite dependent on the length of the summary, with longer summaries allowing personalisation to become more apparent. There was little difference between the weighting and biased training set methods: the weighting technique was slightly better, but the weights had to be selected heuristically, and were domain-dependent, while the biased training set method was independent of the application domain.

We also presented results of a study measuring the accuracy of summary duration compared to different users' requirements: to our knowledge, the first study of its kind. We found that while single-event based summarisation was more accurate than context-group based summarisation, since an event's duration is of finer granularity than a context group's, this advantage decreased significantly with summary length. The mean percentage error between the actual and preferred summary length also decreased as the summary length increased. We introduced the concept of utility, a measure of fulfilment of user requirements which was found to increase with summary length. Using a new coherence measurement based on the causal relationships between events in the summary, we examined the trade-off between coherence and personalisation. We found that while this was an issue for single-event based summarisation, when context groups were used, the weighting method of personalisation actually increased coherence and the biased training set method mitigated the reduction in coherence due to personalisation to a large extent. These results quantitatively supported our hypothesis that the use of context groups improve coherence

in summarisation.

In Chapter 7, preliminary results showed that it was possible to generate a meeting summary using an ontology mapping and training data from the soccer domain. Again, the context-group based method was better than the single-event one. These findings supported our proposal that the Markov chain method to model causally related events would be applicable across domains and could potentially be used in any domain whose summary output contains items that are causally related to each other.

8.2 Future work

There are a number of directions that could be taken in work following on from this thesis:

Input and ontology enrichment

One of our original design criteria in section 1.1 was that it should be possible to eventually couple the system to a real audio-visual input. Although the many audio, video or closed-caption event recognisers currently available offer scope for further development, it would more be interesting to investigate how the system deals with the propagation of a recognition error. For example, if we were only 80% certain that the event was a *Goal*, how would this affect its priority of inclusion in the summary? Since we have taken a template-mining approach to extraction of information from ticker-tapes, our ontology has by necessity been limited; with an expanded ontology, could more event classes be recognised from an audio-visual signal, or would recognition technology instead narrow the ontology further? Could the advances in behaviour modelling from the computer vision community be used to recognise more complex, abstract concepts such as skill or humour? Although detecting or tracking an event is possible with current technology, there is a huge gulf between this and interpreting higher level semantics such as intentionality. For example, a *Foul* could be identified

via image analysis techniques, but far more research must be carried out before the description "deliberate foul by Player X on Player Y" can be extracted from the data. We have seen from our knowledge elicitation study that these descriptions are important in determining which instance out of many from the same class will be included in the summary; for example a particularly skilled *Shot* is shown over a more mundane one, and this would be an interesting area to research.

Another of our design criteria, the mobilisation of background knowledge about the domain, was addressed in Chapter 5. However, further work is needed to establish whether there is an optimal trade off between the size of the feature space and the additional information provided by the background knowledge features. For example, if only two or three additional features were used, and a larger training set was available, would performance improve?

Summarisation

In Chapter 5 we found that automatic clustering of events into context groups, using either a Markov chain or Hidden Markov Model, reduced performance compared with the use of the manually annotated context groups. There is room for improvement in this area, experimenting with categorisation of context groups into different types and then using a separate HMM for each type of context group. Profile or Hidden semi-Markov Models may also offer a way forward on this problem.

One major limitation that we have encountered in time resource allocation of events is the fixed granularity of event duration in our experiments. Following the lead of Crampes et al. (1998), a hierarchical method to allow editing at the frame level rather than event level could be researched. This would then require a more sophisticated resource allocation algorithm. A stochastic model of the time location of a particular event would also have to be developed. For example, an event could be given a fixed centre point in time. Its duration would then extend forwards and backwards in time from that point, according to some probability distribution, so that

frames further in the future would be less likely to belong to that event. In practice, if an audio-visual summary is being produced, silences in the audio commentary would signify the event cut-off point. This is another area of practical implementation that needs to be addressed: the co-ordination of the video with the audio, where full sentences or self-contained phrases represent the audio events.

Combining CBR and Markov methods

The Markov chain method gave better results than the Case based reasoning technique because it captured the local causal structure of the summary. Our implementation of case based reasoning on the other hand applied case based reasoning globally, retrieving a whole soccer game and its summary at a time. An area for future work would be to address this granularity problem by splitting each case into subcases, for example based on phases of each team's possession of the ball or violent episodes, and retrieving individual subcases from different games. Another opportunity to take advantage of the strengths of both CBR and Markov models and save on the computational complexity of personalising many different summaries would be by using CBR to match the current user's profile to the profile of a user who had already had a summary personalised for them via the Markov method. In this way, the personalised summary could be reused.

Personalisation

One fertile area for future research is in personalisation. Our biased training set method could be improved by tailoring the training set more carefully to include more examples of events involving the favourite player or club, and fewer irrelevant events. The issue of user feedback is also worth exploring. A Profile HMM could improve context group clustering by employing user feedback to estimate the probability of Delete and Insert states. This would facilitate 'long-range' context groups, that is, non-sequential events that are still causally related. For example, if a *Foul* causes a

Booking, and then, some time later in the game, another *Foul* causes a *SendingOff* event for the same player, all four events are actually causally related, but our current system would not reflect this.

Computer vision behaviour recognition also comes into its own at the personalisation stage. Different interpretations of the facts of a game, such as whether an event was a *Foul* or not, hotly contested by fans of opposing teams, would allow a deeper level of personalisation.

Another worthwhile avenue for study in personalisation is in subjective testing and user satisfaction. How well do our example user profiles match the requirements of real users, and how well do our personalised summaries live up to their expectations? Users' impressions of the trade-off between additional content and money or time resources should be evaluated in order to develop improved resource allocation algorithms. The utility measure we have developed would also benefit from the inclusion of additional features more accurately representing user satisfaction.

The user profile ontology could be developed to include additional properties concerning preferred output medium, and this is another area for personalisation: how the semantic events output by our system can be presented in different media. In the current system, the events are most easily mapped on to video events, using the *start time* and *duration* properties: key frame and closed caption, or an audio representation might require some adjustments to the summary content.

Narratives

We have only partially been able to address the issue raised in Chapter 3 of structuring the summary as a narrative. While we would argue that the context groups represent local narrative episodes, future work could combine them with a macro-level narrative structure, and other types of event relationships, such as those suggested by Lehnert (1981): motivation, actualisation, termination and equivalence. How exactly to extract these relationships automatically from the data remains an

open research question.

The coherence measure we developed in Chapter 6 could be improved to include these other event relationships as well as causality, and perhaps benchmarked against other approaches to coherence in the literature. We raised the question in Chapter 6 of whether our coherence metric actually measured what Mani et al. (1998) termed 'cohesion': the connectedness of the information, rather than 'coherence': the overall argumentative structure. If so, how can the coherence of a higher level narrative structure be measured?

Multiple Domain Summarisation

The final, and probably most compelling direction of future research is that of generalisation. The preliminary experiment in Chapter 7 extending our summarisation method to another domain raises many questions. Does the 'distance' of the second domain from soccer affect the quality of the result? For example, can a rugby match be summarised more easily than a business meeting using our method? Would a "mixed" training set comprising training examples from more than one domain improve generality? It would be interesting to test the hypothesis that our method is applicable to any domain except those whose summary output contains items that have no causal relationship to each other.

The ontology mapping itself also warrants further investigation: can such a high level semantic mapping (as we have created by hand in Chapter 7) be learnt automatically, for example, by inference from OWL rules or using similarity between causal relationships of events? How much knowledge of the two domains would be required by such an automatic system? There is also the influence of the ontology mapping error to consider, as we have already shown that the summary output is very sensitive to mapping choices. It will become more and more important to address such ontology mapping problems in the future, as the semantic web becomes a stronger influence on information filtering techniques.

Bibliography

A. Aamodt and E. Plaza. Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches. *AI Communications*, 7(1):39–59, 1994.

L. Agnihotri, N. Dimitrova, J. Kender, and J. Zimmerman. Study on Requirement Specifications for Personalized Multimedia Summarization. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 757–760, 2003.

A. Aleven and K. D. Ashley. How different is different? Arguing about the significance of similarities and differences. In L. Smith and B. Faltings, editors, *Advances in Case Based Reasoning: Proceedings of the Third European Workshop*, number 1168 in Lecture Notes in Artificial Intelligence, pages 1–15. Springer Verlag, 1996.

J. Assfalg, M. Bertini, C. Colombo, A. del Bimbo, and W. Nunziati. Semantic Annotation of Soccer Videos: Automatic highlights identification. *Computer Vision and Image Understanding*, 92(2-3):285–305, 2003.

M. S. Bachler, S. J. Buckingham Shum, D. C. De Roure, D. T. Michaelides, and K. R. Page. Ontological Mediation of Meeting Structure: Argumentation, Annotation and Navigation. In *Proceedings of the First International Workshop on Hypermedia and the Semantic Web*, HyperText'03, Nottingham, UK, 26-30 August 2003.

L. Bainbridge. Asking Questions and Accessing Knowledge. *Future Computing Systems*, 1:143–150, 1986.

M. Bal. *Narratology: Introduction to the Theory of Narrative.* University of Toronto Press, 2nd edition, 1978.

S. Barnes. Boyhood hero who gave England his best shot. *The Times*, 21 February 1996.

J. L. Bentley. Multidimensional Binary Search trees used for Associative Search. *Communications of the ACM*. 18(9):509–516, 1975.

T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 279, 2001.

T. Bray, J. Paoli, and C. M. Sperberg-McQueen. Extensible Markup Language (XML) 1.0. Technical report, World Wide Web Consortium Recommendation, 1998.

K. Brooks. Do Story Agents use Rocking Chairs? In *ACM Multimedia*, pages 317–328. ACM Press, 1996.

S. Brüninghaus and K. D. Ashley. The Role of Information Extraction for Textual CBR. In D. W. Aha, I. Watson, and Q. Yang, editors, *Case-Based Reasoning Research and Development: Proceedings of the 4th. International Conference on Case-Based Reasoning (ICCBR-01)*, number 2080 in Lecture Notes in Artificial Intelligence, pages 74–89, Vancouver, Canada, 30 July - 2 August 2001. Springer Verlag.

L. Capus and N. Tourigny. A Case-Based Reasoning Approach to Support Story Summarization. *International Journal of Intelligent Systems*, 18:877–891, 2003.

S. F. Chang, T. Sikora, and A. Purl. Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):688–695, June 2001.

J. Conroy and D. P. O'Leary. Text summarization via Hidden Markov Models and pivoted QR matrix decomposition. Technical Report CS-TR-4221, University of Maryland, February 2001.

M. Crampes, J. P. Veuillez, and S. Ranwez. Adaptive Narrative Abstraction. In *Proceedings of the ACM Conference on Hypertext*, pages 99–105. Pittsburgh, June 1998.

