CricketLinking: Linking Event Mentions from Cricket Match Reports to Ball Entities in Commentaries

Manish Gupta
Microsoft, India
gmanish@microsoft.com

ABSTRACT

The 2011 Cricket World Cup final match was watched by around 135 million people. Such a huge viewership demands a great experience for users of online cricket portals. Many portals like espncricinfo.com host a variety of content related to recent matches including match reports and ball-by-ball commentaries. When reading a match report, reader experience can be significantly improved by augmenting (on demand) the event mentions in the report with detailed commentaries. We build an event linking system *CricketLinking* which first identifies event mentions from the reports and then links them to a set of balls. Finding linkable mentions is challenging because unlike entity linking problem settings, we do not have a concrete set of event entities to link to. Further, depending on the event type, event mentions could be linked to a single ball, or to a set of balls. Hence, identifying mention type as well as linking becomes challenging. We use a large number of domain specific features to learn classifiers for mention and mention type detection. Further, we leverage structured match, context similarity and sequential proximity to perform accurate linking. Finally, context based summarization is performed to provide a concise briefing of linked balls to each mention.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*Data Mining*; H.4.0 [**Information Systems Applications**]: General

Keywords

Cricket data analysis, