Generating Multiple Summaries Based on Computational Model of Perspective

by

Alice H. Oh

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2008

Author .....
Department of Electrical Engineering and Computer Science
September 1, 2008

Certified by .....
Howard Shrobe
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chairman, Department Committee on Graduate Students

Generating Multiple Summaries Based on Computational Model of Perspective

by

Alice H. Oh

Submitted to the Department of Electrical Engineering and Computer Science
on September 1, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Every story about an event offers a unique perspective about the event. A popular sporting event, such as a Major League Baseball game, is followed by several summary articles that show different points of view. The goal of this research is to build a computational model of perspective and build a system for automatically generating multiple summary articles showing different perspectives.

My approach is to take a neutral summary article, reorder the content of that summary based on event features extracted from the description of the game, and produce two new summaries showing the local team perspectives. I will present an initial user survey that validated the hypothesis that content ordering has a significant effect on the users' perception of perspective. I will also discuss collecting and analyzing a parallel corpus of baseball game data and summary articles showing local team perspectives. I will then describe the reordering algorithm, the implementation of the system, and a user study to evaluate the output of the system.

Thesis Supervisor: Howard Shrobe
Title: Principal Research Scientist

Acknowledgments

This thesis would not have been possible without my kind and caring friends, family, and colleagues. The past several years I spent at MIT were filled with such beautiful memories because of them, and I will miss them greatly as I leave most of them behind (only physically) to go onto my next adventure.

The members of the AIRE group were wonderful from the beginning to the end. I thank Krzysztof Gajos, Kevin Quigley, and Stephen Peters for their cheerful smiles. There were many fun moments with UROPs and Master's students as well, and I have to give special thanks to my office mates, Max Van Kleek, Harold Fox, and Gary Look. I also enjoyed having lunch and conversations with Randy's group, Sonya Cates, Aaron Adler, Mike Oltmans, Jacob Eisenstein, and Tom Ouyang. The members of TiG were always kind and helpful with all kinds of problems I had.

I appreciate all the help, encouragements, and comments from my committee, Patrick Winston and Regina Barzilay, and I especially thank my advisor, Howie Shrobe, for everything he has done for me. I thank Trevor Darrell who worked with me for the first couple of years at MIT.

I thank Alex Rudnicky, my Master's advisor at CMU, for showing and opening the door to research for me. Your words of trust and encouragement remind me, every now and then, that I can succeed in research. I also thank all the friends and faculty at CMU LTI for introducing me to the wonderful world of computational linguistics.

My friends in Boston3040 were in charge of entertaining me through the nights when the stress level was at the highest. The times I had with them playing tennis and having wonderful conversations will never be forgotten. I won't name them all here, but they are like family to me, and I thank each and every one of them with all my heart.

I thank my friends in KGSA EECS who have cheered me up many times. I am especially grateful to Eunjong Hong and Jeewook Kim who have endured many years at CSAIL with me. Congratulations to both of them for their dissertations.

My parents-in-law and sister-in-law were supportive of my studies all throughout,

always generous with babysitting, and I am very lucky to have them in my life. My own two sisters, Claudia and Kathy, will be my best friends for life, and I am thankful for them and their beautiful families. My parents always say they thank me for being their daughter, but I cannot express in words, how thankful I am, for their never-ending love and support.

The two most important and beautiful people in my life, Taesik and Herin Jamie, I thank you for your smiles and kisses. I am blessed to have Taesik as my husband, friend, colleague, and love for life. He has shown me how to be a better person, both at home and at school. Jamie is just a delightful daughter, and she gives me joy and love much bigger and better than I could ever imagine before she was conceived. It has been amazing to watch her grow as I watched my research turn into this thesis, and I dedicate this thesis to her with all my love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis defines the notion of perspective in concept-to-summary generation for a baseball game, proposes one approach for modeling and automatically generating multiple perspectives, and shows a system that implements the model of perspective.

Before we begin describing the details of the problem, let us consider the following two excerpts about a major league baseball (MLB) game between the Boston Red Sox and the Toronto Blue Jays in 2005:

> Things looked almost too good to be true for the Blue Jays in last night's fourth inning as they were laying another whupping on the defending champions.
>
> That's because Toronto's four-run lead and a faltering Boston Red Sox pitcher in David Wells really were too good to last. The Jays slowly saw their dream become a nightmare by the middle innings as Wells regained his footing and the biggest Boston bats came to life.
>
> Reality descended completely in the seventh as Manny Ramirez, contained throughout the series, lofted a fly ball towards a mysterious right field corner of this park that wound up being a decisive home run.
>
> "Right there, I'm thinking it's a fly ball to right field," Jays pitcher Pete Walker, reverting to a relief role in this one, said after Ramirez's two-run blast handed Toronto a 6-4 loss in front of 35,268 fans at Fenway Park.

"So then, I started watching the ball and it started floating towards the right field corner. Just from what I saw, the way it kept going out there, I had a feeling it was going to sneak out."

The right-field corner at Fenway is a short porch with a waist-high fence that has turned many a fly ball into a round-tripper. It's just one of the advantages the Red Sox have used over the years and another reason why no lead here is ever safe.

But there are two bigger reasons the Jays are back to six games behind Boston, instead of four. One was the 11 runners the Jays wound up stranding, including a pair in the ninth when Gregg Zaun flied out to right on a Keith Foulke pitch he just got under. Foulke had also induced Reed Johnson to fly out with two on to end the eighth.

–From the Online Edition of *Toronto Star*, July 3, 2005.

A season is all about evolution, a player finding himself, and, when called for, management finding someone new. Plenty of that was on display yesterday at Fenway Park on what was, beginning to end, an eventful, entertaining day in the Back Bay.

There was Matt Mantei going – he's out a minimum of 3-4 weeks, and quite possibly the season, with ligament damage in his left ankle. There was Pawtucket lefthander Abe Alvarez coming – to replace Mantei, though Alvarez's stay is unknown. There was David Wells fighting himself at times, then fighting the umpires, completing his night after 6 2/3 innings by getting ejected.

There was Manny Ramirez continuing to find his swing, launching his ninth homer in his last 17 games to snap a 4-4 tie and vault the Sox to a 6-4 lead, which is how it ended against the pesky Blue Jays. And there was Keith Foulke – in his first appearance since being taken out of the yard by Cleveland's Travis Hafner, then taking his anxiety out on the fans – inheriting two runners with two outs in the eighth and getting Sox killer

Reed Johnson to fly to right. He then held the Blue Jays scoreless in the ninth, though he allowed two singles.

In fact, in the search for new beginnings, Foulke came out to a new theme song last night. Scrapped was "Mother," by Danzig, a song Bronson Arroyo chose for the Sox closer. Now Foulke comes jogging out of the bullpen to Hank Williams Jr.'s "Country Boy Can Survive."

"It's time," Foulke said of the musical change. "You listen to the song and you'll understand why."

Foulke was in position for the save, his 15th, thanks to Ramirez, who powered the Sox to only their third win in 10 games this season against Toronto. Ramirez did his damage with one vintage swing in the seventh with David Ortiz aboard and nobody out.

Blue Jays reliever Pete Walker left a hanger on the outside corner, and Ramirez stepped toward the mound, not the ball, and simply flung his bat at it. The ball started toward right fielder Alex Rios before fading into the fandom behind Pesky's Pole. The homer gave Ramirez 45 against the Blue Jays, the most against Toronto by any player since the team entered the American League in 1977.

–From the Online Edition of *The Boston Globe*, July 3, 2005.

The two excerpts above are from the online news sources for the two teams: the *Boston Globe* and the *Toronto Star*. Although they are written about the same game, the two articles are distinctively different in what and how they choose to tell in the stories. The first article by the Toronto newspaper focuses on their initial lead being blown away by a Red Sox homerun, and the missed opportunities that the Toronto offense had throughout the game. The second article by the Boston newspaper focuses on the team's various players either stepping up to fill their roles or failing to do so, with Manny Ramirez and Keith Foulke being highlighted for their roles in this game. This simple comparison shows that the same set of events that happened during and around the game can be told in multiple versions of narrative. The overarching goal

of this thesis is to look at this problem of generating multiple summary articles of a single event.

One big obstacle to this goal is that there are many dimensions along which the articles differ. Some of those dimensions are based on content and are somewhat easier to identify, such as the set of players discussed, specific game events mentioned, and the team and player statistics in the game and previous games in the season. Others are more difficult to pinpoint and almost impossible to measure, such as the writer's opinions about the team, predictions for the remaining games in the season, and emotions toward the home team and the opposing team. In fact, besides these two articles, there are usually several more articles in each of the two newspapers on the same game that also differ significantly from these articles along various dimensions. Looking beyond the two newspapers, naturally there are other local and national newspapers that also feature articles written about the same game. If you add online sports news, personal blogs, and discussion boards on the Internet, there are literally hundreds of stories that people tell about one single baseball game.

That problem of taking one baseball game and generating hundreds of articles that are significantly different, in itself, is both not interesting and too difficult. It becomes much more tractable and interesting if I can identify one important dimension along which the articles should differ. I chose to look at *point-of-view*, or *perspective* as I will call it throughout the thesis, for that one important dimension.

1.1 Motivation

The motivation for this work comes from the general observation that people enjoy reading multiple stories about an event. At first, this seems counterintuitive in this fast-paced world loaded with all kinds of information right at the fingertip of the user. Why would he want to spend time reading multiple stories about one subject when he can read one story each about multiple subjects? One answer is offered at the BBC (British Broadcasting Company) news website:

> Our users tell us that one of the things they value most about our service

Figure 1-1: BBC lists and links to other articles on the Internet.

> is our policy of linking openly to other websites.
>
> These links offer access to more detailed information, the chance to compare sources or check out a different perspective on the same story.
>
> – From BBC News on the Web at http://www.bbc.co.uk

This quote refers to a section in some of their news stories that link to other stories about the same event. Figure 1-1 is an example of that section.

A similar approach is taken by the Major League Baseball (MLB) website at http://www.mlb.com. For every game, they offer a comprehensive summary of the game, including all the facts of the game (game description and box scores) as well as two sides of the story, one for each team that played in that game. Figure 1-2 shows an example of this.

Google News (http://news.google.com) is another example of offering users several articles on a single topic (Figure 1-3). They also specify how many other articles they can find using Google search, and that number is sometimes in the several hundreds.

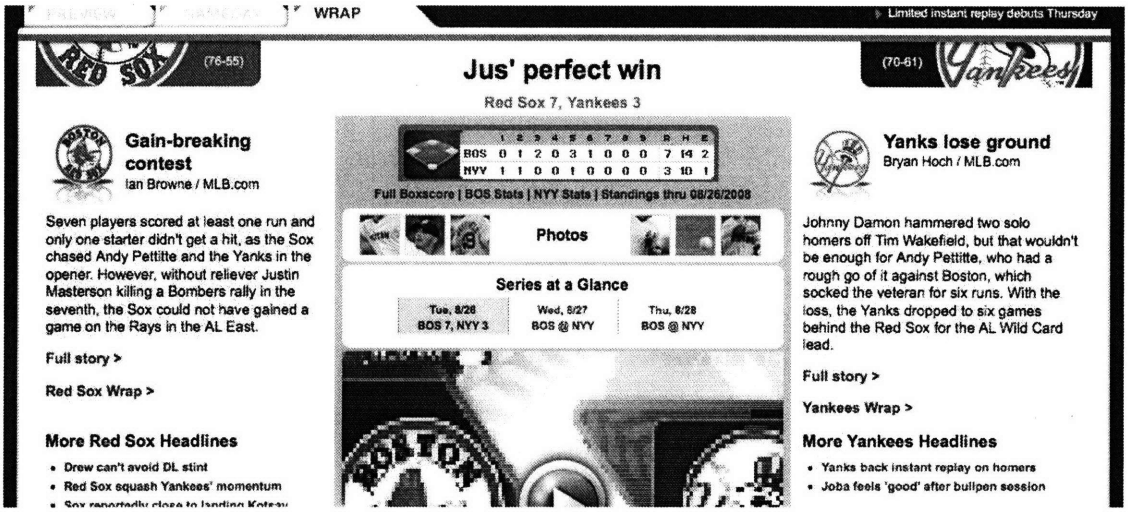


Figure 1-2: The Major League Baseball website features two wrap-up stories for every game.

Clinton Rallies Her Troops to Fight for an Obama Victory

New York Times - **1 hour ago**

By PATRICK HEALY DENVER - With her husband looking on tenderly and her supporters watching with tears in their eyes, Senator Hillary Rodham Clinton deferred her own dreams on Tuesday night and delivered an emphatic plea at the Democratic National ...

Video: Clinton's Speech Praised CBS

Free Press voter panel responds to speech Detroit Free Press

FOXNews - Reuters - Voice of America - Newsweek

all 7,848 news articles »

Figure 1-3: Google News links several online articles on the same topic and also tells the user how many other articles the Google search has found on that same topic.

1.2 Perspective

It is first necessary to define what is meant by the terms *perspective* and *multiple perspectives*. The definition of perspective in this thesis is somewhat different from a more traditional meaning of *perspective* or *point-of-view* in literature.

> **point of view** The perspective from which a story is presented to the reader. The three main points of view are first person, third person singular, and third person omniscient.[18]

There is work by [41] which uses that definition of perspective, where a computational system tries to infer the narrative character whose point of view is presented in each sentence.

Our definition is much closer to that used in [23], where they look at ideological perspectives of online articles on political, social, and cultural issues. They look at the political domain of the issues between Israel and Palestine, and they try to infer, for each online article, whether it is written from the Israeli perspective or the Palestinian perspective.

This is an important problem, especially as the amount of textual information available via Internet becomes larger every day. For every topic, there are many well-written articles worth reading, but because of the huge amount of text, it is difficult to identify which articles to read. With well-known sources such as the online versions of large newspapers (e.g, *The New York Times*), the general perspective and attitude of the journalists can be inferred just by knowing the source. However, with more personal blogs and smaller-scale online journalism becoming more ubiquitous and important, it is often difficult to know the perspective of an article without actually reading the article, and for current events news stories where only partial stories are told initially, the reader would not be sure which side's story they are reading. Hence, work such as [23] that tries to automatically identify the perpsective of an article is interesting and pertinent. On the flip side, it would be useful to be able to automatically generate stories from multiple perspectives. Simply for applications, if a user wants to read about an event from a certain perspective, he would simply

ask for an article to be written from that perspective, and an automatic generation system would produce an article to suit his needs. An important side-effect of such application building would be that we would be able to gain a deep understanding of the computational model behind generating multiple perspectives.

This thesis looks at just that problem in the domain of baseball games. That is, I examine the home team vs. visiting team articles, come up with an algorithm for generating such articles, and build a prototype system. I assume that the two opposing perspectives are expressed in the local newspaper articles of the two teams, and I assume that the neutral perspective is expressed in the Associated Press articles published on an ESPN website (www.espn.com). I confirmed these assumptions via a user study, then I identified some key factors contributing to an article having a certain perspective.

I model this problem as an instance of text-to-text generation (see [4]), a sub-problem within natural language generation (NLG). NLG encompasses the vast problem of automatically generating text. Most NLG systems divide the generation process into content planning and surface realization. Content planning spans the tasks of choosing, ordering, and structuring the content into paragraphs. Surface realization takes that planned content and produces sentences using either pre-made templates or syntactic and lexical selection rules.

1.3 Challenges

An initial challenge of this work is the problem definition itself. There are infinite ways to generate a summary article following any event, and since it is impractical to come up with an algorithm that will try to generate as many of those as possible, it is necessary to define the problem by deciding on a dimension along which the output should vary. To make the problem easier to solve, a widely used dimension is ideal, but to make the problem interesting, a conceptual and generalizable dimension should be chosen, rather than some arbitrary feature of the domain. For example, it would be easy to say that the dimension should be the set of players mentioned in the story,

but that dimension is not flexible enough to be generalized for all teams and varieties of game situations. An added challenge is that I do not assume that a user model of the audience is known, other than the very coarse model that represents only the team that they are supporting. A lack of knowledge about the user implies that the generated output should exhibit behavior that mimics human-generated output with a wide audience in mind rather than targeting a specific user group with a known set of interests [30]. Also, to take advantage of the widely available data and well-specified game rules of the domain, automatically built knowledge of the game should be used, as well as neutral articles as the starting point. Other concept-to-text algorithms such as [21] also take advantage of widely available domain data, so lessons can be learned from those previous systems and be applied to this problem as well. The last challenge is to identify and propose a novel algorithm for a subtask within NLG that would work well for this problem.

1.4 Problem Definition

I can look at the problem of multiple perspective generation at every level of NLG– from content selection to lexical selection, but for this thesis, I further specify the problem in two ways. First, rather than generating the articles from scratch, the system takes an article written from a neutral perspective and makes transformations on that article to produce two other articles, each from a different perspective. While the assumption that there exists a neutral article to begin with may be significant, it is not an unrealistic one, as neutral, or close to neutral articles are abundant for a variety of news topics if one considers news sources such as Associated Press (AP) or Reuters to be credible and near-neutral sources. In one sense, this is a simplification of the problem because the generated output is a transformation of a document that is already existing and written by a human writer. On the other hand, this specification may enable the generation algorithm to be more generalizable if we can take advantage of the neutral article in a way such that the process of domain knowledge acquisition can be eliminated. If we can isolate the transformation algorithm to be domain-

independent given the neutral article, this specification would extend the algorithm to be much more powerful. Although the prototype tested in this thesis does not include that extension, I will show a preliminary experiment that looks promising.

Secondly, I chose to focus on the part of NLG that deals with content ordering. Content ordering is an important subtask within NLG, and much work has been done in it, but most of it has been pairwise ordering constraints, in which the algorithm would decide whether sentence A should come before or after sentence B. I propose a re-ordering algorithm that considers more than two sentences (or units of content) at once, and I assert that content ordering alone can contribute to significant changes in perspective. A more detailed explanation comes in a later chapter that explains why working only with content ordering is enough for our prototype.

The problem definition boils down to the following hypothesis:

> **Ordering Hypothesis:** Ordering of the content alone contributes significantly to the perspective of a story. Hence, you can generate multiple perspectives by taking a pre-determined set of content and reordering it.

I will show, through the rest of this thesis, that this hypothesis is true. The first step is running an initial user survey about what factors of a text contributes to perspective. The results of that survey indicates that ordering is a significant factor in perspective. Then, the second step is building a system that is based on this hypothesis. The system is essentially an implementation of content re-ordering algorithm. The third step is evaluating that system such that, if the results of the system output show the desired perspective, then I can say the ordering algorithm above is valid.

1.5 Contributions

The major contributions of this thesis are presented and discussed in detail in the following chapters in this order:

- **multiple perspectives for summary generation**: In the present chapter, I have defined the problem of generating multiple summary articles of a baseball game with different perspectives. To make the problem tractable using a computational model, I have narrowed it down to a content planning problem within text-to-text generation, and to justify that, I proposed the Ordering Hypothesis.

- **collecting and analyzing data for studying perspective in the baseball domain**: In Chapter 2, I will discuss how I collected and analyzed data for studying multiple perspectives. The data consists of automatically downloaded game data and local perspective articles to constitue a parallel corpus. This chapter explains an important step in transforming textual domain descriptions into feature vectors used in our computational domain model, as well as aligning sentences in the parallel corpus with feature vectors in the domain model.

- **describing user studies used for identifying potential sources of perspective**: Chapter 3 presents the first user study in which users were asked to rate various versions of the local team articles and neutral articles. By modifying the original parallel corpus in four steps and having the subjects rate the modified articles, the study was able to confirm the validity of the Ordering Hypothesis.

- **showing content reordering as an effective way to generate multiple perspectives**: Chapter 4 presents the details of the reordering algorithm and how it is implemented in the prototype system. It describes the different ordering strategies found in the corpus, and how the ordering strategies are chosen using a statistical weighting scheme.

- **evaluating the prototype to show that the reordering algorithm works**: Chapter 5 discusses the user study for evaluating our prototype. The system output was compared against baseline summaries, and statistical tests show that the users rated the system-produced summaries as showing the desired

perspectives.

The last two chapters, 6 and 7 present discussions on related work and concluding remarks. Chapter 6 situates this thesis within the related work in the areas of user modeling, sentiment analysis, perspective analysis, and cognitive science and media studies. Chapter 7 presents concluding remarks including contributions of this work and future directions.

Chapter 2

Corpus Collection and Analyses

A substantial part of this thesis work was in choosing the domain, then collecting and analyzing data for that domain. Although it is important to show generality and extensibility of the model and algorithms by applying them across different domains, it is first necessary to show that a newly defined problem can be solved in a specific domain, in a proof-of-concept way. I chose the domain of baseball to serve this purpose, and in this chapter, I will elaborate on the details of the domain and describe the processes for automatically collecting baseball game data and corpus of news articles. I will then illustrate how I used the game data to extract a semantically rich domain model. Then, I will show the two stages of using the corpus of articles to discover one way to model multiple perspectives. The first stage is aligning the sentences in the articles with the corresponding game events, represented as feature vectors, in the domain model. The second stage is finding patterns of the feature vectors depending on perspective, thereby discovering a model of perspective based on the corpus and the domain model.

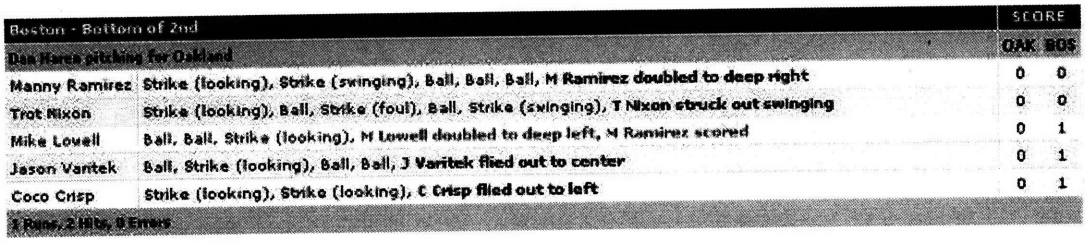
2.1 Baseball Domain

Many previous works in NLP choose sports, such as soccer [1], basketball [31], and other sports (*cite*) as the domain in which to test the ideas. There are a few good reasons for choosing sports over other possible domains. First, there is a large body

of data to work with. Every day, there are many sporting events taking place, and more importantly, being talked about in newspapers, television, and the Internet. That results in hundreds and thousands of documents and transcripts to collect and analyze. Secondly, unlike most other domains, sports games have well-defined rules about possible events, timeline, and entities. For example, a baseball game has nine innings, or eighteen half-innings. The two opposing teams take turns playing offense and defense for each half-inning. The teams are made up of nine active players, and their positions are pre-specified. For each batter coming onto face the pitcher, there is a finite set of outcomes (e.g., homerun, strikeout) of that pitcher-batter interaction. Thirdly, despite a well-defined set of rules, sports domains are fairly complex and rich. There are different types of entities, players, teams, groups of players, coaching staff, and they interact in ways that are analogous to everyday interactions among entities in the non-sports domains. The events and timeline are also complex, in that events in a game, and games in a season, can be organized hierarchically. In baseball, pitches make up an at-bat, at-bats make up a half-inning, two half-innings make up an inning, innings make up a game, and games make up a season. Those event and time units are best represented in a hierarchical model, in which it would be possible to compute and identify important relationships and transitions. Lastly, although the rules and hierarchies are artifically constructed in the sports domain, many of the same types of rules and hierarchies exist naturally in other domains. Interactions among people and organizations, chronological ordering of events and their relationships, and unwritten but unambiguous rules of interactions are ubiquitous in non-sports domains, and hence many of the questions and answers discussed in this thesis are applicable to other domains.

2.1.1 Game Data

The Major League Baseball (MLB) has 30 teams within the United States and Canada, and each team plays approximately 160 games per season. I have collected data for approximately 600 games from the 2005 and 2006 MLB seasons. For every MLB game, the website of MLB (www.mlb.com) publishes game data consisting



Boston - Bottom of 2nd		SCORE	
Dan Haren pitching for Oakland		OAK	BOS
Manny Ramirez	Strike (looking), Strike (swinging), Ball, Ball, Ball, M Ramirez doubled to deep right	0	0
Trot Nixon	Strike (looking), Ball, Strike (foul), Ball, Strike (swinging), T Nixon struck out swinging	0	0
Mike Lowell	Ball, Ball, Strike (looking), M Lowell doubled to deep left, M Ramirez scored	0	1
Jason Varitek	Ball, Strike (looking), Ball, Ball, J Varitek flied out to center	0	1
Coco Crisp	Strike (looking), Strike (looking), C Crisp flied out to left	0	1
1 Runs, 2 Hits, 0 Errors			

Figure 2-1: Pitch by Pitch Log of a Baseball Game

of two documents. The first document is a game log (see figure 2-1) , which is a complete list of *atBats* in the game (see 2.1.2 for definition of *atBat*. There are at least 3 *atBats* per half of an inning (top or bottom), and there are at least 9 innings per game (except in extreme weather conditions), so there are at least 54 *atBats* per game, but usually more. In our corpus, the average number of *at-bats* is 76.2 per game. The second document is a boxscore, which is a list of each batter and pitcher's performance statistics for the game. Currently I do not use the boxscore documents in this work.

The game log is a complete pitch-by-pitch account of the game events. It describes what happened for each pitch that the pitcher threw during the game. It includes individual pitch-level outcomes such as strike or ball, as well as outcomes that result in the end of the current *at-bat*, such as a strikeout, a hit (e.g., single, double, homerun), and various non-strikeout outs (e.g., foulout, lineout). It also lists any runners on base, whether they advanced to the next base on the play, and whether anyone crossed the home plate to add to the score. Each line also includes the home and away team scores at the end of the play, and the number of outs (0, 1, or 2) at the end of the play. As such, each game log can be turned into an accurate and complete model of events in the game, and patterns over those events can easily be computed. Some example patterns that can be computed over the events include two-out scoring events, bases-loaded third out, eighth or ninth-inning blown saves. Section 2.2 has a more detailed description of the domain model and pattern computations.

It must be noted, however, that a game log does not contain events that are not directly from the game. There are events that take place outside of the game itself but are closely related to the game, such as player injuries, trades, and coaching

decisions. Those events are often very important and thus are described frequently in game summary articles. The game summary articles also contain many player and coaching staff quotes, expressing opinions and insights about the game. I chose not to include the non-game events and quotes in this sytem, as the main focus of the thesis is not about summary generation, but about perspective generation in summaries. If the system can generate multiple perspectives with only the game event descriptions, then it is not necessary to include the non-game events and quotes. Those extras may be further studied as one of the next steps of this work.

2.1.2 Baseball Rules and Terms

The website of MLB (www.mlb.com) has a wealth of information on baseball rules and terminology. Here I will present the ones that are used in my system.

- **pitcher**: the player on the defensive team who throws the ball. There is a *starting pitcher* who starts pitching from the very beginning and pitches for three to nine innings. He is the most important pitcher.
- **batter**: the player on the offensive team who tries to hit the ball. Almost every batter also plays a defensive position, such as *catcher*, *left-fielder*, etc.
- **atBat**: a batter's turn in the batter's box consisting of a set of balls thrown from the pitcher to the batter such that the outcome is either an out or an advancement of the batter to a base including the home base, which would be a home run.
- **baseHit**: a hit in which the batter safely advances to a base.
- **walk**: a set of four *balls*, as opposed to *strikes*, that automatically advances the batter to the first base.
- **RBI (runs-batted-in)**: a play in which one of the offensive players (either already on base or the batter himself) safely reaches the home base and scores.

- **inning**: a set of *atBats* that result in three outs makes up a half-inning. An inning consists of two half-innings, the first half is called *top*, and the second half is called *bottom*.

2.2 Domain Model

In addition to the game logs, the MLB website (http://www.mlb.com) has team rosters, listing all the player names and coaching staff in each team. Using the team rosters, game logs, and basic knowledge of the structure of the baseball games, I built a hierarchical model of the game, divided up into two parts, entities and events.

2.2.1 Entities

Entities are individual players, coaching staff, groups of players or staff, and the entire team. The entities are structured hierarchically, where a team is made up of players and coaching staff, and each player and staff member can belong in one or more groups. Each player is represented by his first name, last name, and defensive position (e.g., pitcher, first baseman), such that player lookup can be done by a combination of those fields. Groups are formed dynamically for each game based on entity and event features. Dynamically formed groups are based on defensive position (e.g., pitchers) or game performance (e.g., batters with RBIs in the game). Groups are useful for computing group performance, such as the pitchers' combined earned runs (ER) or strikeouts. The system computes the group performance metrics but does not yet use the group performance analysis in the generated summaries.