J. J. Daniels. *Retrieval of Passages for Information Reduction.* PhD thesis, University of Massachusetts Amherst, September 1997.

M. Davis and M. Travers. A Brief Overview of the Narrative Intelligence Reading Group. In M. Mateas and P. Sengers, editors, *AAAI Fall Symposium on Narrative Intelligence*, pages 11–16. AAAI Press, November 1998.

S. Deeble. My Work Space. *The Guardian*, page 18, 28 August 2004.

D. Diaper. *Knowledge Elicitation: Principles, Techniques and Applications.* Ellis Horwood, 1989.

Z. Ding and Y. Peng. A Probabilistic Extension to Ontology Language OWL. In *Proceedings of the 37th Hawaii International Conference on System Sciences*, page 4011, Big Island, Hawaii, Jan 5-8 2004. IEEE.

A. Divakaran, K. A. Peker, R. Radhakrishnan, Z. Xiong, and R. Cabasson. Video Summarization using MPEG-7 motion activity and audio descriptors. Technical Report TR-2003-34, Mitsubishi Electric Research Laboratory, May 2003.

A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to Map between Ontologies on the Semantic Web. In *Proceedings of the International World Wide Web Conference*, pages 662–673, 2002.

M. Dørum Jære, A. Aamodt, and P. Skalle. Representing Temporal Knowledge for Case-Based Prediction. In S. Craw and A. Preece, editors, *European Conference on Case Based Reasoning*, number 2416 in Lecture Notes in Artificial Intelligence, pages 174–188. Springer Verlag, 2002.

A. J. Duineveld, R. Stober, M. R. Weiden, B. Kenepa, and V. R. Benjamins. Wonder Tools? A Comparative Study of Ontological Engineering Tools. *International Journal of Human Computer Studies*, 52(6):1111–1133, 2000.

A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic Soccer Video Analysis and Summarization. *IEEE Transactions on Image Processing*, 12(7):796–807, 2003.

K. A. Ericsson and H. A Simon. *Protocol Analysis, Verbal Reports as Data*. MIT Press, Cambridge, Massachusetts, USA, 1993.

B. Erol, D-S. Lee, and J. Hull. Multimodal Summarization of Meeting Recordings. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, Baltimore, Maryland, 6-9 July 2003.

A. Evans. User-Centred Design of a Mobile Football Video Database. In *Proceedings of the Second International Conference on Mobile and Ubiquitous Multimedia*, pages 43–48, Norrkoping, Sweeden, December 2003.

A. M. Ferman, J. H. Errico, P. van Beek, and M. I. Sezan. Content-based filtering and personalization using structured metadata. In *Proceedings of the Joint Conference on Digital Libraries*, page 393, Portland, Oregon, 13-17 July 2002.

S. Fine, Y. Singer, and N. Tishby. The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, 32(1):41–62, 1998.

P. Foltz, W. Kintsch, and T. K. Landauer. The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Processes*, 25(2):285–307, 1998.

I. Frank. Football in Recent Times: What We Can Learn From the Newspapers. In *Proceedings of the First International Workshop on RoboCup*, volume 1395 of *Lecture Notes in Artificial Intelligence*, pages 216–230, Nagoya, Japan, 1997. Springer Verlag.

J. S. Garofolo, C. D. Laprun, M. Michel, V. M. Stanford, and E. Tabassi. The NIST Meeting Room Pilot Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 26-28 May 2004.

D. Gatica-Perez, I. McCowan, M. Barnard, S. Bengio, and H. Bourlard. On Automatic Annotation of Meeting Databases. In *IEEE International Conference on Image Processing*, Barcelona,Spain, 14-17 September 2003.

F. Gebhardt, A. Voß, A. Gräther, and B. Schmidt-Belz. Reasoning with Complex Cases. *International Series in Engineering and Computer Science*, 393, 1997. Kluwer Academic Publishers.

T. R. Gruber. A Translational Approach to Portable Ontology Specification. *Knowledge Acquisition*, 5:199–220, 1993.

K. M. Gupta, D. W. Aha, and N. Sandhu. Exploiting Taxonomic and Causal Relations in Conversational Case Retrieval. In S. Craw and A. Preece, editors, *European Conference on Case Based Reasoning*, number 2416 in Lecture Notes in Artificial Intelligence, pages 133–141. Springer Verlag, 2002.

I. Gurevych, R. Malaka, R. Porzel, and H. P. Zorn. Semantic Coherence Scoring using an Ontology. In *Proceedings of the Joint Human Language Technology and Northern Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 88–95, Edmonton, Canada, 2003.

M. Harville, M. Covell, and D. Wee. An Architecture for Componentized Network-Based Media Services. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 2, pages 333–336, Baltimore, Maryland, 6-9 July 2003.

V. Hatzivassiloglou and K. McKeown. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 172–182, 1993.

M. A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, 1994.

E. Hovy, N. Ide, R. Freerking, J. Mariani, and A. Zampolli. Multilingual Information Management: Current Levels and Future Abilities. Technical report, National Science Foundation, April 1999. Chapter 3, Cross-lingual Information Extraction and Automated Text Summarization.

J. Hunter. Adding multimedia to the Semantic Web - Building an MPEG-7 Ontology. In *Proceedings of the International Semantic Web Working Symposium*, pages 261–281, July 30 - August 1 2001.

A. Jaimes, T. Echigo, M. Teraguchi, and F. Satoh. Learning Personalized Video Highlights from Detailed MPEG-7 Metadata. In *IEEE International Conference on Image Processing*, pages 133–136, 2002.

H. Jing, K. McKeown, R. Barzilay, and M. Elhadad. Summarization Evaluation Methods: Experiments and Analysis. In *AAAI Symposium on Intelligence Summarization*, Stanford University, California, March 23 - 25 1998.

Y. Kalfoglou and M. Schorlemmer. Information Flow-based Ontology Mapping. In *Proceedings of the First International Conference on Ontologies, Databases and Applications of Semantics*, pages 1132–1151, October 2002.

Y. Kalfoglou and M. Schorlemmer. Ontology mapping: The State of the Art. *The Knowledge Engineering Review*, 18(1):1–31, 2003.

A. Kidd. *Knowledge Acquisition for Expert Systems, A Practical Handbook.* Plenum Press, 1987.

S. Kim, H. Alani, W. Hall, P. Lewis, D. Millard, N. Shadbolt, and M. Weal. Artequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web. In *Semantic Authoring, Annotation and Knowledge Markup Workshop*, European Conference on Artificial Intelligence, pages 1–6, Lyon, France, 2002.

J. L. Kolodner. Reconstructive Memory: A Computer Model. *Cognitive Science*, 7 (4):281–328, 1983.

J. L. Kolodner. Improving Human Decision Making through Case-Based Decision Aiding. *AI Magazine*, 12(2):52–68, Summer 1991.

J. L. Kolodner. *Case-Based Reasoning.* Morgan Kaufmann, 1993.

P. Koton. *Using experience in learning and problem solving.* PhD thesis, Massachusetts Institute of Technology, 1989.

J. Kramer, S. Noronha, and J. Vergo. A User-Centered Design approach to Personalization. *Communications of the ACM*, 43(8):45–48, 2000.

A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *Journal of Molecular Biology*, 235:1501 – 1531, 1994.

L. V. Kuleshov. *Kuleshov on Film: writings by Lev Kuleshov.* University of California Press, 1974. ISBN 0520026594. R. Levaco, editor.

O. Lassila and R. R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. Technical report, World Wide Web Consortium Recommendation, 1999.

M. Lawson, N. Kemp, M. Lynch, and G. Chowdhury. Automatic extraction of citations from the text of English language patents: An example of template mining. *Journal of Information Science*, 22(6):423–436, 1996.

D. Leake and D. Wilson. Case-base maintenance: Dimensions and Directions. In P. Cunningham, B. Smyth, and M. Keane, editors, *Proceedings of the Fourth European Workshop on Case-Based Reasoning*, pages 196–207, Berlin, 1998. Springer Verlag.

D. Leake and D. Wilson. When experience is wrong: Examining CBR for changing tasks and environments. In *Proceedings of the Third International Conference on Case Based Reasoning*, pages 218–232, Berlin, 1999. Springer Verlag.

W. G. Lehnert. Plot Units: A Narrative Summarization Strategy. *Cognitive Science*, 4:293–331, 1981.

M. Lenz and H.-D. Burkhard. Case Retrieval Nets: Basic ideas and extensions. In G. Görz and S. Hölldobler, editors, *KI-96 Advances in Artifical Intelligence*, volume 1137 of *Lecture Notes in Artificial Intelligence*, pages 227–239. Springer Verlag, 1996.

C. A. Lindley and F. Nack. Hybrid Narrative and Categorical Strategies for Interactive and Dynamic Video Presentation Generation. *The New Review of Hypermedia and Multimedia*, 6:111 – 145, 2000.

J. M. Mandler. *Stories, Scripts and Scenes: Aspects of Schema Theory.* John M. MacEachran Memorial Lecture Series. Lawrence Erlbaum Associates, 1984.

I. Mani and E. Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI'97)*, pages 622–628, 1997.

I. Mani and E. Bloedorn. Machine Learning of Generic and User-focused summarization. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI '98)*, pages 821–826, 1998.

I. Mani, E. Bloedorn, and B. Gates. Using cohesion and coherence models for text summarization. In *Proceedings of the AAAI Spring Symposium on Intelligence Text Summarization*, pages 69–76, 1998.

I. Mani, D. House, G. Klein, L. Hirschman, T. Firmin, and B. Sundheim. The TIPSTER SUMMAC Text Summarization Evaluation. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, page 77. Bergen, Norway, 1999.

S. Marchand-Maillet. Meeting Record Modelling for Enhanced Browsing. Technical Report 03.01, University of Geneva Center for Computer Science, Computer Vision and Multimedia Laboratory, March 2003.

M. Maudlin. *Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing.* Kluwer Press, September 1991.

M. T. Maybury. Generating Summaries from Event Data. *Information Processing and Management*, 31(5):735–751, 1995.

M. T. Maybury, W. Greiff, S. Boykin, J. Ponte, C. McHenry, and L. Ferro. Personalcasting: Tailored broadcast news. *User Modeling and User-Adapted Interaction*, 14:119–144, 2004.

D. L. McGuinness, R. Fikes, J. Rice, and S. Wilde. An Environment for Merging and Testing Large Ontologies. In A. G. Cohn, F. Giunchiglia, and B. Selman, editors, *Principles of Knowledge Representation and Reasoning.* Morgan Kaufmann, 2000.

D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. Technical report, World Wide Web Consortium Recommendation, February 2004.

K. McKeown, J. Robin, and K. Kukich. Generating Concise Natural Language Summaries. In M. Maybury and I. Mani, editors, *Automatic Text Summarization*. MIT Press, 1999.

