Generating Baseball Summaries from Multiple Perspectives by Reordering Content

Alice Oh
MIT CSAIL
32 Vassar St.
Cambridge, MA 02139 USA
aoh@mit.edu

Howard Shrobe
MIT CSAIL
32 Vassar St.
Cambridge, MA 02139 USA
hes@csail.mit.edu

Abstract

This paper presents a reordering algorithm for generating multiple stories from different perspectives based on a single baseball game. We take a description of a game and a neutral summary, reorder the content of the neutral summary based on event features, and produce two summaries that the users rated as showing perspectives of each of the two teams. We describe the results from an initial user survey that revealed the power of reordering on the users' perception of perspective. Then we describe our reordering algorithm which was derived from analyzing the corpus of local newspaper articles of teams involved in the games as well as a neutral corpus for the respective games. The resulting reordering algorithm is successful at turning a neutral article into two different summary articles that express the two teams' perspectives.

1 Introduction

Stories about events are written in many different perspectives, or points-of-view. For example, following a baseball game, multiple articles are written that summarize the game from different perspectives. Although they are describing the same game, readers feel differently about the articles and may prefer to read a certain perspective over all the others. We have explored what factors contribute to the differences in perspective in these event summary stories and how we can automatically plan content to generate multiple summaries of a baseball game written from different perspectives. The end goal of this work is to build a system that takes as input a factual description of a baseball game and a neutral article about the game, then produces two other articles, each from a particular team's point of view. There is previous work such as (Robin and McKeown, 1996) on automatic summary generation of sports games, but our work goes further to generate multiple summaries.

It is first necessary to define what is meant by perspective and multiple perspectives. The definition of perspective in this work is somewhat different from a more traditional meaning of perspective in literature, such as the third-person perspective discussed in (Wiebe and Rapaport, 1988). Our definition is much closer to that used in (Lin and Hauptmann, 2006), 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. For our work, we are looking at the domain of baseball games, so we focus on the article's perspective in terms of the home team versus the visiting team. We first assume that the two opposing perspectives are expressed in the local newspaper articles of the two teams, and we assume that the neutral perspective is expressed in the Associated Press articles published on an ESPN website (www.espn.com). We confirmed these assumptions via a user study, then we identified some key factors contributing to an article having a certain perspective. The next section explains our corpus and user studies.

2 Corpus

The Major League Baseball (MLB) has 30 teams within the United States and Canada, and each team plays approximately 160 games per season. We have collected data for hundreds of games from the 2005 and 2006 MLB seasons. The corpus is divided into two sets. The first is factual descriptions of the games in quantitative form and simple natural language text, and the second is journalistic writings from online news sources.

2.1 Game Data

For every MLB game, the website of MLB (www.mlb.com) publishes game data consisting of two documents. The first is a game log (see figure 1) , which is a complete list of *at-bats* in the game, where each *at-bat* is a set of pitches thrown from a pitcher to a batter such that the batter either gets out or advances to a base at the completion of the *at-bat*. There are at least 3 *at-bats* 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 *at-bats*, but usually more. In our corpus, the average number of *at-bats* is 76.2 per game. The second is a boxscore, which is a list of each batter and pitcher's performance statistics for the game. Currently we do not use the boxscore documents.

The game log is parsed using simple regular expression type patterns to turn each *at-bat* into a feature vector. We have defined 22 features: *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* are 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 if the current play is one that advances the runner one base, the previous feature vector's *onFirst* gets copied to the current *onSecond*. While we tried to identify features that are important for analyzing and generating multiple perspectives, later sections will show that some of them were not used, as they were not significant variables for our content ordering algorithm.

2.2 Online Articles

In addition to game logs and boxscores, we 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 we 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). We use the AP articles as our neutral source. We 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).

3 From Neutral to One-Sided Perspective

We are building a system that takes a neutral article and turns it into an article with a one-sided perspective, so we looked at whether we can use the same content of the neutral article and still produce a non-neutral perspective. Surprisingly, looking at the articles in terms of the game events, there were many games where the three articles overlap quite a bit in the (*at-bats*) that are mentioned in the articles. That is, the neutral and the home/visit team articles all describe the same set of game events, but they still manage to present them such that the readers notice the differences in perspective.

To compute the overlap of content, the articles were first tagged with player names and part-of-

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 1: Pitch by Pitch Log of a Baseball Game

Games	All	Home	Visit
41	215	23	21
Ave	5.24	0.56	0.51

Table 1: Number of at-bats described in all three articles, at-bats only in the home team articles, and at-bats only in the visit team articles for 41 games.

speech tags, and simple pattern matching heuristics were used to automatically align the sentences in the articles with game events. The player names were downloaded from the MLB team sites accessible from www.mlb.com, and the POS tagging was done with the Stanford POS tagger (Toutanova and Manninig, 2000). 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 articles hand-annotated by the first author, we 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. 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 one of the teams. 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, games that are close in score and number of hits, we found a high overlap of content among all three articles. Table 1 lists the number of *at-bats* reported in common in all three articles.