2.2.2 Events

The events in the game are also organized hierarchically. The smallest unit is each pitch the pitcher throws. Then, the pitches make up an *atBat*, a series of pitches to a particular batter. Three or more *atBats* make up a half-inning, and two half-innings make up an inning, and finally, nine or more innings make up a game. The first step in

building a model of the game events from the game log is parsing the log such that each *atBat* is turned into a feature vector using simple regular expression type patterns. These are the features used in the system: *inningNumber, atBatNumber, pitchCount, homeScore, visitScore, team, pitcher, batter, onFirst, onSecond, onThird, outsAdded, baseHit, rbi, doubleplay, runnersStranded, homerun, strikeOut, extraBaseHit, walk, error, typeOfPlay.*

Some of these features, such as *batter* and *typeOfPlay* are extracted directly from each line in the log that is being transformed into a feature vector. Some of the features, such as *inningNumber*, *team*, and *pitcher* span multiple contiguous *at-bats* and are extracted from the current line or in one of the lines going back a few *at-bats.* The remaining features, such as *onFirst*, *outsAdded*, and *runnersStranded* are derived from looking at the feature vector of the previous *at-bat* and following simple rules of the baseball game. For example, *onSecond* is derived from looking at the previous feature vector's *onFirst* value and whether the current play is one that advances the runner one base. If *onFirst* is not null and the current play advanced runners, then the previous feature vector's *onFirst* gets copied to the current *onSecond.* While I tried to identify features that are important in a baseball game, later sections will show that some of them were not used for analyzing and generating multiple perspectives, as only a subset of the features were significant variables for our content reordering algorithm. Here are descriptions of all the features and how they are computed from the game logs.

- inningNumber: the ordinal number for an inning; 0 (top of first), 1 (bottom of first), 2 (top of second), ... This is extracted directly from the first line in each half-inning.

- atBatNumber: the ordinal number for the current *atBat.* This is a counter that increments for each *atBat* line in the log.

- pitchCount: the number of pitches a pitcher throws for a particular *atBat.* This is a count of strikes, balls, and fouls that are listed in the line of the *atBat.*

- homeScore: the current score (before the end of the current *atBat* for the home team. This is extracted from the line.
- visitScore: the current score (before the end of the current *atBat* for the visiting team. This is extracted from the line.
- team: the three-letter name of the offensive team. The two team names are extracted from the beginning of the log, and for each half-inning, the offensive team switches.
- pitcher: the name of the current pitcher. This is extracted from the line of the current *atBat*.
- batter: the name of the current batter. This is extracted from the line of the current *atBat*.
- onFirst: baserunner on first base, null if no one is on. This is parsed from the previous line.
- onSecond: baserunner on second base, null if no one is on. This is parsed from the previous line.
- onThird: baserunner on third base, null if no one is on. This is parsed from the previous line.
- outsAdded: an integer value between 0 and 3 for the number of outs this AtBat has generated. This is computed as the difference between the number of outs in the previous *atBat*, and the number of outs after the current *atBat*.
- baseHit: an integer value between 0 and 3. 0 for no hit, 1 for a single, 2 for a double, and 3 for a triple. This is parsed from the current line using keywords "singled", "doubled", and "tripled".
- rbi: an integer value for the number of runs this AtBat has generated. This is computed as the difference between the score in the previous *atBat* and the score of the current *atBat* line.

- doublePlay: a boolean value. True if this AtBat resulted in a double play, causing two outs to be added. This is parsed using the keyword "double play".

- runnersStranded: an integer value between 0 and 3 for the runners on base when this AtBat has ended to add the final (third) out of the inning. This is determined by looking at whether this *atBat* is the last in the half-inning, and whether onFirst, onSecond, or onThird is non-null.

- homerun: a boolean value. True if the batter hit a homerun, adding one or more points to their team's score.??This is identified by the keyword "homerun" in the current *atBat* line.

- strikeOut: a boolean value. True if the AtBat ended with three strikes, adding one out.

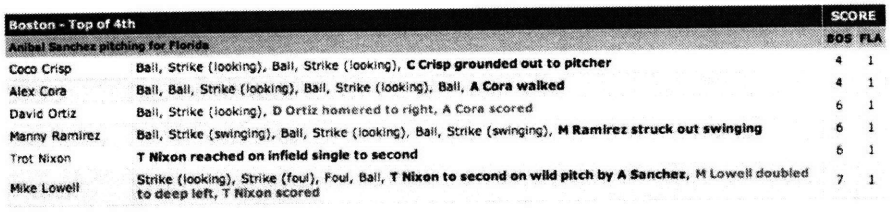
- extraBaseHit: a boolean value. True if this atBat resulted in an extra base hit (a double or a triple), False otherwise. This is identified by keywords "double" and "triple".

- walk: a boolean value. True if the atBat ended with four balls, advancing the batter to first base and other baserunners if applicable.

- error: a boolean value. This is identified by the keyword "error".

- typeOfPlay: the final outcome of the *atBat*. Possible values include strikeout, walk, foulout, lineout, popout, single, dobule, triple, homerun, fielderschoice, etc.

Figure 2-2 shows an excerpt from a game log and how the lines are parsed into feature vectors. Because of space limitations, this example leaves out several features and shows the most interesting features. Occasionally, spelling errors and other abnormalities in the game log causes the feature vectors to be partially incorrect, but more than 99% of the time, the game logs are parsed correctly into the feature vectors.



Boston - Top of 4th		SCORE	
Anibal Sanchez pitching for Florida		BOS	FLA
Coco Crisp	Ball, Strike (looking), Ball, Strike (looking), **C Crisp grounded out to pitcher**	4	1
Alex Cora	Ball, Ball, Strike (looking), Ball, Strike (looking), Ball, **A Cora walked**	4	1
David Ortiz	Ball, Strike (looking), D Ortiz homered to right, A Cora scored	6	1
Manny Ramirez	Ball, Strike (swinging), Ball, Strike (looking), Ball, Strike (swinging), **M Ramirez struck out swinging**	6	1
Trot Nixon	**T Nixon reached on infield single to second**	6	1
Mike Lowell	Strike (looking), Strike (foul), Foul, Ball, **T Nixon to second on wild pitch by A Sanchez**, M Lowell doubled to deep left, T Nixon scored	7	1

batter	pitcher	play	inn	atbat	s	b	h	v	1st	r	o
crisp	sanchez	grndout	4	18	2	2	4	1	none	0	0
cora	sanchez	walk	4	19	2	4	4	1	none	0	1
ortiz	sanchez	homerun	4	20	1	1	4	1	cora	2	1
ramirez	sanchez	strkout	4	21	3	3	6	1	none	0	1
nixon	sanchez	single	4	22	0	0	6	1	none	0	2
lowell	sanchez	double	4	23	2	1	6	1	nixon	1	2

Figure 2-2: An example of how a game log is parsed into feature-vectors.

2.2.3 Computing Over Feature Vectors

The feature vectors contain much information, but some simple computations can be done over the feature vectors to gain more insight into the domain. Here are some examples of computing over feature vectors:

- Two-out scoring plays: This is computed by looking at the values of the feature *outs* and the feature *rbi*. If *outs* is two, and *rbi* is greater than 0, then this feature is set to true, otherwise false.

- Lead-changing plays: This is a boolean value, set to true if $teamAscore - teamBscore$ has different sign (negative vs positive) for the current *atBat* and the next *atBat*.

- Runners stranded: This is computed by looking at the values of the *onFirst*, *onSecond*, and *onThird*, and the value of *outs* of the current and next *atBats*. If any *onXX* has a non-null value, and *outs* at the end of this *atBat* is three, then this is set to true.

- Number of extra-inning hits: This is a count of *atBats* in the half-inning for which the value of *playType* is double, triple, or homerun.

These are just examples, and there are many more of these higher-order features that can be computed by looking at the simple features of the *atBats*. The domain model being used in the current system is flexible to allow these features to be computed and added to the model for richer analysis.

2.3 Corpus of Summary Articles

In addition to the game logs and boxscores which serve the purpose of automatically building domain models, I use online newspaper articles to build the corpus from which to learn how the summaries are written from multiple perspectives. Since, for every game, there are several articles written and published about the game, all from different perspectives, collecting and analyzing those articles would reveal ways of generating multiple perspectives based on the same set of events.

2.3.1 Choosing Sources

Following a baseball game, many online and print newspapers publish stories based on that particular game. Even in a single newspaper, there may be several articles about the game. Additionally, there are sports and personal blogs that also publish online stories on the same game. To constrain the corpus such that data collection is practical and data is consistent in terms of perspective, it makes most sense to collect the main wrap-up stories from the major local newspapers of the two opposing teams. That way, the two local team perspectives can be the target for the system to model and generate. A simple and reasonable assumption would be that the contributing factors to multiple perspectives are the major differences between the two opposing teams' local articles. Of course, even if the corpus is constrained to the main stories in the major local newspapers, there are confounding variables, such as the specific journalist's style or the editor's biases, as well as the overall tone and attitude of the newspaper toward its hometown team. However, by taking several different sources and searching for common factors among them, much of those issues can be eliminated.

boston.com THIS STORY HAS BEEN FORMATTED FOR EASY PRINTING

RANGERS 6, RED SOX 5 The Boston Globe

Same, sad Foulke song

Red Sox closer gives it up to Rangers in ninth

By Nick Cafardo, Globe Staff | July 5, 2005

ARLINGTON, Texas -- Keith Foulke says when he's out in public people call him Kevin (Faulk). Maybe that's because the diminutive Patriots running back might be a better choice right now to close games for the Red Sox.

Yet another Foulke disaster occurred at steamy Ameriquest Field last night as he blew a 5-4 lead in the ninth, allowing the winning run on Kevin Mench's first-pitch RBI single to left with the bases loaded in a horrifying 6-5 loss to the Texas Rangers.

Foulke's teammates and manager Terry Francona continue to support the embattled closer, but how long can they remain patient?

"He's going to get through one of these and that's going to be the end of it," said Sox captain and catcher Jason Varitek. "He's done it over and over in this situation and it just didn't happen [last night]."

Foulke's reduced velocity has been much publicized lately, but last night it seemed all right. On one pitch he hit 89 miles per hour. Varitek remains unconcerned about velocity.

"It's [about] getting comfortable and getting into his zone, because he's a guy who can pinpoint," Varitek said. "He's got three gears on his changeup. It's been a tough time for him. There's no plainer thing to say than to say we need him. We ain't going anywhere without him. Period. We've got to keep getting him out there and it's going to happen for him."

Varitek went so far as to say, "I think he keeps getting better. He's just not able to get away with any mistakes. People see blood. I just think he's going to get through one of these and that's

Discuss | Subscribe | Archives

Late fireworks spark Rangers' 6-5 win

09:35 AM CDT on Tuesday, July 5, 2005

By BEN SHPIGEL / The Dallas Morning News

ARLINGTON – A week ago, this wouldn't have happened.

Nope. Not with the Rangers mired in a tailspin, losing games near and far, in blowouts and nail-biters.

But this is a resilient bunch. The Rangers have won five of their last six, including Monday night's 6-5 win against the Red Sox that provided a joyful ending to what has been quite an interesting first half.

Despite leaving 11 men on base – and twice leaving the bases loaded – the Rangers came back from a 5-3, eighth-inning deficit thanks to some clutch hitting by a pair of All-Stars and a potential reserve who received some passionate stumping from manager Buck Showalter.

It marked only their second victory in 35 chances when trailing after eight innings, and, shocker of shockers, they didn't even hit a home run.

Here's how the rally unfolded:

Tuesday, 7:05 p.m.
TV: Ch. 27
Radio: KRLD-AM 1080; KFLC-AM 1270 (Spanish)
Tuesday's scouting report
Weather
About the Rangers

File

Rangers/MLB

SportsSay: The blog of SportsDay

Rangers 6, Boston 5

Figure 2-3: An example of a pair of local newspaper articles (online versions) for parallel corpus.

Hence, I collected articles published on several online news sources. The MLB website (www.mlb.com) publishes two articles for every game, written for each of the two teams in the game. Each team has a unique sportswriter covering that team for the entire season, so I use the MLB articles as one of our sources with the home/visit team perspective. The ESPN website (www.espn.com) also has articles for every MLB game including the main summary articles from the Associated Press (AP). I use the AP articles as our neutral source. I also collected online local newspaper articles for MLB teams in the American League East Division: Boston Red Sox (The Boston Globe at www.boston.com), New York Yankees (The New York Times at www.nytimes.com), Baltimore Orioles (The Washington Post at www.washingtonpost.com), Toronto Blue Jays (The Toronto Star at torontostar.com), and Tampa Bay Devil Rays (The Tampa Tribune at tampatrib.com). See figure 2-3 for an example of a pair of local newspaper articles (online versions) on the same game.

2.3.2 Alignment

While the game logs are simple to parse into feature vectors representing baseball events and entities, newspaper articles are much harder to analyze. In order to make connections between an article and the domain model of the game built from the game logs, the sentences must be aligned with the game event feature-vectors derived from the game log. For example, a paragraph below describes events in the game, and the sentences in the paragraph can be aligned to the *at-bats* in the game.

> Podsednik started the three-run 10th inning by drawing a leadoff walk from reliever Ambiorix Burgos (2-4). Podsednik moved to second on Burgos' balk, the third of the series for the Royals (37-69), and scored on Crede's well-placed grounder past the pitcher. Ross Gload doubled home a run and Brian Anderson singled home an insurance tally, making Angel Berroa's home run off of closer Bobby Jenks (29th save) in the ninth nothing more than statistical padding.

There is previous work on sentence-to-game event alignment, most notably by Snyder [35] who uses statistical learning algorithms on American football data to achieve successful alignment results. I use a much simpler technique of tagging and keyword-based matching. The articles were first tagged with player names and part-of-speech tags, and simple pattern matching heuristics were used to automatically align the sentences in the articles with game events. The player names were extracted from the entity model of the baseball domain model, and the POS tagging was done with the Stanford POS tagger [39]. Pattern matching heuristics looked for co-occurrences of tags and words within a certain window (e.g., {player} AND "homerun" within 3 words), and the results from applying those heuristics were aligned with the *at-bat* feature vectors computed from the game log. Testing on 45 hand-annotated articles, I achieved a precision of 79.0% and recall of 79.2% for alignment. The average number of *at-bats* in those hand-annotated articles was 8.

Figure 2-4 shows an example of how the sentences are aligned to feature vectors.

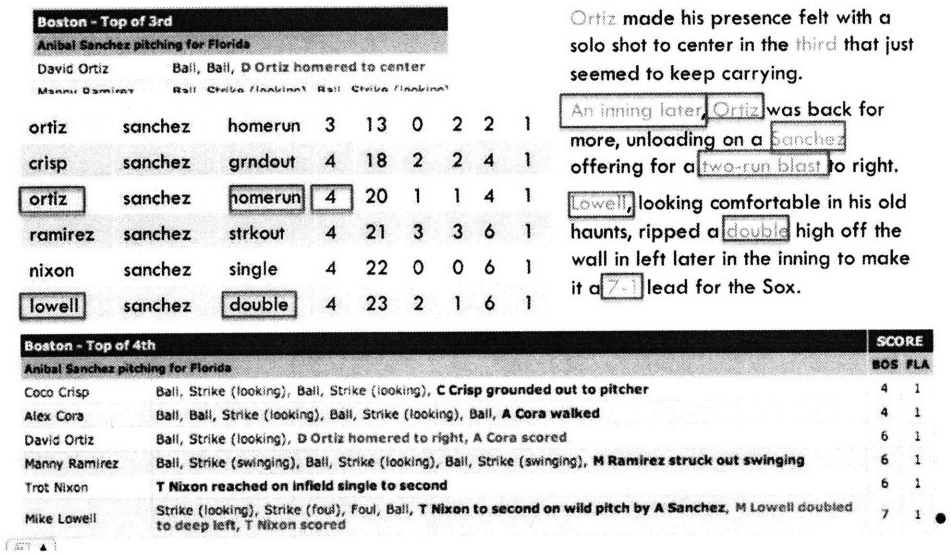


Figure 2-4: An example of sentence to feature-vector alignment.

2.3.3 Content Overlap Analysis

In trying to discover the differences among the local team articles and AP article for the same game, I looked at the overlap of content among the articles. The percentage of overlapping content varies widely, mostly due to the way the games unfolded. For example, many games are one-sided where one team simply dominates, and there are just not enough events that are positive for the losing team. For those games, the losing team's newspaper merely reports the result of the game without describing the events of the game in detail. However, many games are close in score and number of hits, and for those games I found a high overlap of content among all three articles. Table 2.1 lists the number of *atBats* reported in common for a local article and the AP article for the same game, averaged over 20 article-pairs. The first column shows the percentage of *atBats* that are mentioned only in the AP article, the second column shows the percentage of *atBats* that are mentioned only in the local article, and the third column shows the percentage of content mentioned in both articles. Repeated occurrences of the same *atBat* was counted only once.

	AP	Local	AP, Local
Globe	15.5	23.3	72.4
NYTimes	13.7	19.2	78.2
WashTimes	18.2	15.5	80.3
MLB Red Sox	12.4	18.2	82.4
MLB NYY	14.4	18.7	80.3

Table 2.1: Percentage of non-overlapping and overlapping content between local and AP articles.

Chapter 3

User Study I

This chapter presents a user study that was carried out in order to verify the definition of the problem as discussed in the previous section. The overall goal of this research is to model perspective in game summaries and build a prototype system that can automatically generate summaries from multiple perspectives. However, since that problem is much too broad, to make the problem more tractable, section 1.4 proposed a hypothesis that would justify solving a sub-problem that is much narrower in scope. The motivating factor for this hypothesis was the observation that, in the parallel corpus of neutral and local team perspective articles, much of the game event content overlaps among the three different perspective articles, but the content seems to be organized in different ways. This user study was designed to test the following hypothesis:

> **Ordering Hypothesis:** Ordering of the content alone contributes significantly to the perspective of a story. Hence, you can generate multiple perspectives by taking a pre-determined set of content and reordering it.

To test this hypothesis, this user study takes the parallel articles from the corpus and modifies them in stages such that the modified versions would reveal whether

1. original articles are judged to have different perspective
2. modified articles with only content aligned with the domain model retain the different perspectives of the original articles

3. modified articles with only overlapping content retain the different perspectives of the original articles

4. modified articles that use the same surface realization (sentence structure, lexicalization) retain the different perspectives of the original articles

To cut down on the number of articles per subject, steps 3 and 4 were combined in this study. If this had produced results that showed that different perspectives were no longer preserved, I would have separated out the two steps to see where the loss had occurred, but since the results were positive, I can assume that combining steps 3 and 4 did not lose the study's effectiveness in testing the ordering hypothesis.

Section 3.2 elaborates the steps 1 through 4 above and shows how the articles are modified to test the ordering hypothesis.

3.1 Setup

Since the study is fairly simple in design and requires no special hardware or instructions, I conducted the entire study remotely through the Internet. The study was approved by the Committee on Use of Humans as Experimental Subjects (COUHES) at MIT. The wording of the questions was internally reviewed to prevent confusions and confounding variables.

3.1.1 Web-Based Survey

The study was all done through a web-based survey using CGI scripts. Scripts were run off of the MIT CSAIL web servers, and the answers input by the subjects were automatically recorded into text files. The surveys were not timed, and subjects were told so, but the time information was automatically collected through the CGI scripts. The time information was not used in analyzing the results, and there was no significant variation among different subjects or article types in the task completion times.

3.1.2 Participants

Eight subjects participated in the study using twelve games. They were recruited through an email list used primarily for voluntary user studies. They were all MIT students and researchers, ages 18 and up, native speakers of English, who watch major league baseball (MLB) at least once in a season. They were asked for their favorite MLB teams, but that information was not used for analysis because the surveys were made up of games of a variety of teams, and being a fan of one team did not make significant differences in the perspective judgments. Subjects were paid ten dollars in cash or online shopping gift certificate. There were four women and four men.

3.2 Articles and Conditions in the Survey

For all four conditions, subjects were asked to rate each article on a scale of 1 to 5, where 1 is strongly Team A perspective, 3 is neutral, and 5 is strongly Team B perspective. For exact wording of the survey as well as the original and modified articles used, see Appendix A.

The games were chosen from our MLB database of games such that various teams are represented, and various game outcomes are represented. Hence, games that are one-sided, as well as close games, extra-inning games, and games with major milestones (e.g., the starting pitcher's winning streak) are included and randomly assigned order in the web-based survey.

For each game, the three original articles are from the two opposing teams' local newspapers (online editions) and the Associated Press (AP) article as published on http://www.espn.com. Although the articles are modified and presented here in the order from the original articles to the fully modified (overlapping content) articles, they appear in random ordering in the user surveys. This is because the users may read the original article, remember parts of it, and be affected by the perspective of that article when rating the perspective of the modified version.

3.2.1 Confirming Local Team Perspectives

The baseline condition is the comparison between the perspectives of the original articles. Since there is not a good way to define what perspective is, I take an approach that, measuring what the users perceive is a good way to quantify perspective. When a user reads a baseball summary article and says that it seems to have been written from Team A's perspective, then I assume that article was written from Team A's perspective, and I average those numbers across all the subjects to get a measure of perspective of an article. Of course, there is also the source information, so I can simply assume, without user testing, that articles from Team A's local newspaper was written from Team A's perspective. That is the assumption used for collecting the corpus. Here, I am using the first part of the user study to ensure that assumption is valid, and confirm that our user study design draws out valid ratings from the subjects, and at the same time, come up with a quantitative metric for perspective. So, to confirm that the home team and the visit team perspectives of the local team articles are correctly perceived, I simply presented the AP and local newspaper articles to subjects and asked them which team the articles were written for.

Here is an excerpt from an original version of the summary article.

> Schilling was again beset by the long ball in the third, and this time it was Crawford putting a solo shot over the wall in right to make it 2-0. The Red Sox cut that lead in half with yet another mammoth homer from Ortiz, whose towering shot sailed over the wall in right in the fourth.
>
> "We righted the ship and we did some things, and David continues to be the best hitter in the game," Schilling said.
>
> The Red Sox tied it in the sixth when Kevin Youkilis lofted a sacrifice fly to the warning track in left.
>
> Schilling found himself in a sizable mess in the bottom of the sixth with the bases loaded and just one out. One of those hits was an infield single by Travis Lee that bruised Schilling on the right hand when he tried to barehand it. But then he got fired up, striking out B.J. Upton and Tomas

> Perez on 96-mph heaters to end the inning. In as demonstrative a moment as Schilling has had all year, he wildly pumped his right fist as he walked off the mound.

This excerpt will be used in the next two sections to illustrate how it would be modified for the other two conditions.

3.2.2 Aligned Content

An intermediate stage between the baseline condition and the final testing condition is the "aligned content" condition, where the original articles are modified such that they contain only sentences that describe the game events (*at-bats*). That is, player quotes, commentary about the team or players' historical performances, and any financial or personal news were removed from the articles. This condition tests whether the game event-aligned content alone is enough to deliver the same perspective as the original article. This is an important step because the original articles do contain a substantial amount of player quotes and other extra-game information. In future research, I may try to incorporate the extra-game information as well, but that requires either adding onto the domain model to include events outside of the game or adding to the system the capability to analyze and generate sentences that are not aligned with the domain model.

Here is one example of how the excerpt from the previous section would be modified in this condition by discarding the sentences that are not aligned with the events in the domain model

> Schilling was again beset by the long ball in the third, and this time it was Crawford putting a solo shot over the wall in right to make it 2-0. The Red Sox cut that lead in half with yet another mammoth homer from Ortiz, whose towering shot sailed over the wall in right in the fourth.
>
> The Red Sox tied it in the sixth when Kevin Youkilis lofted a sacrifice fly to the warning track in left.

> Schilling found himself in a sizable mess in the bottom of the sixth with the bases loaded and just one out. One of those hits was an infield single by Travis Lee that bruised Schilling on the right hand when he tried to barehand it. But then he got fired up, striking out B.J. Upton and Tomas Perez on 96-mph heaters to end the inning.

The quote in the second paragraph of the original excerpt was removed, and the last sentence of the excerpt was also removed because it does not align with any game event. However, phrases such as "whose towering shot sailed over the wall in right" at the end of the first paragraph was left in even though that information is not available from the automatically built domain model. This is because discarding extra-game information was done at the sentence level, so if any part of the sentence aligned with the game events, then the entire sentence was left in the modified article.

3.2.3 Overlapping Content

In the last condition, the article from the second condition is further modified in two steps to produce the final article to test the validity of the ordering hypothesis. The ordering hypothesis says that same content can carry different perspectives depending on how it is arranged and ordered. Hence, the first step is to make the content of the three articles the same by keeping only the sentences that are about the same game events. If there are sentences that are aligned with game events that appear only in that article, then those sentences are discarded.

In the second step, I replaced all the sentences with slot-filling templates, such that all the articles shared the same surface form of sentences. This means that the only difference among the three articles is the ordering of the content. Here is an example of how the final modifications are made to the excerpt in the previous section.

> Crawford (TOR) hit a one-run home run in the third inning to make it 2-0. Ortiz (BOS) hit a one-run home run in the fourth inning to make it 2-1.

Youkilis (BOS) hit a one-run sacrifice fly in the sixth to make it 2-2.

Lee (TOR) hit a single in the sixth. Bases loaded, Schilling (BOS) struck out Upton (TOR). Bases loaded, Schilling (BOS) struck out Perez (TOR) to end the inning.

It is worth discussing how to design the templates for the user survey as well as for the system that would generate the sumary articles. For this survey, I hand-crafted the templates carefully such that the sentences are the same across the three different perspectives. They may not be the same templates across different games because it was not clear, at this point, how best to design the templates, but for the purposes of this survey, it is only important that the three perspective articles (Team A, Team B, and AP) use the same surface form, so that the subjects' perspective ratings can be compared across the three perspectives. Section 5.4 discusses in more detail how the templates were crafted for the system that generates the summaries, and although the templates used for this user study were slightly different in the words and sentence structures used, the basic ideas are the same.

3.3 Results and Discussions

Tables 3.1 and 3.1 show the average perspective ratings over the eight subjects for Games 1, 2, and 3. Each column is the intended perspective of the original and modified articles, and each row represents one of the three conditions described in 3.2. To prove the validity of the Ordering Hypothesis, we are looking for perspective ratings to be significantly different, with Team A column being close to 1, AP column being close to 3, and Team B column being close to 5. As shown in the table 3.2, the average perspective ratings do show that trend, and for games 1 and 2, the ratings are significantly different for the three perspectives, for all conditions, as analyzed using ANOVA at $p < 0.05$.

The results for Game 4, however, show the same trend, but the ratings are not significantly different across the three perspectives. The perspective ratings for the

Game 1	Team A	AP	Team B
Original	1.38	3.50	4.63
Aligned Content	1.50	3.25	4.38
AP Content	1.75	3.13	4.00
Game 2	Team A	AP	Team B
Original	2.25	2.75	3.75
Aligned Content	2.38	3.63	3.88
AP Content	2.50	3.38	3.63

Table 3.1: Perspective ratings, averaged over eight subjects, for Games 1 and 2. Columns are the intended perspectives of the original articles, and rows are the modifications made for each condition. ANOVA results show significant difference among the three perspectives, at level $p < 0.05$.

Game 3	Team A	AP	Team B
Original	1.38	3.50	4.63
Aligned Content	1.50	3.25	4.38
AP Content	1.75	3.13	4.00

Table 3.2: Perspective ratings, averaged over eight subjects, for Game 3. The ANOVA results do not show significant differences in perspective ratings for the last condition.

AP article and Team B article are not much different, especially for the last condition, AP Content, where we have just the ordering information preserved. This is because the content in the original AP article is decidedly one-sided. In cases where the game is a significant one-sided win, there is usually very little, in terms of the game events, to talk about for the losing team. Thus, taking the same content and reordering contributes some, but not much, to generating the desired perspective. For this type of games, the factors that contributed to the Team B perspective in the original condition is most likely the way the events are interpreted, shown by player and coaching staff quotes, or extra-game events and issues, such as what the team has to do to win next time. These one-sided games happen in about quarter of the total games, so it is not an insignificant portion, but the others, for which the Ordering Hypothesis is valid, make up 75% of the games. Moreover, as the results show in table 3.2, the difference in perspective, although not statistically significant, does carry over to the last condition.

To sum up the results of the user survey, although one game out of five did not show significantly different perspective ratings for the final condition, four of five games did. This shows that the Ordering Hypothesis is valid, because in the final condition, the articles were modified using the same content as in the neutral article, replacing the sentences with canned templates, and preserving only the content ordering information from the original perspective articles.

Chapter 4

Reordering Algorithm

This chapter presents the reordering algorithm, the driving force behind the multi-perspective generation system described in Chapter 5. After a short summary of the background on ordering algorithms, the chapter is divided into two sections. The first section discusses ordering strategies that were identified in the parallel corpus described in section 2.3. The second half of this chapter discusses choosing the best ordering strategies given a desired perspective, where the optimal choice is one that maximizes the sum of weights learned from the corpus.

4.1 Background

Content ordering is a well-studied problem within natural language generation. It assumes that content selection has been already done, and the problem is selecting the optimal ordering of the content such that the resulting text is coherent and easy-to-read. Content ordering is part of many applications such as spoken dialogs [36] and multi-document summarization [4]. In the earlier works, planning was the method of choice (cf. [8], [26]). Recent trends have turned toward methods that learn ordering constraints from the corpus (cf. [12], [20]). In all of these applications, however, they look for the single most effective ordering for delivering the content in an easy-to-read and accurate way. The ordering problem here in this thesis differs in that the solution should be more than one ordering, and those orderings must make the generated text

to exhibit different perspectives.