J. F. Meech. Narrative Theories as Contextual Constraints for Agent Interaction. In *Narrative Intelligence Symposium*, AAAI Fall Symposium Series, pages 38–43. November 1999.

P. Mitra and G. Wiederhold. Resolving terminological heterogeneity in ontologies. In *Proceedings of the ECAI'02 Workshop on Ontologies and Semantic Interoperability*, Lyon, France, July 2002. European Conference on Artificial Intelligence.

V. O. Mittal and C. L. Paris. Context: Identifying its elements from the Communication point of view. In *Proceedings of the IJCAI-93 Workshop on Using Knowledge in its Context*, Chambry, France, August 1993.

K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002.

C. J. Needham and R. D. Boyle. Tracking Multiple Sports Players through Occlusion, Congestion and Scale. In T. Cootes and C. Taylor, editors, *British Machine Vision Conference*, volume 1, pages 93–102, 2001.

M. Nilsson and M. Sollenborn. Advancements and Trends in Medical Case-Based Reasoning: An Overview of Systems and System Development. In *Proceedings of the 17th International FLAIRS Conference. Special Track on Case-Based Reasoning*, pages 178–183, Miami, USA, May 2004. AAAI.

N. F. Noy and D. L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. Technical Report KSL-01-05, Stanford University, 2001.

N. F. Noy and M. Musen. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In *Proceedings of the 17th National Conference on Artificial Intelligence, AAAI '00*, Austin, Texas, 2000.

N. F. Noy and M. Musen. PROMPTDIFF: a fixed-point algorithm for comparing ontology versions. In *Proceedings of the 18th National Conference on Artificial Intelligence, AAAI '02*, pages 744–751, Edmonton, Canada, 2002.

D. O'Sullivan, B. Smyth, D. C. Wilson, K. McDonald, and A. Smeaton. Improving the quality of the personalized electronic program guide. *User Modeling and User-Adapted Interaction*, 14:5–36, 2004.

L. Portinale, P. Torasso, and P. Tavano. Speed-up, quality and competence in multi-modal reasoning. In *Proceedings of the Third International Conference on Case-Based Reasoning*, pages 553–564, Berlin, 1999. Springer Verlag.

V. Propp. *Morphology of the Folktalke.* University of Texas Press, 1968.

Protégé 2000. The Protégé Project, 2000. http://protege.stanford.edu.

L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

L. R. Rabiner and B. H. Juang. An Introduction to Hidden Markov Models. *IEEE Acoustics Speech and Signal Processing Magazine*, 3(1):4–16, January 1986.

D. R. Radev. W. Fan, and Z. Zhang. WebInEssence: A Personalized Web-Based Multi-Document Summarization and Recommendation System. In *Workshop on Automatic Summarization, North American Chapter of the Association for Computational Linguistics, NAACL'01*, Pittsburgh, PA, June 2001.

I. Reid and A. Zisserman. Goal-directed Video Metrology. In R. Cipolla and B. Buxton, editors, *Proceedings of the 4th European Conference on Computer Vision, LNCS 1065. Cambridge*, volume II, pages 647–658. Springer, April 1996.

C. K. Reisbeck and R. C. Schank. *Inside Case-Based Reasoning.* Lawrence Erlbaum Associates, Hilsdale, N.J. USA, 1989.

D. Riecken. Personalized views of Personalization. *Communications of the ACM*, 43 (8):27–28. 2000.

S. Rougegrez. Similarity Evaluation between Observed Behaviours for the Predicition of Processes. In *Topics in Case based Reasoning*, number 837 in Lecture Notes in Computer Science, pages 155–166. Springer Verlag, 1994.

D. Sadlier, N O'Connor, N. Murphy, and S. Marlow. A Framework for Event Detection in Field-Sports Video Broadcasts based on SVM generated Audio-Visual Feature Model. Case-Study: Soccer Video. In *Proceedings of the International Workshop on Systems, Signals and Image Processing*, Poznan, Poland, September 2004.

H. Saggion, H. Cunningham, K. Maynard, D. Bontcheva, O. Hamza, C. Ursu, and Y. Wilks. Extracting Information for Information Indexing of Multimedia Material. In *Proceedings of the Third Language Resources and Evaluation Conference*, Las Palmas, Spain, May 27 - June 2 2002.

R. Schank. *Dynamic memory: a theory of reminding and learning in computers and people.* Cambridge University Press, 1982.

R. Schank and R. Abelson. *Scripts, Plans, Goals and Understanding - An Inquiry into Human Knowledge Structures.* Lawrence Erlbaum Associates, 1977.

F Shook. *Television Field Production and Reporting, 3rd ed.* Longman, 2000.

P. Singh and B. Barry. Collecting Commonsense Experiences. In *Proceedings of the Third International Conference on Knowledge Capture, K-CAP 03*, pages 154 – 161, Sanibel Island, FL. October 2003.

P. Smagorinsky. The Reliability and Validity of Protocol Analysis. *Written Communication*, 6(4):463–477, 1989.

A. Smeaton, H. Lee, and N. E. O'Connor. TV News Story Segmentation, Personalisation and Recommendation. In *AAAI 2003 Spring Symposium on Intelligent Multimedia Knowledge Management*, Stanford University, Palo Alto, CA, 24-26 March 2003.

M. A. Smith and T. Kanade. Video Skimming for Quick Browsing based on Audio and Image Characterization. Technical Report CMU-CS-95-186, Carnegie Mellon University, July 1995.

R. A. Smith, A. W. Fitzgibbon, and A. Zisserman. Improving Augmented Reality using Image and Scene Constraints. In *Proceedings of the 10th British Machine Vision Conference*, pages 295–304, Nottingham, 1999.

B. Smyth and P. Cotter. A Personalized Television Listings Service. *Communications of the ACM*, 43(8):107–11, 2000.

B. Smyth and L. McGinty. The Power of Suggestion. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence, IJCAI'03*, pages 127–132, Acapulco, Mexico, August 9-15 2003.

B. Smyth and E. McKenna. Building Compact Competent Case-Bases. In *Proceedings of the Third International Conference on Case-Based Reasoning*, pages 329–342, Berlin, 1999. Springer Verlag.

P. Smyth. Clustering Sequences with Hidden Markov Models. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing 9*. MIT Press, 1997.

H. Sundaram and S. F. Chang. Condensing Computable Scenes using Visual Complexity and Film Syntax Analysis. In *IEEE International Conference on Multimedia and Expo, ICME'01*, Tokyo, Japan, 2001.

T. Trabasso and L. L. Sperry. Causal relatedness and importance of story events. *Journal of Memory and Langugage*, 24:595–611, 1985.

L. Truss. Wimbledon remain the True Princes at Palace. *The Times*, 24 January 1997.

C. J. van Rijsbergen. *Information Retrieval*. Butterworths, Woburn, Massachusetts, 1979.

L. Vanderwende, M. Banko, and A. Menezes. Event-centric summary generation. In *Proceedings of the Document Understanding Workshop, DUC 04*, Boston, May 6-7 2004.

I. Watson and F. Marir. Case-Based Reasoning: A Review. *The Knowledge Engineering Review*, 9(4):327–354, 1994.

D. Wettschereck and D. W. Aha. Weighting Features. In *Proceedings of the First International Conference on Case Based Reasoning Research and Development*, Lecture Notes in Artificial Intelligence, pages 347 – 358. Springer Verlag, 1995.

D. Wilson. *Case-base maintenance: The husbandry of experience.* PhD thesis, University of Indiana, 2001.

R. M. Young. Creating Interactive Narrative Structures: The Potential for AI Approaches. In *The Working Notes of the AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment*, Stanford, CA, March 2000.

S. Zelikovitz and H. Hirsch. Integrating Background Knowledge into Nearest Neighbour Text Classification. In S. Craw and A. Preece, editors, *European Conference on Case Based Reasoning*, number 2416 in Lecture Notes in Artificial Intelligence, pages 1–5. Springer Verlag, 2002.

L. Zelnik-Manor and M. Irani. Event Based Analysis of Video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 123–130, December 2001.

J. Zhu and Q. Yang. Remembering to add: Competence-preserving case-addition polices for case base maintenance. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI '99*, pages 234–241, Stockholm, Sweeden, 1999.

Appendix A

Knowledge elicitation study

A.1 Knowledge Elicitation interview transcript

Interview conducted 24/4/02, 2pm with a New Media Development Producer, BBC Sport

What training do you have?

If someone's going to be an AP[1] or a Producer, they've got to have training in how to construct a package, like how to make something when all the feeds come in live, how to editorialise that and make it into a package. Looking for the relevant points, how to construct it, how to use different pictures to illustrate a story that you might not have pictures for. They might have some basic editing skills as well, they might have been taught how to use an Avid, or a Media Composer machine. And they would also have - a lot of the stuff we make here is specially shot material so you go out with a camera and shoot your own bits and pieces and you'd be given training on how to structure that so you went out knowing what you wanted and came back with the relevant shots to be able to construct the story that you're trying to do.

Could you talk me through the making of football highlights?

The game will come in in its entirety on a live feed into Television Centre and it will be cut into a highlights package from here or it could be out on location. Now on location, they've got all the different cameras around the ground. Obviously they've got the director sitting in the back of an OB[2] truck deciding which camera he's going to take and he's probably got a good idea of which camera he's going to go to next, depending on where the ball's moving around the pitch. Within that truck they've got people working on Slow Mo and Replay so they're taking the game into a machine called an LSM. It can store and edit at the same time, so it's still recording and you can spool back within it and clip up the section that you need. That's how they would provide an instant replay of a goal line incident or a goal and the Slow Mo machine would do the same. It can store it while it's playing out.

Is there a standard number of cameras?

It depends on who's covering the ground, which ground it is, how many cameras

[1] Assistant Producer

[2] Outside Broadcast

you're allowed in. Sky would probably have more cameras than we would have, but it would depend on the importance of the match as well. For the FA Cup Final you're going to have more cameras there than for a third round game.

What sort of numbers? 2? 6?

I actually don't know. It depends. There would be some static cameras that don't move, and there would be some that are operated. Probably about 12.

So there's the director in the OB truck and you guys back here. What's the difference between what you're doing?

If there's people at the live game, then it's actually going out live on air. And if it's a highlights package, a highlights programme, then you're just taking that live feed in from somewhere and you could be taking it in from another broadcaster and you're just going to turn that round into a twenty minute highlights of a ninety minute game. So in that instance you're looking for the talking points of the game, the goals, the big fouls, the controversial incidents. And then you're looking to use the other camera angles of the same incident. So say somebody's been tackled, then you'd want to see the reverse angle, or the side angle, whatever angle it is that gives you the better view of what actually happened and similarly, in an isolated shot where the camera's just been trained on one person rather than a wide shot. Or in a Slow Mo or a Super Slow Mo where the action's been slowed down so you can really analyse what's going on.

Does it make a difference if you have to get a highlights out in five minutes for the next programme?

Yes. The highlights that you'll see in a half past ten programme when the games have been in the afternoon, you've got a lot longer to edit them, so you can be finer with them. When we're doing the interactive service, the match is going on at the same time, we've got highlights running behind it, so if you've missed the first ten minutes of the match, you can click highlights and watch a minute's worth behind that ten minutes, so you can catch up on the exciting minute of the game. What makes that really hard is if you're taking in a game, you've often got the clean effects feed which is the match atmosphere, the actual sound within the ground, you've got a feed with the commentators on it and the summariser, and often they're not talking to pictures, they're talking to the flow of the game, so you could well have an incident where somebody scores, and the summariser's going blahdyblah and you want to make an edit on the pictures because you've come to a shot change, but you can't because the person's talked right the way over the top of it, so that's often where the difficulty lies. And if you had a lot longer to do that, you could mess around with the sound and pull things down and fade them up, whereas if you're doing it quickly you have to think, "Right, we'll just use this: normal commentary for the first goal and just clean effects for the replay, because someone's talked all over it and we can't chop and change about."

What criteria are you given to decide what goes into the highlights? Is it time?

We don't have criteria. I mean, I've never been given any criteria. From an interactive point of view, when we're on air, on a ninety minute game, our highlights aren't allowed to be more than five minutes long. Because otherwise that takes people away

from the game. Whereas if you're doing a ninety minute game and it's just a standalone game, depending on the quality of the game, you'd be looking at somewhere between ten and twenty minutes. I would say, in football, you'd have to put all the goals in, but in rugby, you wouldn't put all the kicks in, because a kick is pretty much the same each time. And if the goals have come from static play, you know, a set piece is done, they're quite quick, whereas if a goal has come from a big movement, where there's been a lot of build up from a team, something like Manchester United play particularly well, then you'd want to demonstrate more of that. So it depends on the kind of game it is, and I think it comes down to knowing it. I know that's not much help, but when you watch it you can actually tell. Some incidents, where you'd expect people to score, like glaring misses, where they really should have scored, they would go in as well. Controversial decisions, funny bits like dogs running on the pitch, all that kind of stuff. You want to create some kind of atmosphere. But also create a piece which reflects the game as well. So say Man United have won 5-Nil, but the other team haven't played all that badly, you wouldn't just want to show five goals. You'd have to try and reflect the balance of the game, say they were playing a team that were really low down the division, but they actually did themselves justice and they defended really well. It could have been 15-Nil, but they defended really well so it's only 5-Nil, so you'd have to try and reflect that.

Do you write the commentary as well?

If the highlights have been cut live, then often not, but if you're doing a Match of the Day or a highlights program then you would write a voice-over for the presenter and he would then voice up over the clips that you had chosen. So you could decide how long the clip wanted to be, or you could tailor what you were writing to fit the length of the clip and the amount of replays. That takes a lot longer than just cutting it with the actual sound on it because you need to rehearse it and you need to make sure the emphasis is right on the script and that the words do really match the pictures because you're actually cutting them as a sequence rather than just bolting them together.

How long would that take?

Again, it would depend on how quickly you wanted it. You could probably turn it round in about an hour, an hour and a half, but you could spend a lot longer on it. I mean I've done tennis matches and golf rounds and rugby games as we've gone and we've done that as the game's been happening. So we've voiced all the way through a set and had it on air before the next set started in a tennis game. So that's a minute and a half turn around. But we've voiced all the clips as the tennis game was going on. We've decided that one game was... so someone was 3-3, then it was 4-3 so we've voiced lots of different clips, then dropped the ones we didn't need so it still kept a story. That was using a really clever piece of equipment called an AirBox which allows you to loop material as a series of separate clips which are joined together but you can remove clips from within that loop, and still keep the loop on air. It's a clever piece of equipment. It's made by a Belgian company. We use that all the time on the interactive service.

What does the edited material get used for? Apart from sports programmes or the news. Quiz shows?

We walked past the library, where every tape is catalogued and logged, like referenced. They deal with sales from everywhere. Overseas, in the country, lots of other broadcasters in the UK and Europe. If we've filmed it and own the rights to it, we would then sell on.

Would you store the whole of the match? Or would you store highlights?

Yes, the whole of the match, and the highlights. Because you may well find, in a year's time that someone wants to make a documentary about one player. And what they need is just some loose shots of them running around the pitch or the field. They wouldn't necessarily be in the highlights but they would definitely be stored within the game.

What metadata do you have with the game

None.

So there's no tagging of it?

Nothing.

So you've got the title... is it just written down on it? Every programme has a programme number, which is allocated within the BBC and that's how we record and track things. And there is a tape number on the tape as well, so we cross-reference the tape number with the programme number. But that's the only way. We've looked at metadata solutions, but because our library is so extensive and enormous it would be - it wouldn't be impossible, but it would be almost unthinkable. If you think about implementing a system when we're still adding so much... we realise we've got to do something, but trying to work out quite what is a bit hard.

Is it all on analogue tape or is it digitised?

No, it's on a variety of different formats, which is another problem. Some of it's on old one inch tape, the actual reel-to-reel stuff, some of it's on pre-digital format and for about the last three years it's been on digital format.

So on the tape you've just got the programme number and the tape number? There's no - who's in the match or...?

In the library, when you know the tape number, there's a series of paper logs and you can go and check those, and at the relevant timecode it goes and tells you what it is, and that's all on a computer database as well, so you can search that. So you've got it on paper and on a computer database. But things change all the time.

So whoever writes in the description of the programme just puts in what they want, there's no standard?

No there's no... the standard shorthand, if you're logging a game: say you're sitting and there's a game coming in, you've got a log sheet in front of you which is set out in a certain way. You write your name and the date on it, so if someone's got a query with it, they come and talk to you. And there's also shorthand, for different kinds of shot, like a Mid Shot, a Long Shot, a Wide Shot, a Close Up, a Big Close Up, a Medium Close Up. They're all abbreviated: MCU, BCU that everybody would know if you were to use them. Or if you were logging a tennis game, if a shot's from the right hand side or the left hand side or a volley or a half volley. There's all that shorthand that people know. Or a slow Mo, a Super Slow Mo, an ISO. And you would put down as many of the names of the players as you could, and the presenters

and the summarisers, things like that.

But it's just up to the person that's writing it?

Yes

So when people are watching TV highlights, what sort of things do you think they are looking for? What do they want to see? What makes a "good" highlights programme?

A good highlights programme is content-led. If you've got a rubbish game, if you've got a goalless draw, it's really hard to make an interesting set of highlights out of that. If you watch any highlights game, from something that's on Sky, something that's on ITV. People really like to watch goals in a highlights programme. You could say, "For Match of the Day this week, we're just going to show you all the goals from Saturday" and a lot of people would be happy with that. Because that's all they really want to see. But then people who are more fans might say: "I actually want to see *my* team's highlights, I'm not too bothered about everybody else's, but I do want to see everybody else's goals". And that's something that's been very popular when it's been done as a Goal Round Up on the FA Cup or the Interactive Service. It's something that people will use and do like. I would say that, and Magic Moments. When we were doing the highlights last year of golf, this guy got a hole-in-one and we put that on a highlights loop, and people crowded round the loop, because they wanted to see the hole-in-one. So it's talking points. So say like if a big player gets injured, like Beckham's injury, people will be watching that rather than the goals that they scored that night, I'm sure. That tackle.

There was one match where there was seven goals and they didn't show them all in the highlights. So are there better goals and worse goals?

I'm quite surprised they didn't show all of them. I'd have thought they'd put them all in. There are Wonder Goals like Michael Owen against Argentina, Ryan Giggs against Arsenal. Like this season you had Zola in the FA Cup against Norwich. That was a brilliant goal. There are goals that really stand out. People will see those again and again, but all goals are important at the end of the day in football.

How do you know when to start showing the cut?

That just comes from watching the flow of the play. Each goal will start with a move somewhere. Often it starts from the other end of the pitch or it will start from a pass in the centre of midfield where someone does something that is a little bit out of the ordinary. But then if you take it and you can sometimes watch it and think, "Well actually no, that feels a bit short. It happens too quickly." Then you can take it back a bit further. And it's knowing when to come out. How many times do you want to see him pulling his vest up to the crowd? All that kind of stuff. It's - how long does it feel right? It is very kind of... there's a certain kind of personal judgement to all of it, I think.

Can you show me some of these? I'd also like to see what's a not-so-good edit.

I can show you some downstairs, no problem.

Does the BBC have plans to offer more personalised content?

Not that I'm aware of at the moment.

Is it alright if I mention the name of the BBC when I write up my report?

Sure, yes. Could we have a look at it as well?
Yes, and is it possible for you to let me use any example footage? Or not? Is it all copyright?
I'll have to check. I imagine it's all copyright. But if it's for research purposes you might be able to get around it, if you were to approach the governing body rather than us. I'll give you some contact details.

Is there any internal documentation I might be able to get a copy of? Like your edit log?
I can show you my log sheets.
...
[Explanation about PhD research]
If you've got the football pitch and the halfway line and once it's gone across into that area *(indicating line parallel to penalty area line.)* If you could calibrate it so that once the ball has crossed those lines, that anything there was automatically clipped up. If the ball was moving in that direction *(indicates towards the goal.)* So that once the ball had crossed into the penalty area, then it might be worth clipping that up. Invariably, that is where most of the action takes place. And then if, I don't know if you could manually override it, if you needed to start when the ball was back there, then you could go back and get that extra little bit on the back of it. But I would say that once the ball's in that area, that's really important.

What about the sound? Obviously there's cheers when there's been a goal, but what about before? Is there a volume difference?
There often is a kind of build up towards a goal. Maybe you could do it on the commentators. Certainly, if you listen to any South American commentators, you can see them getting really excited when someone's coming towards a goal, and they get faster and louder and when someone scores they really go for it. You could work on a model like that. You would find the same here, they do get more animated when it looks like someone - "Oh my word!" that kind of stuff.
Interview ends

A.2 Protocol analysis transcript

Kick off was at 13:00, however the highlights loop began a few minutes before this.

12:58:37 Flags (i.e. team crests) shown for 25 seconds.

12:59:02 Team line up.

12:59:06:17 to 13:00:43:14 A separate, longer sequence was also clipped up, which showed the team line up followed by kick off.The editors had difficulty with this clip, they couldn't edit it any shorter as the commentators "burbled on", despite having instructions to pause between comments, to ease the task of the highlights editors. They decided, "Not much we can do, just take the whole lot. We'll get rid of it soon anyway." This was then two minutes of footage repeated in a loop.

13:08 Manchester United scored for the first time. There was a buzz of activity throughout the editing suite. This was clearly the most important event type for all concerned. Clip 61 was tagged with the label "MU 7 Goal 1" (7 being the shirt number of David Beckham, although Ryan Giggs, number 11 actually scored the goal - this was either a mistake on their part, in the limited time available, several labels were misspelled, or referred to David Beckham, who lined up the shot for Giggs), and ran for 30:12 seconds. The most difficult part of the task was dealing with the audio commentary, the choice of exact edit point had to be on where the commentator paused in a sentence. The slow motion replay of the goal was also included in the highlights (although shortened from the original live replay.) There were two additional shots showing different angles of Giggs' goal. Then the editors added in three seconds of the First Half title graphic into the loop, and removed the team line up sequence.

13:17 A Beckham shot on goal was deemed "not that great" and not included. The decision was taken three minutes after it happened, as prior to that they were still busy editing the goal.

13:24 Beckham's shot on goal, which hit the top bar was clipped up for inclusion. Between Giggs' goal and Beckham's attempt, there was a close-up on Alex Ferguson (Manchester United's manager). One editor advised the other, "We don't want to put Ferguson in between those two [shots] and then find you need it somewhere else". That is, continuity on close-ups need not be preserved, as they can be useful for smoothing transitions between shots and events.

13:26 A booking is not included in the highlights.

13:28 An attempt to shorten Beckham's shot on goal from 20 seconds to 12 makes it "too tight - I don't like it", that is, the shot change is too jumpy from one to another.

13:30 Giggs' second goal. The editors use clean effects for the audio instead of commentary. "There's not a lot we can do to clean it up. Maybe work on it at half time."

13:34 "Quite a good replay that's just gone through [of the second goal]". They clip it up, but the clean effects on this one are different (less cheering) as it is several minutes after the goal was actually scored. "Not sure the replay works... such a big difference in atmosphere". But this second replay is also included. One editor asks the other whether to include a Beckham shot described by the commentator as "a scorcher" or a Scholes shot which hits the post. The other editor instructs him to include both.

13:38 The highlights loop so far: First Half graphic; Manchester Utd Goal 1; Manchester Utd Goal 1 slo mo; Giggs goal (from side angle); Giggs goal (from reverse angle); Beckham shot; Manchester Utd Goal 2; Manchester Utd Goal 2 slo mo: Beckham skilled shot; Scholes shot.

13:44 Lee Bowyer (West Ham) has a shot on goal. The editor states, "That's livened things up a bit." This shot is included, because it's the only West Ham attempt on goal in the first half. The clip duration is 29 seconds, labelled as: "Whattack 1b".

13:49 Half time There are there is 2 minutes 27 seconds on the highlights loop so far. "We can always tell when we're doing something right because the analysis at half time is the same as our highlights" That is, the editors know almost as well as a professional footballer (who presents) which are the important events.

14:06 Second half A controversial foul is not declared a foul, even though one editor exclaims "How can that *not* be a foul?" It is not included in the highlights.

14:07 Manchester United Goal 3.

14:09 Manchester United Goal 4. The second half graphics are included. and then the two goals and their slow motion replays. The goals are 15-20 seconds long and the replays 6 or 7 seconds. The highlights loop now contains: First Half graphic; Manchester Utd Goal 1; Manchester Utd Goal 1 slo mo; Giggs goal (from side angle); Giggs goal (from reverse angle); Beckham shot; Manchester Utd Goal 2; Manchester Utd Goal 2 slo mo; Beckham skilled shot; Scholes shot; West Ham Shot; Second half graphic; MU Goal 3; MU Goal 3 slo mo; MU Goal 4; MU Goal 4 slo mo.

14:15 They get rid of the Beckham shot as it's too long: "It's like War and Peace." Giggs' first goal's additional two angles are removed. They reclip the fourth goal slo mo, to make it shorter, and discuss whether to reduce the West Ham shot on goal.

14:17 Clip up the fifth goal, from the pass back and kick out, and a slow-motion replay. Commentary is, "West Ham are being taken to the cleaners."

14:23 Cut the fourth goal shorter. because the highlights package is getting too long.

14:27 The sixth goal and a slow-motion replay are clipped up. Although the package is now 4 minutes, this is justified for a six goal game, "We could even go to 4:30" but they ask, "Is there any more on the front of the Neville goal [that we could cut]?" The verdict is that, "It's a good watch. People should come away from the main match and come straight to this."

14:40 A second West Ham shot on goal is included for 13 seconds.

14:50 They clip up the end of the match, including the final whistle. This is edited more carefully as there is now more time. Some action is added in prior to the whistle, so that it becomes longer than the 6 frames which was considered "messy". After the loop plays for the final time, the highlights end on the original team crests.

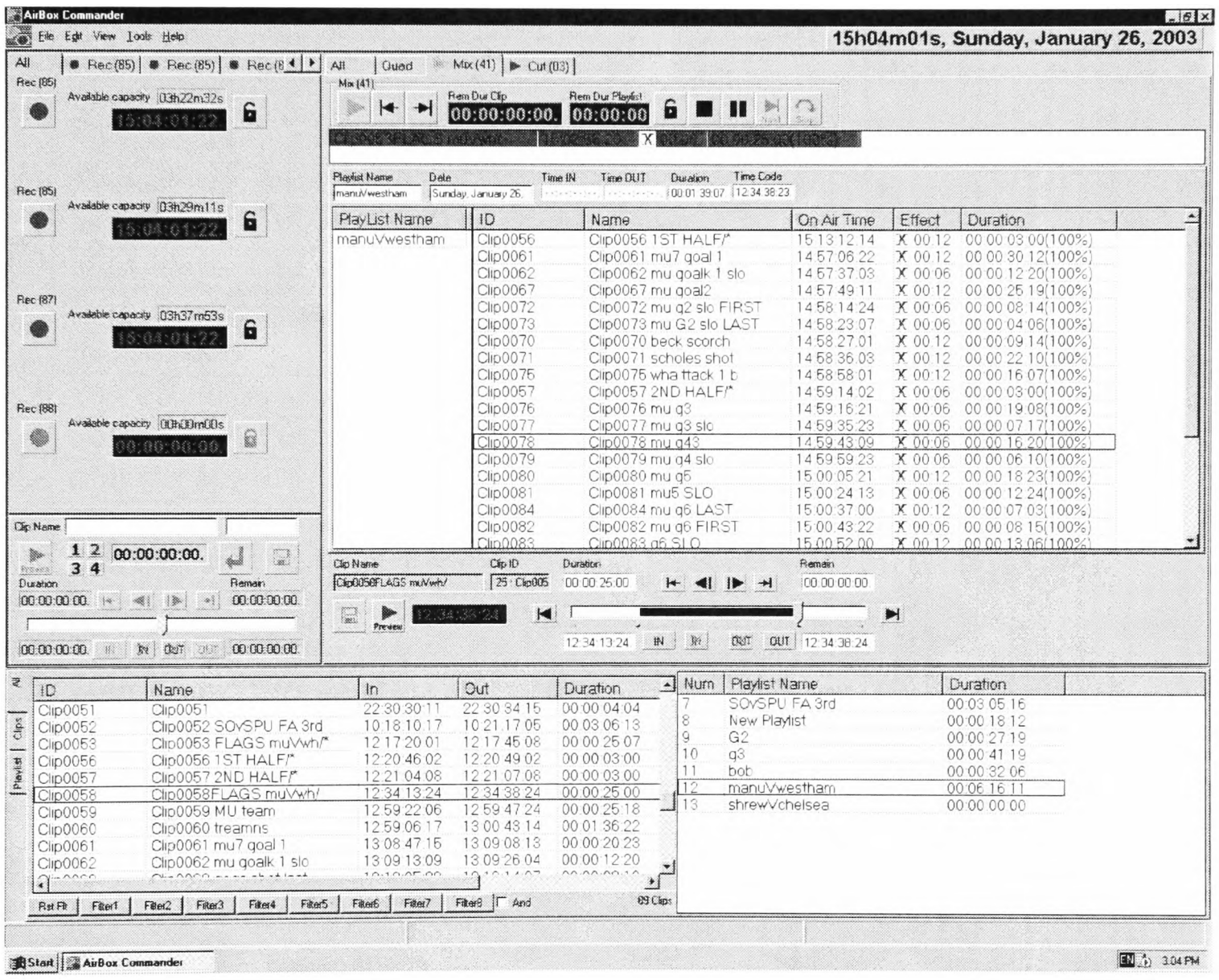


Figure A.1: A screen shot from the software used to edit football highlights

Shot Description	Length (in seconds:frames)
First Half Flags	3:00
Manchester Utd Goal 1	30:12
Manchester Utd Goal 1 Slow Motion	12:20
Manchester Utd Goal 2	25:19
Manchester Utd Goal 2 Slow Motion First	8:14
Manchester Utd Goal 2 Slow Motion Last	4:06
Beckham Scorcher Shot	9:14
Scholes Shot	22:10
West Ham Shot on Goal 1	16:07
Second Half Flags	3:00
Manchester Utd Goal 3	19:08
Manchester Utd Goal 3 Slow Motion	7:17
Manchester Utd Goal 4	16:20
Manchester Utd Goal 4 Slow Motion	6:10
Manchester Utd Goal 5	18:23
Manchester Utd Goal 5 Slow Motion	12:24
Manchester Utd Goal 6 Angle 1	7:03
Manchester Utd Goal 6 Angle 2	8:15
Manchester Utd Goal 6 Slow Motion	13:06
West Ham Shot on Goal 2	17:00
Final midfield action and Whistle	6:09
Start Flags	25:00

Table A.1: Final Edit Decision Log

Appendix B

Markov modelling details

B.1 Hidden Markov Model algorithms

B.1.1 Forward-backward algorithm

To compute $P(O|\lambda)$ directly, given a sequence $O = O_1, O_2, ...O_T$ and an N state Hidden Markov Model λ, would involve $2TN^T$ calculations [Rabiner and Juang (1986)]. The forward-backward algorithm requires many fewer calculations (only N^2T) by taking advantage of an iterative procedure where either a forward variable, $\alpha_t(i)$ or a backward variable, $\beta_t(i)$ is defined:

$$\alpha_t(i) = P(O_1, O_2, ...O_t, i_t = q_i|\lambda) \quad \text{(B.1)}$$

$$\beta_t(i) = P((O_{t+1}, O_{t+2}, ...O_T|i_t = q_i, \lambda) \quad \text{(B.2)}$$

That is, the forward variable is the probability of the partial observation sequence up to time t and state q_i at time t, given the HMM λ, while the backward variable is the probability of the partial observation from time t until the end of the sequence (at time T), given state q_i at time t and the HMM λ.