Based on the corpus analyses we surveyed users to see whether we can identify the important factors that contribute to differences in perspective.

First, to confirm that the home team and the visit team perspectives of the local team articles are correctly perceived, we simply presented the AP and local newspaper articles to users and asked them to guess which team the articles were written for. As expected, users identified the local team perspective with ease and confidence. Then, we took out all sentences except ones that describe the the game events (*at-bats*). Player quotes, commentary about the team or players' historical performances, and financial and personal news were some of the content that were removed from the articles. Users were asked to guess which team the articles were written for, and again, they were able to identify the local team perspectives. We then removed sentences describing game events that did not overlap with the content in the neutral article, and again, users identified the local perspectives. Finally, we replaced all the sentences with canned surface forms, such that all the articles shared the same surface form of sentences and preserved only the ordering of the content. This last experiment, albeit with less confidence than the previous ones, still produced users' perception of local perspective for the non-neutral articles. 8 users participated in the study using 12 games, and table 2 summarizes the results of these user surveys. All 8 users rated all 36 articles, 3 articles for each game, but the ordering of the articles was randomized. For all four conditions, users were asked to rate each article on a scale of 1 to 5, where 1 is strongly home team perspective, 3 is neutral, and 5 is strongly visit team perspective.

4 Feature-based Ordering Strategies

Following the results from the user study, we used a corpus-driven approach to identify the ordering strategies that contribute to the different perspectives. We looked at the games for which the three

Condition	Home	AP	Visit
Original	1.75	2.75	4.06
Events Only	1.75	2.90	3.85
Overlapping	2.02	2.75	3.85
Ordering	2.18	2.83	3.83

Table 2: Users' ratings on how they perceived perspective. They rated using a 1 to 5 scale, where 1 is the home team perspective, 3 is neutral, and 5 is the visiting team perspective. For all lines, t-test for the users' ratings of home team articles and visit tema articles show a statistically significant difference at the level $p < 0.05$.

articles have highly overlapping content and studied how the content is organized differently. We segmented the articles into topic segments (e.g, paragraphs) and noticed that the three articles differ quite a bit in the topics that hold the content together. These topics can be expressed simply by the feature values that are shared among the *at-bats* that appear together in the same segment. Below is an example of two different orderings of *at-bats* based on feature values. The first segment (lines 1a, 2a) of the first ordering shares the same values for the features $pitcher$, $team$, $inning$, and R (score added by that play). The second segment (lines 3a, 4a) shares $pitcher$, $batter$, and $team$.

	Pitcher	Batter	Team	inn	type	R
1a	Johns	Damon	Bos	1	hr	1
2a	Johns	Ramir	Bos	1	dbl	1
3a	Schil	Jeter	Nyy	4	dbl	0
4a	Schil	Jeter	Nyy	6	hr	2

The second ordering shows the same content arranged in different segments, where both segments are organized based on the value of $type$ of play. This is a frequent pattern in our corpus that seems to be responsible for the different perspectives of the articles.

	Pitcher	Batter	Team	inn	type	R
1b	Johns	Damon	Bos	1	hr	1
2b	Schil	Jeter	Nyy	6	hr	2
3b	Johns	Ramir	Bos	1	dbl	1
4b	Schil	Jeter	Nyy	4	dbl	0

Since there are many features, we need to identify the features to use for assigning the *at-bats* to appear in the same segment. We 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 we 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, we take the *at-bats* 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 *at-bats* are arranged in chronological order.

5 Conclusion

We presented a content ordering algorithm that takes a neutral article of baseball and produces two other articles from the two teams' perspectives. We showed that just by reordering, we can induce different perspectives, and we used a corpus for discovering the different ordering strategies. In the future, we will refine our reordering algorithm, carry out a full evaluation, and also look at other factors that contribute to perspective such as content selection and surface realization. We will also look at another domain, such as the socio-political conflict in the Middle East discussed in (Lin and Hauptmann, 2006), to see whether similar reordering patterns appear in those articles.

References

Wei-Hao Lin and Alexander Hauptmann. 2006. Are these documents written from different perspectives? A test of different perspectives based on statistical distribution divergence. *Proceedings of the 42th annual meeting on Association for Computational Linguistics*.

Jacques Robin and Kathleen McKeown. 1996. Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence*.

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

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