4.2 Ordering Strategies

As the first step in trying to solve the ordering problem, I looked to the parallel corpus described in section 2.3 to discover what kind of ordering strategies were used by the journalists. Before discussing each of the ordering strategies, let us look at an example of the same content in different ordering. Here are excerpts from two articles written on the same game.

> Damon helped create the winning run after reaching base on a one-out single against closer Chris Ray in the ninth. With two outs, Damon tried to steal second, and it appeared as though catcher Ramon Hernandez had thrown him out, which was the call made by umpire Lance Barksdale.
>
> But the ball popped out of second baseman Brian Roberts' glove, and after Damon pointed that out to the ump, he was called safe, giving Jeter a chance to come through.
>
> Jeter, who hit third for the first time since Sept. 28, 2003, had already driven in the go-ahead run in the seventh, only to watch the Yankees' bullpen give up the lead in the eighth.
>
> Melvin Mora hit a hard grounder to third, where Miguel Cairo, playing in place of A-Rod, had trouble controlling the ball. Cairo picked it up and fired to first for the second out, but the tying run came home on the play.
>
> The Yankees regained the lead in the seventh, as Jeter singled in the go-ahead run against Todd Williams. Melky Cabrera later scored on a Kurt Birkins wild pitch, giving New York a two-run lead.
>
> Ron Villone threw a scoreless seventh, but after allowing a one-out hit by Lopez in the eighth, he was pulled in favor of Scott Erickson. Erickson hit Jeff Conine to put the tying run on base, then served up a double to Luis Matos, scoring Lopez.

Torre intentionally walked Roberts to load the bases, bringing in Farnsworth to face Mora. Cairo's misplay on the grounder tied the game, but after the Yanks intentionally walked Miguel Tejada, Farnsworth came back and retired Hernandez for the third out.

Damon and Jeter took it from there, combining to produce the go-ahead run in the ninth. Farnsworth made it stand up, retiring Baltimore in order to close out the game.

– From the MLB New York Yankees, June 2, 2006

The game turned with two outs in the ninth inning, when Baltimore catcher Ramon Hernandez made a perfect throw to nab Johnny Damon on an attempted steal. Second baseman Brian Roberts put the tag down, but Damon slid and knocked the ball from his glove. Damon was ruled out – then safe – and wound up scoring the decisive run in a 6-5 win.

New York's Derek Jeter wound up driving Damon in with a soft single to right field, but the steal grabbed most of the attention. Second-base umpire Lance Barksdale made an emphatic out call, but things changed on the tag attempt. When the second baseman brought his glove up, without the ball, the umpire changed his mind.

Baltimore closer Chris Ray was walking off the mound after the throw, but he had to go back to face Jeter. The right-hander, who's still 14-for-14 in save opportunities, said Jeter beat him with a good piece of hitting.

Jeter did the same thing in the seventh, when he hit a single to break a 3-3 tie. New York (32-21) scored another run on a wild pitch from Kurt Birkins, but the Orioles forced a tie game in the eighth. With one on and one out, the Yanks went to Scott Erickson, who hit a batter and gave up a run-scoring single. The O's tied the game on a ground ball to third base.

Shortstop Miguel Tejada singled twice in the early innings and scored Baltimore's first two runs. The Orioles (25-30) didn't score again until

the sixth, and they did it with small ball. Corey Patterson dropped a two-out bunt up the first-base line and reached on an error. Patterson stole second on a pitchout and scored on a subtle single up the middle.

The Yanks did their early damage via the long ball. Andy Phillips cracked a solo shot in the fifth to put them on the board, and one inning later, Jason Giambi gave the road team a brief one-run lead with a two-run blast over the right-field scoreboard. Giambi hit a long foul right before his homer, and Baltimore starter Kris Benson left after Patterson tied the game.

– From the MLB Baltimore Orioles, June 2, 2006

The two excerpts are from the MLB website (http://www.mlb.com), where for each game, a team journalist from each of the two teams writes a story for that team site. Comparing the two excerpts, several differences are visible. First, as noted in section 2.3.3, much of the content overlaps between the two. This is true even though the two articles are from the two local team perspectives. Ordering of the content, however, differs quite a bit, thus again confirming the hypothesis that ordering is a significant factor in deciding perspective of an article. One thing to note about ordering is that the same event (e.g., Damon's steal) appears in different contexts. That observation is key to identifying the ordering strategies below.

4.2.1 Feature-Based Content Ordering Strategies

In the following sections, I will illustrate several ordering strategies found in the corpus. The strategies are for a segment of an article in the corpus, usually spanning one or more paragraphs. The segments themselves must be ordered, and that will be discussed in 4.2.2.

Chronological Ordering

An ordering strategy based simply on the chronological ordering of events is the easiest for the baseball domain where the chronology of events (*atBats*) is clearly defined.

Barzilay, et al. [5] has found that chronological ordering works well for multidocument summarization where the article is mostly event-based. Ironically, a purely chronological ordering strategy is not used very frequently, even if we ignore repeated content where the most important events are mentioned at the very beginning and/or end of the article and consider only the middle portion of the article. In our corpus, only two out of ten, on average, articles show a chronological ordering. The following excerpt is an example of such an article where a portion of the article is purely chronological and is counted as an instance of the chronological ordering strategy.

> Wakefield allowed just a hit and a walk through three innings before falling behind 3-1 in the fourth. The Yankees loaded the bases with no outs on a single by Derek Jeter and walks to Jason Giambi and Alex Rodriguez.
>
> Hideki Matsui's groundout to Youkilis drove in one run and Robinson Cano singled in two.
>
> Boston tied it in the fourth when Alex Cora and Youkilis singled before Loretta bunted into a forceout at third. Ortiz loaded the bases with a single before Manny Ramirez singled in one run and Nixon tied it with an RBI groundout to first.
>
> –From espn.com article on Boston vs New York game on May 1, 2006

A clear advantage of using this strategy is that, given the set of events to be included in the article and the description of the game in the form of inning-by-inning game log, the ordering of the content is trivial.

Inning-Based Ordering

A subset of the chronological ordering strategy is an inning-based ordering strategy where a set of events from the same inning are grouped together into one segment, and those events within the segment are chronologically ordered. The following is an excerpt that shows the inning-based ordering strategy.

> The Rays fought back in the seventh with consecutive singles by Toby Hall, Aubrey Huff and Damon Hollins to cut the lead to 6-2 before Lee's sacrifice fly made it 6-3. Gathright then hit one off the left-field wall for what appeared to be an RBI double, but Hollins fell rounding third base and did not score.
>
> With runners on second and third and two out, Julio Lugo stepped to the plate and hit a ball deep to left field that looked like a three-run homer over the Green Monster. Unfortunately for the Rays, the ball took a hard left just before reaching the foul pole. Lugo then struck out to end the inning.
>
> –From the mlb.com Tampa Bay Devil Rays article on May 26, 2006

In an inning-based ordering strategy, not all $at-bats$ need to be included in the text, but usually the $at-bats$ that advance the baserunners and those that add to the score are included. To organize the content in this strategy, simply look at the value of the *inningNum* feature of events, then group those that have the same *inningNum* value. This is a very frequent strategy, as the corpus shows about eight out of ten articles in which a portion of the article uses the inning-based strategy.

Player-Based Ordering

Another frequently used ordering strategy is based on entities, such as a single player or a set of players (e.g., relief pitchers). Sometimes, although rare in a non-commentary article, an entire story is based around a single player and his performance in the game. A monumental milestone, such as breaking the most number of homeruns by any player in history, may elicit several articles, regardless of team perspective, on a single player. Here is a more commonplace example of player-based ordering strategy.

> The two moonshots that Glaus provided paved the way for Toronto's latest victory. The third baseman's second shot – his eighth of the season

– came on a 2-0 offering from Baltimore reliever LaTroy Hawkins. With Toronto trailing 3-2, Glaus sent the pitch over the 25-foot wall in right field for a three-run blast.

His 4-for-5 showing also helped snap a two-week long slump. Entering Monday's game, Glaus had hit just .154 since April 18 and saw his average drop from .348 to .259. The solo homer that he hit off Baltimore starter Erik Bedard in the second inning was his first in 10 days.

"I got a couple pitches up and I was able to take advantage of it," Glaus said. "It's just one of those days. I was able to find some holes and hit some balls on the barrel, which was nice. It's been a while."

Glaus came a few feet shy of having three home runs in the sixth inning, when he sent another pitch from Bedard off the wall in left-center field. He later scored when Shea Hillenbrand chipped in an RBI single. Glaus added a ground-rule double in the ninth, and finished the day with four RBIs. –From the mlb.com Toronto Blue Jays article on May 1, 2006

This strategy requires looking at a few different features because a player may have been involved in an *atBat* event as a *batter*, *pitcher*, or as a baserunner (features *onFirst*, *onSecond*, and *onThird*). In the example above, Toronto player Glaus is the *batter* in most of the *atBats*, but in the sentence "He later scored when Shea Hillenbrand chipped in an RBI single", Hillenbrand is the *batter* and Glaus is a baserunner.

PlayType-Based Ordering

Another frequently used strategy is based on types of play, or, the result of the *atBat*. For example, all the homeruns in a game may be grouped together into a paragraph. Some play types, such as *doubles* or *double plays* are more frequently used to group the content together, rather than others such as *foul-outs* or *singles*. The following excerpt shows the play type *doubles* being used as an ordering strategy. Often, the

play type feature is used in conjunction with the *team* or other entity-based feature. Thus, all the *strikeouts* by one team or one pitcher may be grouped together.

> All four runs off Josh Towers came on run-scoring doubles by Mike Lowell, Adam Stern and Kevin Youkilis. –From the mlb.com Toronto Blue Jays article on April 11, 2006

This strategy is rather simple to identify, as only the *playType*, and sometimes *team* and *player* features are used as the grouping feature.

Scoring-Based Ordering

One of the most important events in a baseball game are *atBats* that add score. A run-scoring *atBat* happens when the batter and/or a baserunner safely reaches the home plate on a hit, a walk, or a defensive error. There are hits, walks, and defensive errors that do not result in runs scored, so a run-scoring play is rarer in a game than a non-run-scoring play, and those scoring plays deserve extra attention. Furthermore, the runs scored determine the winning team and the losing team, so keeping track of when and how the runs are scored is important in describing the events of the game to the readers of the article.

> Aside from Sexson's home run in the sixth, Johjima – who had the first three-hit game of his career – had an RBI double in the inning. Betancourt bounced a two-run double into left field and Jose Lopez later added an RBI triple. –From the mlb.com Seattle Mariners article on May 1, 2006

Team-Based Ordering

Another entity-based ordering strategy is using the *team* feature. This is not very discriminative, since there are only two teams in the game. However, it is most often used with another grouping feature, such as *scoring*. There are often extra-game features, such as the team's overall performance or rank in the division, but since this thesis does not deal with extra-game features, that type of content is discarded and not considered in ordering decisions.

The Yankees regained the lead in the seventh, as Jeter singled in the go-ahead run against Todd Williams. Melky Cabrera later scored on a Kurt Birkins wild pitch, giving New York a two-run lead. –From the mlb.com New York Yankees article on June 2, 2006

4.2.2 Learning and Using Grouping Features

Table 4.1 shows an illustration of the different ordering strategies used in grouping feature vectors. The actual feature vectors have many more features, but for illustrative purposes, only a subset is shown here. Recall that the feature vectors are sentences in an article in the parallel corpus aligned with the domain model for the game associated with the particular article. Since there are many features and ordering strategies, the reordering algorithm needs to identify the features to use for assigning the *atBats* to appear in the same segment. I used a simple counting of most frequent feature values of the corpus to derive these features. This comes from the intuition that the players whose names appear most frequently in the articles for a local newspaper tend to be important topics for those stories. So I aggregate all the local team articles and rank the feature values including pitcher and batter names and play types (e.g., homerun, single, strikeout). To turn a neutral article into a local perspective article, I take the *atBats* that should appear in the article, look at the feature values that are shared among them, and find the highest-ranked feature value for that team. Any remaining *atBats* are arranged in chronological order.

Once the features are learned from the parallel corpus, they are used to find the optimal set of ordering strategies for the entire article in the following way.

- For every possible grouping of content (feature vectors)
    1. For each group
    2. Compute the score $w_f n$
        - w_f is weight computed for feature f that is shared by all the vectors in a group

Batter	Pitcher	Play	Inning	AtBat	onFirst	Runs	Outs
Ramirez	Sanchez	**homerun**	1	3	Cora	1	1
Ramirez	Sanchez	**homerun**	6	48	Ortiz	1	0
Ortiz	Sanchez	homerun	3	14	none	1	0
Ortiz	Sanchez	homerun	4	20	Cora	1	1
Ortiz	Sanchez	double	6	47	Cora	1	0
Varitek	Sanchez	foulout	**4**	21	none	0	1
Nixon	Sanchez	single	**4**	22	none	0	2
Lowell	Sanchez	double	**4**	23	Nixon	1	2

Table 4.1: An illustration of how different features are used for grouping content. The features in boldface are the grouping features for that group of feature vectors. Each set of feature vectors delimited by double horizontal lines represent content in that paragraph.

Batter	Pitcher	Play	Inning	AtBat	onFirst	Runs	Outs
Ramirez	Sanchez	**homerun**	1	3	Cora	1	1
Ortiz	Sanchez	**homerun**	3	14	none	1	0
Varitek	Sanchez	foulout	**4**	21	none	0	1
Nixon	Sanchez	single	**4**	22	none	0	2
Lowell	Sanchez	double	**4**	23	Nixon	1	2
Ortiz	Sanchez	homerun	**4**	20	Cora	1	1
Ramirez	Sanchez	homerun	**6**	48	Ortiz	1	0
Ortiz	Sanchez	double	**6**	47	Cora	1	0

Table 4.2: An example showing how different feature weights would change the groupings. The feature vectors from the previous table are rearranged in this table such that grouping features are chosen differently.

- n is the number of feature vectors in that group

3. Do this for all features shared by the feature vectors in the group

- Repeat for all possible groupings

- Find the grouping that maximizes the score $w_f n$

Because of the grouping feature weights that are learned for each desired team's perspective, the same set of feature vectors are fed into the algorithm above and assigned a different optimal ordering. This is how the system would output different orderings based on the perspective that it wishes to generate. Table 4.2 illustrates how

the same feature vectors in table 4.1 are rearranged according to different grouping features.