To calculate $P(O|\lambda)$ using the forward variable, α_1, the joint probability of each state and the initial observation O_1, is defined:

$$\alpha_1(i) = \pi_i b_i(O_1), \text{ for } 1 \leq i \leq N \quad \text{(B.3)}$$

Next, consider what happens at time t+1: there is a transition from one of the N possible states q_i, to state q_j, with the probability a_{ij}. The product $\alpha_t(i)a_{ij}$ is then the joint probability that the sequence $O_1, O_2, ...O_t$ is observed *and* state q_j is reached at time t+1 via state q_i at time t. If these products are summed over all N possible states, this results in the probability of q_j at time t+1, with all the accompanying previous partial observations (due to the Markovian property). Now in state q_j at time t+1, multiplying by the observations seen in this state, which are produced with probability $b_j(O_{t+1})$, gives:

$$\alpha_{t+1}(j) = P(O_1, O_2, ...O_{t+1}, j_{t+1} = q_j|\lambda) \quad \text{(B.4)}$$

$$= [\sum_{i=1}^{N} \alpha_t(i)a_{ij}]b_j(O_{t+1}) \quad \text{(B.5)}$$

From the definition of the forward variable $\alpha_t(i)$ in equation B.1, it can be seen that $P(O|\lambda)$ at time T, for all states *i=1 to N*, is simply the summation of the terminal forward variables.

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{B.6}$$

So using equation B.5, each forward variable can be iteratively calculated at each time interval, beginning with $\alpha_1(i)$ (equation B.3), until the end of the sequence is reached. In reverse, for the backward variable $\beta_t(i)$, the equations are:

$$\beta_T(i) = 1, \text{ for } 1 \leq i \leq N \tag{B.7}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}(j)), \tag{B.8}$$

$$\text{for t = T-1, T-2....1, and } 1 \leq i \leq N.$$

B.1.2 Viterbi algorithm

The Viterbi algorithm finds the single most probable state sequence, $I = i_1, i_2, ...i_T$, given a certain observation sequence, $O = O_1, O_2, ...O_T$, from a Hidden Markov Model λ. It is similar to the forward-backward algorithm, except that it uses a maximisation over previous states, rather than a summation. Firstly, the best score (highest probability) along a single path at time t, $\delta_t(i) = maxP[i_1, i_2, ...i_t = q_i, O_1, O_2, ...O_t|\lambda]$ is defined, which accounts for the first t observations and ends in state q_i. The steps of the Viterbi algorithm, as described in Rabiner and Juang (1986) are as follows:

1. Initialisation:

$$\delta_1(i) = \pi_i b_i(O_1), \text{ for } 1 \leq i \leq N. \tag{B.9}$$

$$\Psi_1(i) = 0 \tag{B.10}$$

   where the array $\Psi_t(j)$ keeps track of the argument which maximises $\delta_t(i)a_{ij}$ for each t and j.

2. Recursion:

$$\delta_t(j) = max_{1 \leq i \leq N}[\delta_{t-1}(i)a_{ij}]b_i(O_t) \tag{B.11}$$

$$\Psi_t(j) = argmax_{1 \leq i \leq N}[\delta_{t-1}(i)a_{ij}] \tag{B.12}$$

$$\text{for } 2 \leq t \leq T, 1 \leq j \leq N$$

3. Termination:

$$P^* = max_{1 \leq i \leq N}[\delta_T(i)] \tag{B.13}$$

$$i_T^* = argmax_{1 \leq i \leq N}[\delta_T(i)] \tag{B.14}$$

4. Path (state sequence) backtracking:

$$i_t^* = \Psi_{t+1}(i_{t+1}^*) \tag{B.15}$$

$$\text{for t = T-1, T-2.....1.}$$

B.1.3 Baum-Welch algorithm

Finding the most likely Hidden Markov model (in terms of its parameters $\lambda = (A, B, \pi)$) for a particular observation sequence $O = O_1, O_2, ...O_T$ analytically is an NP-complete problem. However, $P(O|\lambda)$ can be locally maximised using an iterative procedure such as the Baum-Welch algorithm, which refines the estimate of λ at each iteration by increasing the probability of O given λ. The Baum-Welch re-estimation formulae, using the notation in Rabiner and Juang (1986), for the HMM parameters π, A and B are:

$$\bar{\pi}_i = \gamma_1(i), \text{ for } 1 \leq i \leq N \tag{B.16}$$

$$\bar{a_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{B.17}$$

$$\bar{b_i}(k) = \frac{\sum_{t=1, O_t=k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{B.18}$$

where :

$$\gamma_t(i) = P(i_t = q_i)|O_t, \lambda) \tag{B.19}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} \tag{B.20}$$

and:

$$\xi_t(i,j) = P(i_t = q_i, i_{t+1} = q_j)|O, \lambda) \tag{B.21}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(O_{t-1})\beta_{t+1}(j)}{P(O|\lambda)} \tag{B.22}$$

At each iteration, $\bar{\lambda} = (\bar{\prod}, \bar{A}, \bar{B})$ is used to calculate $P(O|\bar{\lambda})$, and then $\bar{pi}$, $\bar{a_{ij}}$ and $\bar{b_i}(k)$ can be re-estimated.

Ticker Tape Groups	**Automatically Learnt Context Groups**
0:13 **Goal kick** taken long by Jussi Jaaskelainen (Bolton).	**Goal kick** *start time*: 0:13 *extra time*: 0:00 *Player*: Jussi Jaaskelainen *taken*: long *duration*: 41 **Throw in** *start time* 0:54 *extra time*: 0:00 *Player*: Jay-Jay Okocha *type*: Attacking *duration*: 25
0:54 Attacking throw-in by Jay-Jay Okocha (Bolton).	
1:19 **Corner** from right by-line taken short right-footed by Stelios Giannakopoulos (Bolton).	**Corner** *start time*: 1:19 *extra time*: 0:00 *Player*: Stelios Giannakopoulos *taken*: short right-footed *from:* right by-line *duration*: 18
1:37 **Header** by Kevin Davies (Bolton) from centre of penalty area (6 yards), over the bar. **Goal kick** taken long by Nigel Martyn (Everton).	**Header** *start time*: 1:37 *extra time*: 0:00 *Player*: Kevin Davies *from*: centre of penalty area (6 yards) *resulting in*: over the bar *duration*: 49
	Goal kick *start time*: 1:37 *extra time:* 0:00 *Player*: Nigel Martyn *taken*: long *duration*: 49

Table B.1: An example of context groups clustered using a Markov chain

Ticker Tape Groups	Automatically Learnt Context Groups
2:26 **Foul** by Joseph Yobo (Everton) on Kevin Davies (Bolton). **Free kick** taken left-footed by Simon Charlton (Bolton) from left wing, resulting in open play.	**Foul** *start time:* 2:26 *extra time:* 0:00 *Player:* Joseph Yobo *Team:* Everton *on:* Kevin Davies *duration:* 22 **Free kick** *start time:* 2:26 *extra time:* 0:00 Player: Simon Charlton *taken:* left-footed *from:* left wing *resulting in:* open play *duration:* 22
2:48 **Shot** by Per Frandsen (Bolton) left-footed from centre of penalty area (18 yards), (caught) by Nigel Martyn (Everton).	**Shot** *start time*: 2:48 **save** *extra time:* 0:00 *Player*: Per Frandsen *taken*: left-footed *from*: centre of penalty area (18 yards) *duration*: 24 **Save** *start time*: 2:48 *extra time*: 0:00 *Player*: Nigel Martyn *type*: caught *duration*: 24
3:12 Defending **throw-in** by Ricardo Gardner (Bolton).	**Throw in** *start time*: 3:12 *extra time*: 0:00 *Player*: Ricardo Gardner *type*: Defending *duration*: 27 **Throw in** *start time:* 3:39 *extra time*: 0:00 *Player*: David Unsworth *type*: Defending *duration*: 35
3:39 Defending **throw-in** by David Unsworth (Everton).	

Table B.1: An example of context groups clustered using a Markov chain (*Continued*)

Previous event

Marginals	CurrentEvent:	Assist	Backpass	Block	Booking	Clearance	Corner	Cross	Foul	Free kick	Goal	Goal kick	Handball	Header	Offside	Penalty	Save	Sending off	Shot	Substitution	Throw in
0.003895477	**Assist**	0	0	0	0.007426	0	0.00223	0.002758	0.009703	0	0.014164	0.00285948	0.011905	0.002151	0.00894	0	0	0.066666667	0.006202	0.007092199	0.004577
3.79E-04	**Backpass**	0	0	0	0	0	7.43E-04	3.45E-04	3.13E-04	0	0.002833	4.08E-04	0	0	0.002554	0	0	0	0	0	8.17E-04
0.009204357	**Block**	0	0	0	0.002475	0	0.051301	0.011375	0.008764	0	0.002833	0.00776144	0.005952	0	0.006386	0	0	0	0.015504	0.007092199	0.011934
0.013927192	**Booking**	0	0	0	0.029703	0	0.005204	0.003102	0.012207	0.0589354	0.008499	0.00694444	0.005952	0	0	0.09375	0	0.066666667	0.004651	0.035460993	0.005068
0.07735797	**Clearance**	0	0.181818182	0	0.012376	4.46E-04	0.265428	0.139262	0.092019	0	0.053824	0.03676471	0.095238	0.030108	0.099617	0	6.94E-04	0	0.146253	0.05141844	0.10528
0.04636652	**Corner**	0	0	0	0.009901	0.27228164	0.021561	0.031368	0.010642	0.0097433	0.076487	0.02287582	0.02381	0.275269	0.016603	0.03125	0.090972	0	0.045995	0.015957447	0.01177
0.100006895	**Cross**	0	0	0.322097	0.009901	0.56461676	0.052788	0.028956	0.019092	0.0047529	0.184136	0.09844771	0.071429	0.417204	0.033206	0.125	0.274306	0	0.098708	0.04787234	0.025666
0.11014203	**Foul**	0	0	0	0.641089	0	0.002974	0	0.002817	0.6770437	0.022663	0	0	0	0.002554	0.53125	0	0.4	5.17E-04	0.10106383	0
0.145063431	**Free kick**	0	0	0	0.185644	0.13235294	0.115242	0.182006	0.210642	0.0052281	0.195467	0.13357843	0.178571	0.150538	0.217114	0	0.097917	0.133333333	0.202584	0.111702128	0.190943
0.012169057	**Goal**	1	0	0	0.014851	0	0.005204	0.036539	0.007512	2.38E-04	0.005666	0.00245098	0.017857	0.03871	0.011494	0	0	0.066666667	0.003101	0.019503546	0.006703
0.084390513	**Goal kick**	0	0.363636364	0.003745	0.014851	0	0.055762	0.100655	0.174022	7.13E-04	0.067989	0.04044118	0.232143	0.008602	0.154534	0	0	0.066666667	0.094574	0.086879433	0.159555
0.005791506	**Handball**	0	0	0	0.009901	0	0	0	0	0.0375475	0	0	0	0	0	0.15625	0	0.066666667	0	0.003546099	0
0.016030061	**Header**	0	0	0	0	0.00757576	0.005204	0.001724	0.002817	0	0.005666	0.10130719	0.005952	0.004301	0.005109	0	0.096528	0	0.002584	0.017730496	0.002125
0.026992554	**Offside**	0	0	0	0.00495	0	0	0	0	0.1817966	0.002833	0	0	0	0	0	0	0	0	0.028368794	0
0.001103144	**Penalty**	0	0	0	0	0	0	0	0	0	0.076487	0	0	0	0	0	0.003472	0	0	0	0
0.049641478	**Save**	0	0	0	0.007426	0	0.158364	0.064461	0.071049	2.38E-04	0.082153	0.0253268	0.035714	0.004301	0.090677	0	0	0	0.074419	0.04787234	0.07471
5.17E-04	**Sending off**	0	0	0	0.002475	0	0	0	6.26E-04	0.0011882	0	4.08E-04	0	0	0	0.03125	0	0	0	0.007092199	1.63E-04
0.066705736	**Shot**	0	0	0.674157	0.009901	0.02272727	0.064684	0.00586	0.008764	0	0.031161	0.33333333	0.005952	0.008602	0.012771	0	0.436111	0	0.011886	0.070921986	0.00801
0.019442912	**Substitution**	0	0	0	0.009901	0	0.012639	0.009307	0.013772	0.0206749	0.011331	0.03267974	0.005952	0.006452	0.014049	0.03125	0	0	0.010853	0.147163121	0.029263
0.210872863	**Throw in**	0	0.454545455	0	0.027228	0	0.180669	0.382282	0.355243	0.0019011	0.155807	0.15441176	0.303571	0.053763	0.324393	0	0	0.133333333	0.282171	0.193262411	0.363413