Chapter 5

Multiple Perspective Generation System

The previous chapter presented the algorithm for reordering the content feature vectors. This chapter describes the entire system for generating baseball summaries from multiple perspectives.

5.1 System Overview

Like many other NLG systems, the multi-perspective generation system consists of three main parts: content selection, content ordering, and surface realization. Figure 5-1 shows graphically how the system takes an AP article and a game description, then generates the two perspective articles. On the top left is the AP article, and on the right side of it is the game description parsed into feature vectors. The sentences in the AP article are aligned with the feature vectors to produce a set of feature vectors to be used as content for the perspective articles. This part completes the first step, content selection.

For the second part of the system, the reordering algorithm described in Chapter 4 is used for content planning.

Finally, for the third part of the system, the reordered content, represented as feature vectors, is turned into a set of sentences through the template-based surface

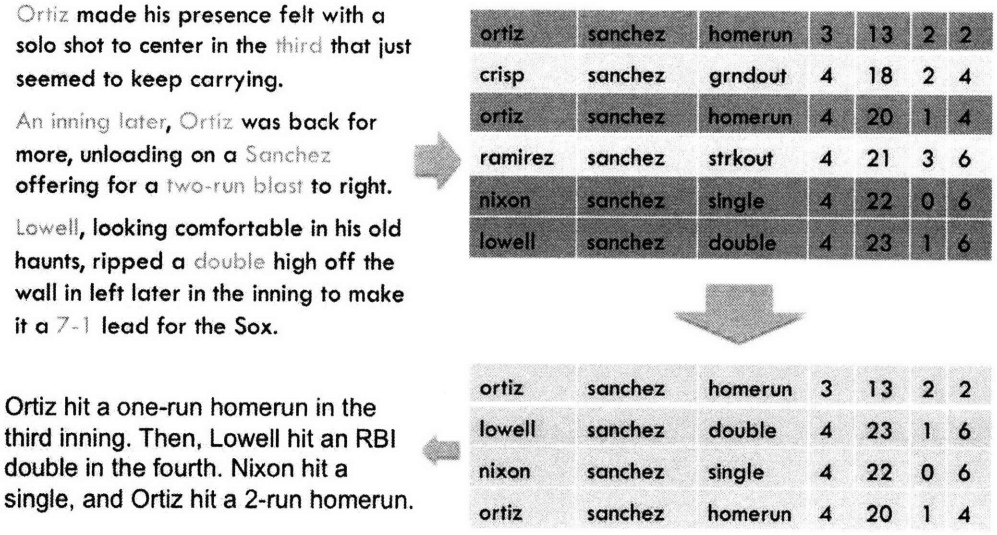


Figure 5-1: An overview of the generation system.

realization described below in section 5.4.

5.2 Content Selection

Since the Ordering Hypothesis, validated using the user study in Chapter 3, states that neutral content can be turned into an article with a local team perspective just by reordering, the system just takes the neutral content (from AP article) for content selection. However, I ran a small experiment to see whether content selection can be automatically done using feature vectors if a neutral article was not in the corpus.

I tagged ten MLB games using AP (Associated Press) articles. For each *atBat* represented as a feature vector, if it is mentioned in the article, it is tagged as *significant*. I used a supervised learning approach using Weka[42], a Java-based tool for experimenting with various machine learning algorithms for inferring whether a feature vector should be tagged *significant* or not. A ten-fold cross-validation on the ten games using the alternating decision tree (ADTree) gave an overall accuracy of 92.8%. Using the confusion matrix, recall and precision on retrieving the *significant* plays were 85.5% and 82.4%, respectively. I used the standard recall and precision definitions in the information retrieval community, where recall measures how much

of the ground-truth the system was able to retrieve, and precision measures how accurate were the retrieved items. The precision and recall numbers are low relative to accuracy because there were many *non-significant* events that are considered only in the accuracy numbers.

This relatively simple experiment shows that content selection for picking out neutral content (as in AP article) is not difficult. This means that, even if the system does not have neutral articles to start with, it can first to content selection based on supervised learning on the features, then use that content as if it came from a neutral article. Since, for this thesis, the corpus already has neutral (AP) articles, the system simply uses the neutral articles for content selection. Besides simplifying the content selection process, this has the added advantage that accuracy and completeness of content is somewhat guaranteed, as much as one can expect from AP articles.

5.3 Content Organization and Ordering

Content organization and ordering is a step in natural language generation that takes a predetermined set of contents, organizes them into paragraphs, and orders them. In a recent book that presents a comprehensive overview of NLG [32], Reiter and Dale discuss *document structuring*, the part of an NLG system that organizes and orders content, as an essential part of NLG for making multisentential text readable and understandable. Content organization and ordering has been studied in the context of many NLP applications including summarization [24] and concept-to-text generation [21]. These and other previous work on content organization and ordering have focused mainly on readability of the generated output, and thus much of the effort has been on discourse and sentence-to-sentence coherence. In contrast, in this thesis, content organization and ordering is used as a means to induce a certain perspective of the generated output rather than focusing on coherence. However, it is shown in (add citation) that there are multiple ways to order content that result in coherent text, so producing multiple orderings to induce different perspectives does not mean that coherence or readability of the text must be sacrificed.

Another major difference of this work is that the content organization and ordering is done globally, whereas much of the previous content ordering research focuses on looking at sentences one by one. One instance of this is to model the ordering problem as a pairwise decision between sentence A and sentence B, deciding whether $A << B$ or $B << A$, where $x << y$ means that x should come before y. Another way to model the problem is to look at it as a Markov process, that is, figuring out what the n^{th} sentence S_n should be, given the $n-1^{th}$ sentence S_{n-1}. Although these two simplified approaches of the problem are used often, they make strong independence assumptions, and the models do not capture interactions that involve more than two sentences. For example, the presence of sentence C may reverse the ordering of sentences A and B, such that $A << B$ is the preferred order over $B << A$, but $B << C << A$ is preferred over $A << C << B$. The organization and ordering algorithm used here takes the entire set of sentences in the summary article and determines the ordering for all the sentences.

Ordering of the content is explained in three steps. First, different ordering strategies found in the corpus are described with examples of each strategy. Then, using statistics from the corpus, the different strategies are compared in the way they contribute to perspective. Lastly, the algorithm for using the ordering strategies to produce multiple perspectives is described.

5.4 Surface Realization Using Templates

Once the content is organized and ordered, the feature vectors must be turned into sentences. There are two steps to do this. The first is slot-filling templates, and the second is simple aggregation. Templates are use in many NLG systems, as they are simple to build requiring very little expertise. They are relatively difficult to maintain if new types of sentences must be added to the system, so many of the large-scale NLG systems use rule-based realizer [13] or stochastic surface realization [2][27]. For the purposes of this research, templates are sufficient, as the types of sentences needed is finite. Once the templates are used to generate sentences, a simple heuristic for

aggregation is used, such that phrases that are repeated in a paragraph are omitted. For example, the following paragraph is aggregated to produce a simpler paragraph:

> Boston scored 4 runs in the second. In the second inning, MLowell (BOS) hit a 1-run double. In the second inning, AStern (BOS) hit a 2-run double. In the second inning, KYoukilis (BOS) hit a 1-run double.

After aggregation, the paragraph becomes:

> Boston scored 4 runs in the second. First, MLowell (BOS) hit a 1-run double. Then, AStern (BOS) hit a 2-run double. Then, KYoukilis (BOS) hit a 1-run double.

After aggregation, the generation process is complete.

Chapter 6

System Evaluation

This chapter starts by a general discussion of issues in evaluating natural language and summary generation systems. Then, it will describe the user study carried out in order to evaluate the performance of the perspective generation system. The results of the user study show that the system is successsful in generating summaries with the desired perspective.

6.1 Evaluation for NLG

It is not easy to evaluate a multiple perspective summary generation system. A summary generation system, even without the additional problem of multiple perspectives, is difficult to evaluate well because there are many aspects of the system that can and should be evaluated. The goal of a summary generation system is to produce accurate, easy-to-read, and concise summaries, so a system-generated summary should be evaluated for its document planning, content selection, and surface realization for accuracy and style. What makes NLG evaluations more difficult in general than most NLU evaluation is that it is difficult to evaluate a generation system automatically, so much of evaluation is driven by human judgments [25]. Recent efforts in devising automatic evaluation metrics in machine translation, such as BLEU [29] and NIST [10] have led to a similar development of automatic evaluation for NLG (cf. [3], [19], [22]), and they have been met with both enthusiasm and criticism. A

proposed automatic metric, such as ROUGE [22], attract trial from the community (cf. [40]) because successful automatic metrics would be useful, efficient, and would serve to compare different systems. However, there has been criticism that automatic metrics based on relatively simple pattern matching, such as n-grams, are not good measures of well-written texts [33]. With the goal of testing the effectiveness of the automatic evaluation metrics, Belz and Hovy compared the automatic metrics with human judgments to see whether they are correlated [6], and they show that some of the automatic metrics do correlate with human judgments, but they conclude that it is best to use the automatic evaluation metrics as a way to supplement human evaluation.

When using human judges to evaluate NLG systems, the judges can evaluate the system output for its own qualities, or they can evaluate the system output comparatively against a baseline or a reference text. Comparative evaluation is useful and efficient in many applications where reference or baseline text is readily available. Comparative evaluation is used in other related fields such as NLG for spoken dialog systems (cf. [27]). For the system here, I chose to use comparative evaluation because of two reasons. First, it is relatively simple to produce baseline and reference texts for comparison. Second, and more importantly, the goal of this research is to produce summaries with different perspectives, so the success of the system is measured in terms of whether the system-produced summaries differ in the human judgments of perspective. Automatic evaluation using metrics similar to ROUGE may be used in the future to analyze how the system output compares to reference texts.

6.2 User Evaluation

A user evaluation of system output was designed similar to the user survey presented in Chapter 3. In order to show whether the perspective generation algorithm described in Chapter 5 was successful, users were asked to read system output as well as reference and baseline summaries to judge the perspective of the summary. The exact wording of the question, as well as the system output and reference summaries

used, is in Appendix B.

6.2.1 Web-Based Survey

As in the first user survey, this user evaluation was conducted on-line using CGI scripts, and the users' answers were automatically collected into text files.

6.2.2 Participants

There were fifteen users who participated in the study. They were recruited using an email list primarily used for recruiting experimental subjects. They were paid $15 for their time, either in cash or Amazon gift certificate. Seven of the fifteen subjects were MIT students or researchers, and the other eight were non-MIT affiliates. They were all ages 18 and up who are native speakers of English. They also agreed that they watch at least one MLB game per season, and they were asked for their favorite team, but that information did not have any effect on the results. There was not one team that a majority of the subjects were fans of. If the subjects were fans of one particular team, it would be possible and useful to design a survey such that the subjects would be asked whether they like one summary over another, where one of them would be written from that team's perspective. This alternative experimental design may be used in future studies for a more discreet way to measure perspective.

6.2.3 Summaries Evaluated

Subjects were asked to judge twenty summaries, five games, four summaries per game. The games were chosen for variety, such that various teams were represented, as well as different types of games: one-sided game, close game, game with a monumental event, and a non-eventful game. The four conditions were chosen to be two baseline summaries and two system output summaries.

- Neutral: This is the AP content organized in the same order as the original AP article. The AP article is modified such that non-aligned content is discarded

(e.g., player quotes), and sentences are replaced with templates described in 5.4.

- Chronological: This is a summary where the content ordering is purely chronological. Content from the AP article are organized into paragraphs by innings, paragraphs are ordered chronologically, and sentences within paragraph are also ordered chronologically. Sentences from the original AP article are replaced with templates described in 5.4.

- Team A: This is the system output written from team A's perspective. This is the AP content arranged in the order to produce the Team A perspective, as described in the reordering section 4.2.

- Team B: This is the system output written from team B's perspective.

The order in which these summaries appeared in the surveys were varied, such that, for each game, any of the four different conditions appeared first, and the rest of the ordering was also counterbalanced. The games were presented in the same order, but that should not cause any confoundings in the results. The subjects' ratings were converted into numeric values, such that a strong Team A perspective gets a score of 1, a neutral perspective gets a score of 3, and a strong Team B perspective gets a score 5.

6.3 Results and Discussions

The subjects' perspective ratings were averaged over all 15 subjects, and analysis of variance (ANOVA) was run to compare the perspective ratings for each condition for each game. Table 6.1 shows the mean values of the perspective ratings for all games and conditions. For games 1, 2, 3, and 5, ANOVA results show that the independent variable (four different conditions) is a significant factor in the perspective ratings. For game 4, there is no sigificant difference among the four conditions. This is an exceptional case, and it is probably because game 4 was a one-sided game, where all the content in the AP article was related to the winning team's offense.

	Game 1	Game 2	Game 3	Game 4	Game 5
Team A	2.75	2.38	2.73	3.50	2.00
AP	3.64	3.86	3.63	3.27	2.75
Team B	3.88	4.27	4.25	3.63	3.3
Chron	3.40	4.13	3.29	3.39	2.29

Table 6.1: Results of ANOVA for User Study 2. The independent variable was a significant factor in games 1, 2, 3, and 5.

Game1	Team A	AP	Team B	Chron
Home		0.03	0.02	0.09
AP			0.45	0.39
Away				0.13

Table 6.2: Results of Pairwise t-test for Game 1.

Next, we ran a pairwise t-test for the 4 games that showed significant effect of the independent variable, games 1, 2, 3, and 5. We tested whether Team A and Team B show significantly different perspective ratings, as well as Team A and AP, Team B and AP, Team A and Chron, and Team B and Chron. We expect that there would be small p-values for the Team A and Team B pair, as well as the Team A/B and the two baseline AP/Chron conditions. As expected, the p-values shown in the tables are small, meaning the system output shows significant difference in perspective ratings for those pairs.

Game1	Team A	AP	Team B	Chron
Home		0.01	0.00	0.02
AP			0.12	0.49
Away				0.67

Table 6.3: Results of Pairwise t-test for Game 2.

Game1	Team A	AP	Team B	Chron
Home		0.02	0.00	0.14
AP			0.06	0.41
Away				0.03

Table 6.4: Results of Pairwise t-test for Game 3.

Game1	Team A	AP	Team B	Chron
Home		0.15	0.00	0.39
AP			0.03	0.31
Away				0.00

Table 6.5: Results of Pairwise t-test for Game 5.

Chapter 7

Related Work

This chapter looks at previous research in psychological literature and media studies, as well as related work in the computational modeling research.

7.1 Multimedia Analysis and Generation

Creating a biased story has been explored in Bocconi [7]. Theirs is a very interesting system that retrieves video interviews based on a user's point of view. When a user wishes to make an argument (e.g., "U.S. should not go to war in Afghanistan"), the system searches the interviews to provide evidence in support of the argument, therefore creating a biased documentary. The goals of our project overlap with theirs, but the appraoches are different. The inspirational part of their system is the use of rhetorical structure in creating a story with which to support an argment. For the interview database, they analyze and annotate the audio manually and use the annotations in retrieving the appropriate interviews. An important part of our system is to understand, using external data, the semantics of the events, thus automatically generating annotations of semantic features for the video clips.

There is a large body of work in sports video analysis. Earlier work was focused on rule-based systems for video indexing [45] [37], and recent projects have used statistical pattern recognition for detecting significant events in sports videos [44] [43]. There is interesting work in sports video summarization [38], but it is mainly

based on metadata, rather than automatic detection of events.

7.2 Sentiment Analysis and Generation

In a related field, one important area of recent progress has been in sentiment analysis. Work such as [28] has identified a critical problem and a well-designed solution for extracting information from the Web. Sentiment analysis is related to this work in that it tries to figure out the viewpoint of a text, but the question it asks is whether a text has a positive or a negative rating toward a product. While much can be learned from the sentiment-analysis community, both our problem and our approach differ quite a bit. First, the problem of measuring bias does not seek an answer from a binary, or even finite, set of choices. Second, our approach looks at the content of the text and how it differs from the content of another text. More details about the approach will be presented in later sections.

7.3 Psychology and Media Studies Literature

There is a large body of psychology literature about perspective-taking (cf. [34]), which is an ability of humans to comprehend someone else's point of view. A large part of this research is about physical viewpoint, and very young children acquire this ability, shown by the fact that he understand what he sees may not be exactly what his mother sees from the other side. Later on, children acquire the ability to do perspective-taking about beliefs. Although he knows there is a piece of candy in a crayon box, his friend, who has not seen the inside of the box, does not think that there is candy in there. All of this is related to language and story-telling, as children who are able to complete persepctive-taking tasks show a good command of second and third personal pronouns [34].

7.4 Perspective Classification

There has been recent work by [23] on classifying perspective. They have collected, from online news sources, articles about the political situation in Israel and Palestine. The online source they use, the Bitter Lemon corpus, is divided well into articles from the Israeli perspective and the Palestinian perspective. That characteristic of the corpus enables them to use the corpus for training a statistical classifier for inferring whether an article is written from the Israeli or the Palestinian perspective. This is certainly interesting work and has close connections to this thesis. One of the interesting aspects of their work is that the corpus is open to others, so comparions can be done with alternative inference algorithms. The reordering algorithm here, for example, can be modified such that it can be used for analyzing perspective of an article. There is certainly much future work to be done for perspective classification, and looking at content planning and ordering is one interesting direction.

7.5 User Modeling

The user modeling community such as [30] has also done similar research in text generation for specified audiences. There are two major differences between their work and this work here. First, the user modeling community assumes a much deeper knowledge of the user preferences and experiences. Here in this work, the only assumption about the potential reader is the team that they prefer. There are no other knowledge required, such as how well the user knows the domain or what kind of language he/she prefers to read. Secondly, much of [30] relies on rules hand-crafted for each type of user. The contribution of this work is that the content reordering strategy is learned through the grouping feature weights which are learned from the parallel corpus.

7.6 Content Ordering

Content organization and ordering is a step in natural language generation that takes a predetermined set of contents, organizes them into paragraphs, and orders them. In a recent book that presents a comprehensive overview of NLG [32], Reiter and Dale discuss *document structuring*, the part of an NLG system that organizes and orders content, as an essential part of NLG for making multisentential text readable and understandable. Content organization and ordering has been studied in the context of many NLP applications including summarization [24] and concept-to-text generation [21]. These and other previous work on content organization and ordering have focused mainly on readability of the generated output, and thus much of the effort has been on discourse and sentence-to-sentence coherence. In contrast, in this thesis, content organization and ordering is used as a means to induce a certain perspective of the generated output rather than focusing on coherence. However, it is shown in (add citation) that there are multiple ways to order content that result in coherent text, so producing multiple orderings to induce different perspectives does not mean that coherence or readability of the text must be sacrificed.

Chapter 8

Contributions and Future Work

This thesis has looked at perspective and generating multiple perspectives in a concept-to-text generation. The major contributions of the work are problem definition, corpus collection, and prototype building and evaluation. The problem of generating multiple perspectives for a game summary has not been explored before, and this thesis has introduced that problem and narrowed it down to a computationally tractable problem by introducing and validating the Ordering Hypothesis. It may seem obvious, but it is an important discovery that neutral content can be transformed into non-neutral articles by regrouping and reordering content. It is also important that the two user surveys established a simple way to measure perspective. A similar survey can be used in situations where the sources of the text are not well known, meaning the perspective of an article is not readily obvious.

The corpus collected and described presents one way to begin thinking about what is needed to study perspective. The baseball domain is a good first step in trying a computational model of perspective, and the parallel corpus of game data and local newspaper articles will serve as a good database of domain models and aligned articles. The features identified are useful for using statistical learning algorithms for content selection and planning, and we have shown that the features are useful in selecting content and ordering the feature vectors.

The reordering algorithm presented in this thesis is much different from previous work in content planning in that the overall goal of the algorithm is to produce

a desired perspective. Furthermore, the reordering algorithm takes advantage of the parallel corpus and uses a corpus-based learning algorithm. This will make the system generally applicable to other domains.

The system takes very simple approaches for content selection and surface realization, and that highlights the importance of content ordering on perspective, and the system evaluation shows that the system produces the desired perspectives successfully. In all, the problem definition, corpus collection, reordering algorithm, and multi-perspective generation system combine to make this thesis a good proof of concept from which to delve deeper into the important problem of perspective generation.

8.1 Content Selection

There are many future directions for this work, and content selection and surface realization, the two other components of NLG, are good candidates. We showed that simple statistical learning can be done on the features to select content for neutral articles. A similar approach can be taken for generating content that has a certain perspective. Although the Ordering Hypothesis states that content ordering is a significant factor in producing multiple perspectives, choosing the content differently may also contribute significantly to multiple perspectives.

8.2 Surface Realization

Surface realization is another component of NLG that can be explored. It is an easy guess to make that lexicalization, aggregation, and syntactic structure will contribute to different perspectives. For example, just a simple change of the sentence "Pitcher X gave up a home run to batter Y" to "Batter Y hit a home run off of Pitcher X" would probably make a signifiant difference in how the user perceives the perspective of that sentence. Similarly, changing the way the named entities are lexicalized, for example "David Ortiz" versus "Big Papi" assumes the readers' preferences for certain teams and players, and will make a significant difference in the readers' judgment of

perspective.

8.3 Other Domains

This thesis explored only one domain, that of baseball summaries. That was a good first-time domain because domain modeling can be done relatively easily, and there is tons of data that can be automatically collected and analyzed. However, because the spots domain is somewhat different from other domains that do not have a well-defined set of rules, it may be interesting to see if a similar system can be built in domains such as politics or finances. I believe important concepts, such as the Ordering Hypothesis, will generalize to other domains, but it will be difficult to implement many parts of the system because certain automatic analyses done for baseball cannot be done as easily in other domains. For example, I used baseball game descriptions to produce feature vectors of events, but such a clean model of the domain is just not available for non-sports domains. Nevertheless, there is much interest in other domains, and I would like to explore some of those as the next step of this work.

8.4 Statistical Learning

There was some simple statistical learning done in this thesis for content selection and content planning. However, much more can be done with different models and algorithms. Since the parallel corpus can be very useful for statistical learning, it would be good to take advantage of that resource. Also, the learing done in this work was very specific to each team. It would be much more desirable to have a more general model based on home versus visiting team, or winning versus losing team, such that it would be not required to have a corpus for a selected team in order to produce a summary article from that team's perspective.

In other domains, such as politics where a well-defined domain is difficult to get automatically, it may be much more important to use statistical learning for many other components of the system. The events themselves can be learned automatically,

such that we can use a group of articles about one event and build a probabilistic event model.

The field of NLP and NLG are evolving such that subjective measures, such as sentiment and opinions, are becoming into focus. Perspective is another one of those subjective dimensions of text, and it will become more and more important to look at perspective as an important problem to look into. If this thesis can convince the NL community of that and serve as a starting point, then that would be the biggest contribution of this work.

Appendix A

System Evaluation

This section has the actual surveys used in user study 1.

In the following 15 pagess, you will read summaries of 5 baseball games, 3 summaries per game, for a total of 15 game summaries. At the top of each page, there will be a table of game results–an inning-by-inning scoring of the game. Look at the game results, then read the game summary below it, and then answer the question at the bottom. There is no right or wrong answer.

> Baseball Summary 1 A different lineup produced the same result for the Devil Rays on Wednesday night.
>
> The Rays tallied just four hits, even though manager Joe Maddon tried to provide his team a fresh outlook when he fiddled with the front end of the lineup, and couldn't find enough production despite a pair of home runs.
>
> A four-run fifth inning was too much for the Rays to overcome as they lost to Minnesota, 7-2, at the Metrodome. It was the club's sixth straight loss – the longest losing streak of the season and worst second-half start in franchise history.
>
> The Rays, who are a season-high 16 games under .500, have dropped nine of 10 game and 11 straight to the Twins since 2004.

"Before the break we looked wonderful, and now we don't look so wonderful," Maddon said. "We have to get better than that."

Minnesota starter Brad Radke (8-7) tossed seven innings and allowed just four hits, including a pair of solo home runs. After struggling in the first couple of months this season, Radke hasn't lost since June 3.

The Rays' four hits tied their season low, done four times previously, including Tuesday night against Francisco Liriano.

"Sometimes there are extenuating circumstances to your demise, and we'll have to just keep battling until we get through it," said Maddon, noting the team's tough-luck run of facing strong pitching.

The new lineup couldn't beat Radke, but it did end a couple of cold streaks for the Rays.

Designated hitter Jonny Gomes, who hit in the second spot for the first time this season, homered in the sixth inning to snap an 0-for-21 streak. The blast was Gomes' 19th home run of the season and his first hit since the All-Star break.

Third baseman Ty Wigginton, who had missed the last four games with a strained back, collected two of Tampa Bay's four hits. He ended an 0-for-16 streak with a single in the fifth.

"It was definitely nice to get back out there," said Wigginton, who added that his back felt fine, even after making a diving stop at third.

Rays starter Jae Seo (0-4) lasted 5 2/3 innings and gave up 11 hits and seven runs. The Twins took the lead with a four-run fifth inning that began with a leadoff homer by Rondell White, who hit another homer two innings earlier to tie the game.

Down, 2-1, with one out and runners on the corners, Minnesota's Nick Punto hit a 1-2 pitch down the right-field line for a triple. He scored on a sacrifice fly two batters later.

"It came down to the at-bat with Punto," Maddon said. "I can't say it was an awful pitch. He gets a breaking ball and put it right down the line. [The Twins] work good at-bats."

Minnesota tacked on two more runs in the sixth with three consecutive two-out hits – the last two off reliever Shawn Camp, who replaced Seo after Jason Bartlett ripped an RBI triple.

Right fielder Greg Norton homered in the second inning, giving the Rays an early 1-0 lead. The ball just cleared the wall in left, reaching the first row of seats. Carl Crawford and Rocco Baldelli – both moved down in the order to third and fourth, respectively – combined to hit 0-for-8.

The Rays look to prevent a four-game sweep Thursday against Johan Santana and try to win one game on the road trip before returning home for series against Baltimore and Anaheim.

"It's real frustrating," Gomes said. "It's not like no one doesn't want to spark the first [win], it's not like no one's trying. We just have to go out and get them and not sit back and watch."

Baseball Summary 2 of 12 Struggling through perhaps his worst season, White hit two home runs Wednesday night to lead the Minnesota Twins to their sixth straight victory, 7-2 over the Tampa Bay Devil Rays.

Since coming off the disabled list on July 15 with a strained left shoulder, White has gone 8-for-14 with three home runs and six RBI. Before that he had no home runs and 16 RBI.

Brad Radke won his fourth straight decision for Minnesota, which has won 18 of 23 overall and 19 of the last 20 at home. Radke (8-7) allowed two runs and four hits in seven innings, while striking out four and walking none.

Greg Norton and Jonny Gomes both homered for the Devil Rays, who have lost a season-high six in a row. Tampa Bay has lost 11 straight to

the Twins dating to 2004.

Jae Seo (0-4) allowed seven runs and a career-high 11 hits in 5 2/3 innings. Seo has lost all four of his starts since being acquired from the Los Angeles Dodgers on June 27.

Norton hit his sixth homer off Radke in the second inning to give Tampa Bay a 1-0 lead, but White hit his first of the game in the bottom of the inning.

White led off the fifth with his third homer of the season, a 410-foot shot that made it 2-1.

The Twins went up 5-1 in the fifth on Nick Punto's two-run triple and Michael Cuddyer's sacrifice fly.

Gomes homered in the sixth to snap an 0-for-19 slump.

Jason Bartlett had an RBI triple and Luis Castillo added a run-scoring single in the sixth to make it 7-2.

Baseball Summary 3 of 12 An impressive night on Wednesday, when Rondell White belted two home runs and a double in the Twins' 7-2 victory over the Devil Rays, officially signaled the rebirth of the power hitter the club had expected when it signed him in the offseason.

White's first homer came in the second inning, when the Twins trailed, 1-0. The 395-foot blast to left field off Devil Rays starter Jae Seo knotted the game at 1.

The game remained tied until White's next at-bat in the fifth. Leading off the inning, White delivered another shot to left – this one carrying 410 feet to put the Twins up, 2-1.

The best part of White's night may have come when he came to the plate for his final at-bat. White was greeted by the fans with a standing ovation, to which White tipped his helmet.

White's homer in the fifth sparked a four-run inning courtesy of a two-run triple by Nick Punto and an RBI sacrifice fly by Michael Cuddyer. The hit by Punto extended his hitting streak to 12 games.

More runs were added to the team's lead in the sixth, as the Twins drove home two on a Jason Bartlett RBI triple and an RBI single by Luis Castillo.

Radke (8-7) allowed just two runs, both coming on homers, on four hits. Besides the two mistakes, Radke was able to show good command, issuing no walks and throwing just 87 pitches over seven innings. Radke has not lost since June 3.

Baseball Summary 5 of 12 The A's picked up what should have been a feel-good win Wednesday, downing the Orioles, 5-1, behind a homer and three RBIs from Frank Thomas, Eric Chavez's first homer in more than a month, and seven brilliant innings from Barry Zito.