Conditional Probability Matrix: P(Current Event Occurred | Previous Event Occurred)

Previous event

Marginals	CurrentEvent	Assist	Backpass	Block	Booking	Clearance	Corner	Cross	Foul	Free kick	Goal	Goal kick	Handball	Header	Offside	Penalty	Save	Sending off	Shot	Substitution	Throw in
0.00392995	**Assist**	0	0	0	0.069767	0	0.054217	0.038462	0.079096	0	0.085366	0	0	0.018692	0	0	0	0.071428571	0.068681	0.068965517	0
3.45E-05	**Backpass**	0	0	0	0	0	0	0	0	0	0.003049	0	0	0	0	0	0	0	0	0	0
3.79E-04	**Block**	0	0	0	0	0	0.006024	0.006993	0.016949	0	0.003049	0	0	0.009346	0	0	0	0	0.005495	0	0
0.00148235	**Booking**	0	0	0	0.023256	0	0.006024	0.003497	0.033898	0.1179775	0.009146	0	0	0	0	0.1	0	0.071428571	0.008242	0.068965517	0
0.004929675	**Clearance**	0	0	0.090909	0.046512	0	0.198795	0.083916	0.084746	0	0.060976	0	0	0	0	0	0	0	0.065934	0.103448276	0.241379
0.005722559	**Corner**	0	0	0	0	0.43356643	0.024096	0.034965	0.039548	0.011236	0.07622	0.01503759	0.083333	0.233645	0	0.033333	0.056604	0	0.046703	0.034482759	0.034483
0.009859349	**Cross**	0	0	0	0.023256	0.24475524	0.054217	0.087413	0.084746	0.005618	0.240854	0.03007519	0.333333	0.411215	0.153846	0.1	0.056604	0	0.167582	0	0.103448
0.006101765	**Foul**	0	0	0	0.581395	0	0.006024	0	0.050847	0.7078652	0.02439	0	0	0	0	0.533333	0	0.428571429	0.002747	0.103448276	0
0.006136238	**Free kick**	0	0	0	0.139535	0.18881119	0.024096	0.052448	0.112994	0	0.103659	0.03759398	0.083333	0.065421	0	0	0.079245	0.142857143	0.063187	0.172413793	0.103448
0.011307226	**Goal**	1	0	0	0.023256	0	0.03012	0.367133	0.107345	0	0.07622	0	0.166667	0.186916	0.153846	0	0	0.142857143	0.076923	0.034482759	0.034483
0.004584942	**Goal kick**	0	0	0	0.046512	0	0.036145	0.066434	0.112994	0	0.042683	0.01503759	0.083333	0.037383	0.307692	0	0	0	0.134615	0	0.034483
4.14E-04	**Handball**	0	0	0	0.023256	0	0	0	0	0.0393258	0	0	0	0	0	0.166667	0	0.071428571	0	0	0
0.003688638	**Header**	0	1	0	0	0.04895105	0.006024	0.031469	0.022599	0	0.030488	0.22556391	0.083333	0.018692	0	0	0.113208	0	0.024725	0	0.034483
4.48E-04	**Offside**	0	0	0	0	0	0	0	0	0.0786517	0	0	0	0	0	0	0	0	0	0	0
0.001034197	**Penalty**	0	0	0	0	0	0	0	0	0	0.079268	0	0	0	0	0	0.015094	0	0	0	0
0.009135411	**Save**	0	0	0	0	0	0.421687	0.090909	0.129944	0	0.097561	0.0075188	0.166667	0.009346	0.230769	0	0	0.071428571	0.211538	0.103448276	0.103448
4.83E-04	**Sending off**	0	0	0	0.023256	0	0	0	0.011299	0.0224719	0	0	0	0	0	0.033333	0	0	0.005495	0.068965517	0
0.012548263	**Shot**	0	0	0.909091	0	0.08391608	0.126506	0.08042	0.073446	0	0.042683	0.64661654	0	0	0.153846	0	0.679245	0	0.090659	0.103448276	0.068966
1.00E-03	**Substitution**	0	0	0	0	0	0.006024	0.024476	0.016949	0.0168539	0.012195	0.02255639	0	0	0	0.033333	0	0	0.008242	0.103448276	0.103448
1.00E-03	**Throw in**	0	0	0	0	0	0	0.031469	0.022599	0	0.012195	0	0	0.009346	0	0	0	0	0.019231	0.034482759	0.137931

Conditional Probability Matrix: P(Current Event Included | Previous Event Included)

Figure B.1: The conditional probability matrices of a) an event occurring in the full length soccer match, given that another event has just occurred and b) an event being included in the highlights, given that another event has just been included. For clarity, matrix elements are shown before the uniform prior is added to the zero elements.

Index	Feature Vector	Event Class	Code
0	event_class	Assist	0
1	by	Backpass	1
2	duration	Block	2
3	start_time	Booking	3
4	extra_time	Clearance	4
5	from	Corner	5
6	taken	Cross	6
7	to	Foul	7
8	resulting_in	Free_kick	8
9	type	Goal	9
10	on	Goal_kick	10
11	booked_for	Handball	11
12	dismissed_for	Header	12
13	off	Offside	13
14	team	Penalty	14
15	reason	Save	15
		Sending_off	16
from	**Code**	Shot	17
Null	0	Substitution	18
right by-line	1	Throw_in	19
left by-line	2		
right wing	3	**type**	**Code**
left wing	4	Null	0
right channel	5	outswinging	1
left channel	6	inswinging	2
centre of penalty area	7	attacking	3
right side of penalty area	8	defending	4
left side of penalty area	9		
right side of six-yard box	10	**taken**	**Code**
left side of six-yard box	11	Null	0
own half	12	short	1
		long	2
to	**Code**	left footed	3
Null	0	right footed	4
short	1	short left footed	5
far post	2	short right footed	6
near post	3	long left footed	7
centre	4	long right footed	8

Table B.2: Event metadata vector representation

reason	**Code**	**resulting in**	**Code**
Null	0	Null	0
injury	1	open play	1
tactical	2	ball out of play	2
		over the bar	3
dismissed for	**Code**	missed left	4
Null	0	missed right	5
second bookable offence	1	passed	6
serious foul play	2	hit the woodwork	7
violent conduct	3		
professional foul	4	**booked for**	**Code**
		Null	0
		unsporting behaviour	1
		dissent	2

Table B.2: Event metadata vector representation (*Continued*)

Appendix C

Business meeting "ticker-tape"

Each paragraph corresponds to one context-group of events. Each sentence is an event, encoded as: *event class, participant, start time, duration, topic.*

MeetingOpening Dave 00:00 60 "Present" Dave John Jan Alison Tony Rae Cathy Eddie Ben. MeetingOpening Jan 01:00 125 "Apologies" Alastair.

AgendaItem 03:05 "Matters Arising".

Information Jan 03:05 130 "Interviewing for new staff member. Tony and I will interview 4 candidates on the 25th". Proposal Jan 05:15 15 "Who else can be there?". Support Rae 05:30 20 "I can". Decision NemCon 05:50 5 "Interview 4 candidates on 25th" Tony Jan Rae.

Information Jan 05:55 5 "The Northampton charity had signed the naming agreement".

Information Jan 07:00 40 "The Trustee Year Plan is done".

Information Tony 07:40 90 "The site plan hasn't been done, now we're in the National Park it's harder to build. I've contacted Planning Aid".

Information Jan 09:10 60 "Discussion of Birthday plans deferred as Alastair isn't here".

Information Jan 10:10 250 "The date of Finance Sub Committee should be before 8th January. But in practice it's unlikely we can squeeze it in before the morning of the 8th". Proposal Dave 14:20 20 "Shall we wait until Chris has done the audit?". Decision NemCon 14:40 10 "Choice of date for Finance Sub Committee deferred until Chris has done the audit".

Proposal Dave 14:50 10 "Are the minutes accepted as correct?". Decision NemCon 15:00 10 "Minutes accepted".

AgendaItem 15:10 "CEO Report".

Question Ben 15:10 25 "What is the role of the artisan support worker?". Answer Jan 15:35 40 "To draw up projects to assist artisan and training. Capacity building to

help networks to form. To work on feasibilty of a tool refurbishment project for long term sustainability". Answer Jan 16:05 40 "To be run by our Partner Organisation. Our charity will write a Memorandum of Understanding with them, and provide £W pa".

Information Jan 16:45 40 "The founder is bad at management; he's an ideals person. Would have been better for him to move aside. But the organisation has lots of potential". Information Jan 17:25 25 "Ali hasn't been on the payroll for a year; but working as a volunteer. No one told us. Poor communication". Information Jan 17:50 65 "The founder has agreed in public that the staff position will go ahead. Therefore Ali is pleased; he believes it will happen". Information Jan 18:55 220 "We've been asked to donate 3 small kits to a school to keep the goodwill of the Minister, which will help our Partner Organisation. It's expedient for us, but it's a genuine project, so it's ok".

Question Dave 22:35 25 "Did the lottery explain why the grant was refused?". Answer Jan 23:00 75 "I can't remember details. Nothing wrong in the application, just too many applications".

Information Jan 24:15 140 "The Charity Commission visit has changed to Tuesday 5th April. We do not have signed forms from the Trustees that they are not bankrupt, over 18 and have no convictions". Proposal Tony 26:35 10 "I will circulate a full declaration to all for signature". Decision NemCon 26:45 5 "Circulate a declaration of eligibility of trusteeship to all trustees for signature" Tony.