A fine win it was for the A's, who maintained their slim lead in the American League West by winning their second consecutive road series.

Chavez, who has been battling tendinitis in both forearms and entered the game batting .133 (12-for-90) over his past 25 games, gave Oakland a 4-0 lead when he took Orioles starter Kris Benson (9-9) deep to right field with one out in the sixth inning.

Thomas, who had given the A's a 2-0 lead with a two-run single with two out in the first, hit his 20th homer of the year two pitches after Chavez's blast, sending a Benson fastball 410 feet into the left-field bleachers. Oakland's third run came when Bradley, who led the A's with three hits, homered to right with one out in the third.

Zito (10-6), who idolized Benson in his late teens, was brilliant in their first head-to-head matchup; Oakland's ace faced three batters over the minimum in the first six innings while Benson was giving up five runs on nine hits and a walk.

The O's finally broke through in the seventh, when Kevin Millar doubled and scored on a bloop single by Ramon Hernandez. Kotsay was then charged with two errors after mishandling Corey Patterson's grounder and flipping past Zito at the bag to put runners at the corners with one out, but Zito got out of the jam by getting Chris Gomez to hit into an inning-ending double play.

Baseball Summary 6 of 12 Zito pitched seven innings of five-hit ball, Frank Thomas homered and drove in three runs, and Oakland defeated Kris Benson and the Baltimore Orioles 5-1 Wednesday.

Milton Bradley and Eric Chavez also homered for the Athletics, who took two of three from Baltimore to improve to 5-2 since the break. After beating up on Boston and Baltimore, the A's get a day off before beginning a weekend series against the Detroit Tigers.

Zito (10-6) helped, pitching five solid innings against the Red Sox in a 15-3 win. He now has 53 wins after the break since 2000 – second in the majors behind Bartolo Colon (54).

Against Baltimore, the left-hander allowed only one runner past first base through the first six innings while Oakland built a 5-0 lead. It was the 14th time in 21 starts he has gone at least seven innings, and the 12th time he yielded two runs or fewer.

Ramon Hernandez drove in a run for the Orioles, now 9-23 when the opposition starts a left-hander.

Benson (9-9) gave up five runs and nine hits in losing his fourth straight start. He yielded all three Oakland homers, but it was only the fourth time in 21 starts this season that the right-hander allowed as many as five earned runs.

Thomas put Benson in a hole in the first inning. With two outs and runners on second and third, Thomas lined a two-run single to left for a 2-0 lead.

Bradley's fourth homer – his first since April 23 – made it 3-0 in the fourth.

In the sixth, Bradley hit a ball down the left-field line but tripped over first base and was tagged out while he headed for second. Chavez followed with his 15th homer, and two pitches later Thomas hit No. 20 for a 5-0 lead.

In the Baltimore seventh, Millar hit a one-out double, advanced on a passed ball and scored on Hernandez's bloop to right. Corey Patterson then reached on an error, but Zito ended his strong outing by getting Chris Gomez to hit into a double play.

Baseball Summary 7 of 12 Byrd threw six strong innings to help end Los Angeles' eight-game winning streak, Ben Broussard hit a two-run homer that snapped Lackey's scoreless string at 30 2/3 innings, and the Cleveland Indians stopped their own five-game skid with a 6-4 victory Wednesday.

Byrd (7-6) allowed three runs, 10 hits and no walks over six innings to improve his career record against the Angels to 3-0, including a three-hit shutout for Kansas City in 2002.

Juan Rivera homered in the fourth for the Angels, his 14th this season and eighth in 15 games – including the two he hit during Tuesday night's 7-5 victory. Vladimir Guerrero had a pair of RBI singles.

Aaron Boone also homered and Jhonny Peralta hit a go-ahead single for Cleveland. Bob Wickman worked the ninth for his 15th save in 18 attempts.

Right fielder Casey Blake preserved Byrd's 5-3 lead in the sixth with a sensational, diving grab of Chone Figgins' slicing fly toward the line with two on.

Lackey (8-6) gave up five runs and 10 hits over 4 2/3 innings after throwing consecutive shutouts against Oakland (one-hitter) and Tampa Bay

(five-hitter). The right-hander, trying for his fourth straight double-digit strikeout game, fanned seven and walked five while working with runners on base each inning.

Lackey threw 45 of his 107 pitches during the first two innings and stranded five baserunners. His luck ran out in the fifth, when Cleveland scored five runs for a 5-2 lead.

Broussard was 1-for-14 lifetime against Lackey before his tying homer, which came after a walk to Victor Martinez. Broussard's 12th of the season was only the second homer given up by Lackey in his last seven starts.

Peralta gave the Indians a 4-2 lead with his third straight hit, a two-run single that landed just inside the right-field line. Rookie Joe Inglett capped the rally with a run-scoring single, his first RBI in the majors.

Boone's fifth homer made it 6-3 in the seventh against Brendan Donnelly.

Appendix B

System Evaluation

This section has the actual surveys used in user study 2.

In the following 15 pagess, you will read summaries of 5 baseball games, 3 summaries per game, for a total of 15 game summaries. At the top of each page, there will be a table of game results–an inning-by-inning scoring of the game. Look at the game results, then read the game summary below it, and then answer the question at the bottom. There is no right or wrong answer.

> Survey 1A Summary 1 of 4 for Game 1
>
> Toronto scored 2 runs in the eighth. Top of eighth, FCatalanotto (TOR) hit a 2-run homerun.
>
> Toronto had hits in the first, second, and seventh. Top of first, VWells (TOR) hit a single. Then, SHillenbrand (TOR) hit into a doubleplay with runners on base. Top of second, AHill (TOR) hit a double. Top of seventh, BMolina (TOR) hit a single.
>
> Boston scored 4 runs in the second. First, MLowell (BOS) hit a 1-run double. Then, AStern (BOS) hit a 2-run double. Then, KYoukilis (BOS) hit a 1-run double.
>
> Boston scored 1 run in the seventh inning. DOrtiz (BOS) hit a 1-run homerun.
>
> The final score was Toronto 3, Boston 5.

Survey 1B Summary 1 of 3

Boston scored 4 runs in the second. First, MLowell (BOS) hit a 1-run double. Then, AStern (BOS) hit a 2-run double. Then, KYoukilis (BOS) hit a 1-run double.

There were 2 homeruns in the game.¡br¿ In the seventh, DOrtiz (BOS) hit a 1-run homerun.¡br¿ In the eighth, FCatalanotto (TOR) hit a 2-run homerun.¡br¿¡br¿

Toronto had 1 hit in the first inning.¡br¿ First, VWells (TOR) hit a single.¡br¿ Then, SHillenbrand (TOR) hit into a double play with runners on base.¡br¿¡br¿

Toronto had hits in the second and seventh.¡br¿ Top of second, AHill (TOR) hit a double.¡br¿ Top of seventh, BMolina (TOR) hit a single.¡br¿¡br¿

The final score was Toronto 3, Boston 5.¡br¿

Survey 1C Summary 2 of 3

Toronto had 1 hit in the first inning. First, VWells (TOR) hit a single. Then, SHillenbrand(TOR) hit into a double play with runners on base.

Boston scored 4 runs in the second. First, MLowell(BOS) hit a 1-run double. Then, AStern(BOS) hit a 2-run double. Then, KYoukilis(BOS) hit a 1-run double.

There were 2 homeruns in the game. Bottom of seventh, DOrtiz(BOS) hit a 1-run homerun. Top of eighth, FCatalanotto(TOR) hit a 2-run homerun.

Toronto had hits in the second and seventh. Top of second, AHill(TOR) hit a double. Top of seventh, BMolina(TOR) hit a single.

Final score was Toronto 3, Boston 5.

Survey 1D Summary 3 of 3

Toronto had 1 hit in the first inning. First, VWells(TOR) hit a single. Then, SHillenbrand(TOR) hit into a double play with runners on base.

Top of second, AHill(TOR) hit a double.

Boston scored 4 runs in the second. First, MLowell(BOS) hit a 1-run double. Then, AStern(BOS) hit a 2-run double. Then, KYoukilis(BOS) hit a 1-run double.

Top of seventh, BMolina(TOR) hit a single.

There were 2 homeruns in the game. Bottom of seventh, DOrtiz(BOS) hit a 1-run homerun. Top of eighth, FCatalanotto(TOR) hit a 2-run homerun.

Final score was TOR 3, BOS 5.

Bibliography

[1] Elisabeth Andre, Gerd Herzog, and Thomas Rist. On the simultaneous interpretation of real world image sequences and their natural language description: The system soccer. In *Proceedings of the 8th ECAI*, 1988.

[2] Srinivas Bangalore and Owen Rambow. Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th conference on Computational Linguistics*, 2000.

[3] Srinivas Bangalore, Owen Rambow, and Stephen Whittaker. Evaluation metrics for generation. In *Proceedings of the first international conference on Natural Language Generation*, 2000.

[4] Regina Barzilay. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation.* PhD dissertation, Columbia University, 2003.

[5] Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55, 2002.

[6] Anja Belz and Ehud Reiter. Comparing human and automatic evaluation of nlg systems. In *Proceedings of EACL 2006*, 2006.

[7] Stefano Bocconi and Frank Nack. Automatic generation of biased video sequences. In *Proceedings of the 1st ACM Workshop on Story Representaiton*, October 2004.

[8] Robert Dale. *Generating referring expressions in a domain of objects and processes.* PhD dissertation, University of Edinburgh, 1988.

[9] Aggeliki Dimitromanolaki and Ion Androutsopoulos. Learning to order facts for discourse planning in natural language generation. In *Proceedings of EACL 2003 Workshop on Natural Language Generation*, 2003.

[10] G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of ARPA Workshop on Human Language Technology*, 2002.

[11] Pablo Duboue and Kathleen McKeown. Content planner construction via evolutionary algorithms and a corpus-based fitness function. In *Proceedings of the Second International Natural Language Generation Conference*, July 2002.

[12] Pablo A. Duboue and Kathleen R. McKeown. Empirically estimating order constraints for content planning in generation. In *Proceedings of the Annual Meeting of Association for Computational Linguistics*, 2001.

[13] Michael Elhadad and Jacques Robin. An overview of surge: A reusable comprehensive syntactic realization component. In *Proceedings of the 8th International Workshop on Natural Language Generation*, 1993.

[14] Paul D Ji and Stephen Pulman. Sentence ordering with manifold-based classification in multi-document summarization. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 526–533, Sydney, Australia, July 2006. Association for Computational Linguistics.

[15] Min-Yen Kan and Kathleen R. McKeown. Corpus-trained text generation for summarization. In *Proceedings of the Second International Natural Language Generation Conference*, pages 1–8, July 2002.

[16] Nikiforos Karamanis and Chris Mellish. Using a corpus of sentence orderings defined by many experts to evaluate metrics of coherence for text structuring. In *Proceedings of the ENLG*, 2005.

[17] Nikiforos Karamanis, Massimo Poesio, Chris Mellish, and Jon Oberlander. Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 391, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[18] Kathleen Kuiper, editor. *Merriam Webster's Encyclopedia of Literature.* Merriam-Webster, 1995.

[19] Irene Langkilde-Geary. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 12th International Natural Language Conference*, 2002.

[20] Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the annual meeting of ACL, 2003*, 2003.

[21] Mirella Lapata and Regina Barzilay. Collective content selection for concept-to-text generation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 331–338, Vancouver, October 2005.

[22] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL*, pages 71–78, Edmonton, Canada, May-June 2003.

[23] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning(Co-NLL-X)*, pages 109–116. Association for Computational Linguistics, 2006.

[24] Nitin Madnani, Rebecca Passonneau, Necip Fazil Ayan, John Conroy, Bonnie Dorr, Judith Klavans, Dianne O'Leary, and Judith Schlesinger. Measuring variability in sentence ordering for news summarization. In *Proceedings of the 11th*

European Workshop on Natural Language Generation, Dagstuhl, Germany, June 2007. Association for Computational Linguistics.

[25] Chris Mellish and Robert Dale. Evaluation in the context of natural language generation. *Computer Speech and Language*, 12:349–373, 1998.

[26] Johanna D. Moore and Cecile L. Paris. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651695, 1993.

[27] Alice Oh and Alex Rudnicky. Stochastic natural language generation for spoken dialog systems. *Computer Speech and Language*, pages 387–407, October 2002.

[28] Bo Pang and Lillian Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[29] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL-2002*, pages 311–318, 2002.

[30] Cecile L. Paris. *User Modeling in Text Generation.* Communication in Artificial Intelligence Series, 1993.

[31] Dragomir R. Radev and Kathleen McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 1998.

[32] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems.* Cambridge University Press, 2000.

[33] Ehud Reiter and Somayajulu Sripada. Should corpora texts be gold standards for nlg. In *Proceedings of the Second International Conference on Natural Language Generation*, 2002.

[34] Marcell Ricard, Pascale C. Girouard, and Therese Gouin Decarie. Personal pronouns and perspective taking in toddlers. *Journal of Child Language*, 26:681–697, 1999.

[35] Benjamin Snyder and Regina Barzilay. Database-text alignment via structured multilabel classification. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2007.

[36] Amanda Stent. Content planning and generation in continuous-speech spoken dialog systems. In *Proceedings of KI99 workshop May I Speak Freely*, 1999.

[37] G. Sudhir, J. Lee, and A. Jain. Automatic classification of tennis video for high-level content-based retrieval. In *Proceedings of IEEE International Workshop on Content-based Accces of Image and Video Database*, pages 81–90, 1998.

[38] Yoshimasa Takahashi, Naoko Nitta, and Noboru Babaguchi. Automatic video summarization of sports videos using metadata. In *Proceedings of Pacific Rim Conference on Multimedia*, 2004.

[39] Kristina Toutanova and Christopher Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 2000.

[40] Lucy Vanderwende, Michele Banko, and Arul Menezes. Event-centric summary generation. In *Working Notes of DUC*, 2004.

[41] Janyce M. Wiebe and William J. Rapaport. A computational theory of perspective and reference in narrative. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*, pages 131–138, Morristown, NJ, USA, 1988. Association for Computational Linguistics.

[42] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[43] L. Xie, P. Xu, and S. Chang. Structure analysis of soccer video with domain knowledge and hidden markov models. *Pattern Recognition Letters*, 25(7):767–775, May 2004.

[44] D. Zhang and S. Chang. Event detection in baseball video using superimposed caption recognition. In *Proceedings of ACM Multimedia'02*, pages 315–318, 2002.

[45] W. Zhou, A. Vellaikal, and C. Kuo. Rule-based video classification system for basketball video indexing. In *Proceedings of ACM Multimedia Workshop*, pages 213–216, 2000.