Information Jan 26:50 30 "Papers for January 8th meeting will be ready mid December, as I am on holiday from 20th December".

Information Jan 27:20 120 "Here are 3 possible website home pages from Nicky. She wants us to say which is our 1st, 2nd and 3rd choice. She is offering to redesign the logo as well". Proposal Dave 29:20 20 "Have we got the right logo for the organisation?". Oppose Rae 29:40 15. Oppose Cathy 29:55 15. Oppose Ben 30:10 15. Support Alison 30:25 15. Support Tony 30:40 15. Oppose Jan 30:55 30 "We always include the text to explain the drawing". Oppose Dave 31:25 35 "It looks a bit dated". Oppose Rae 32:00 30 "It's artistically horrible. The Welsh charity uses an anvil. But it says what we are. It does not stand up on its own". Decision NemCon 32:30 210 "Ask Nicky to look at alternative designs, with the likelihood that we will change it" Jan.

Information Jan 36:00 10 "She won't want to waste time unless we are willing to change". Question Rae 36:10 10 "How many times have we changed it?". Answer Eddie 36:20 30 "Not many, we lightened it about 2 years ago".

AgendaItem 36:50 "Process for the Auditors Report".

Question Dave 36:50 10 "Why is this item on the agenda?". Answer Jan 37:00 10 "Because it was on last year".

Question Dave 37:10 10 "What did we say last year?". Answer Jan 37:20 70 "The Annual Review goes to the printers in December but the Board who have to approve

the auditor's final figures do not meet until January. Last year we approved it by email".

Question Dave 38:30 15 "Can it go via John so he can add comments?". Question John 38:45 15 "When will it be emailed?". Answer Jan 39:00 30 "December 12th. Then posted to trustees a few days later".

Information Jan 39:30 30 "Otherwise it's way out of date. In 2005 we send out the Annual Report for 2003-4. It looks very out of date, but our year ends in September". Question John 40:00 30 "Why not change the financial year end to December? There are no financial advantages to having it in Sepetember". Question Tony 40:30 30 "What effect would it have on the AGM? When would we get the figures?". Answer John 41:00 35 "You would have less time. We would get the figures in April". Information Tony 41:35 20 "We need 35 days before the AGM in May".

Question Dave 41:55 30 "Can we get back to the point? Can we discuss this at the next meeting with Chris?". Proposal Jan 42:25 15 "Do we accept the process as it stands?". Support Dave 42:40 50 "The formal decision is still at the 8th January meeting, but we will look at it earlier. Agreed?". Decision NemCon 43:30 50 "Email auditor's final figures to John on December 12th. If he agrees, post to the trustees a few days later. Formal decision still at the 8th January meeting" Jan.

AgendaItem 44:20 "October Finance Report".

Information John 44:20 100 "The September report should have said we have a shortfall against budget of £X rather than a deficit of X when we actually had a surplus that month". Information Jan 46:00 50 "In the 2 months to October our income was £Y below budget while expenditure was £Z below budget. Leaving us with a net deficit £Y-Z adverse to budget".

Question Dave 46:50 50 "It's useful to know how far out we are against what we said we would do. Is this time of year often like this?". Answer Jan 47:40 50 "Last year went up in October, but to less than where we are now".

Question Rae 48:30 50 "Has our expenditure gone up?". Answer Jan 49:20 50 "No. Our newsletter was late circulating which may have affected income timing". Information John 50:10 15 "But you've approved a further job so it will". Information Ben 50:25 20 "But for a fundraiser, so in 6-12 months we will get more income".

Information Dave 50:45 10 "We need to keep watching our income".

Proposal Tony 50:55 10 "It would be better to have a budget line on the graph". Oppose John 51:05 10 "It will become confusing if we have too many lines". Support Ben 51:15 120 "Can we get the Finance Sub Committee to look at the graphs, so that we can understand our income relative to budget?". Support Dave 53:15 60 "That can be an action item. Agreed?". Decision NemCon 54:15 70 "Review financial graphs so that financial state relative to the budget is clear" FSC.

Question John 54:25 65 "There's an unexplained difference of 25 between the September and August balance sheets. The movement in the bottom line, should

equal the net surplus in September. It does not. Was there a year end adjustment?". Answer Jan 55:30 15 "I'm not sure". Proposal Jan 55:45 65 "I'll follow it up with Chris and Lydia. Can you email me the details?". Decision Nemcon 56:50 55 "Follow up with Chris" Jan "Email Jan the details" John.

AgendaItem 57:45 "Group Kit Numbers Report".

Question Alison 57:45 25 "Do the groups ever get a certificate of appreciation when they've completed say 100 kits?". Answer Jan 58:10 100 "We don't have a record of past years kits and we don't publish it. Some groups only refurbish tools rather than complete kits so we don't want to discourage them. Their tools contribute to the Netley Marsh kits while Netley also contribute to some of the kits attributed to other groups".

Information Dave 59:50 95 "It's useful, but we have to be careful of the data". Information Jan 61:25 30 "Yes, Blackpool's one person while Milton Keynes have 50. How can you compare them?".

Information Dave 61:55 30 "Bob is feeling a bit stretched". Information Jan 62:25 45 "Bob has no one to organise him. He works very hard but does not organise his time and gets stressed". Question Dave 63:10 20 "Does that put you under pressure Jan?". Answer Jan 63:30 15 "You can't put him under too much pressure, but if he has none, he's in fairyland".

AgendaItem 63:45 "Ethical Policy".

Question Jan 63:45 15 "Have you got the updated version?". Answer Ben 64:00 10 "Yes, it's got my comments incorporated".

Question John 64:10 20 "I don't understand the last line all staff and volunteers will pay the full value for tools they purchase". Answer Jan 64:30 30 "It's saying you can't walk off with what you want, it's not ethical".

Proposal John 65:00 30 "We should add for tools they purchase from the charity for their personal use". Support Dave 65:30 30 "OK, with that amendment, can we endorse this?". Decision NemCon 66:00 5 "Ethical policy accepted, with the ammendment to the last line for tools they purchase from the charity for their personal use".

Question John 66:05 10 "What consultants do we use as fundraisers?". Answer Jan 66:15 10 "We don't at the moment. We have in the past, to research potential trusts".

Question John 66:25 10 "Do we pay a fee or a percentage of income generated?". Answer Jan 66:35 10 "Yes and this is saying we wont pay a percentage".

Question Rae 66:45 10 "What about Good Gifts. What are their charges to the charity? The lowest cost gift was 89 which is very high". Answer Cathy 66:55 5 "5 on top of the cost of gift". Answer Jan 67:00 120 "And they didn't ask us this year. We don't do plumbing kits but they put that in again. But we got £A last year".

Question Dave 69:00 40 "Can we move on?".

AgendaItem 69:40 "Update on Trustee Year Jobs".

Information Dave 69:40 20 "This is when you know you ought to have done something and haven't".

Information Jan 70:00 20 "This is on the agenda because you said you'd check what jobs we're meaning to do". Question Dave 70:20 10 "Did I say that?". Answer Rae 70:30 15 "I kind of remember".

Proposal Dave 70:45 115 "I think it's the trustee development, which we'll do in January. As it stands it's a pointless agenda item. The minutes of a meeting should cover the actions. You and I should chase up on Trustee Jobs". Oppose Jan 72:40 55 "But it's over the year. A certain thing has to be done by July, not from meeting to meeting. But you never did a plan for 2004. So there's forward thinking to it". Support Dave 73:35 15 "But the reality is of trustees' free time and getting things done".

Proposal John 73:50 90 "Can you and Jan make sure things that come out of you plan, get on the agenda and hence onto the minutes and hence into our To do lists". Support Tony 75:20 10 "Yes, you keep noticing that you haven't done it". Support Ben 75:30 30 "And when it becomes a priority". Support Jan 76:00 75 "So the year tasks for Employment Sub Committee, Finance Sub Committee and working groups get added to the agenda". Support Eddie 77:15 100 "What about photocopying the Trustee Year Plan on the back of the agenda each meeting?". Decision NemCon 78:55 15 "Photocopy the Trustee Year Plan on the back of each meeting agenda. Add items from the Board, Employment Sub Committee, Finance Sub Committee and working groups' year plans to the agenda" Jan Dave.

AgendaItem 79:10 "Board of Trustees' Development".

Information Dave 79:10 100 "We're half an hour behind schedule, so we'll postpone Board of Trustees development to the next meeting".

AgendaItem 80:50 "Membership".

Proposal Eddie 80:50 45 "2 new members have been proposed: Carol, Southampton and Philip from Emmer Green". Decision NemCon 81:35 10 "Accept as members".

AgendaItem 81:45 "Organisational Strategy".

Information Dave 81:45 10 "Issues arising from this morning's meeting".

Proposal Dave 81:55 110 "I'll write up our evaluation of the organisation strategy. Dave, Tony, Alison, Ben and Jan are on the steering group for Organisation strategy. Who will chair it? Ben?". Oppose Ben 83:45 55 "I'd like to be involved in the partnership group so I don't want to say, "I'm the chair." Whoever's in the Strategy Group needs to see it through for at least a year to AGM 2006". Support Jan 84:40 245 "You don't have to go to all 5 group weekend workshops. They collate the

questionnaire data. The chair it's more about getting other people to do it". Support Dave 86:45 35 "Did you say I'm the chair. It's on the tape". Support Ben 89:20 20 "OK. I'm prepared to do it". Decision NemCon 89:40 10 "Write up the Board's evaluation of the organisational strategy" Dave "The Steering group will develop the new organisational strategy" Dave Tony Alison Ben "Chair the steering group" Ben.

AgendaItem 89:50 "Dates of Next Meetings".

Information Dave 89:50 15 "On the back of the last minutes. Through as far as July 16th". Information Jan 90:05 10 "Haven't got venues and topics". Information All 90:15 30 "Yes there are". Information Dave 90:45 65 "January 8th. In the morning. Chris will be there and we'll discuss finance".

Proposal Dave 91:50 65 "Can we pick up the Board of Trustees development of skills in January?". Oppose Jan 92:55 50 "Chris usually talks for 2 hours on his audit process". Support Dave 93:45 20 "So we can have 45 minutes to 1 hour on our skills development. Then April 2nd can be a training workshop based on what we've identified at Ben's place". Support Jan 94:05 65 "OK, email next time to remind me to put the heating on". Decision NemCon 95:10 100 "45 minutes to 1 hour on Trustee Skills Development on 8th January" Board "2 hour discussion of finances" Board Chris "April 2nd meeting at Ben's".

Information Dave 96:50 50 "I think we're about fully recorded".

AgendaItem 97:40 "Any Other Business".

Proposal Jan 97:40 10 "Any Other Business?". Oppose Dave 97:50 10 "Isn't any. Good!". Decision NemCon 98:00 5 "No Other Business".

MeetingClosing Dave 98:05.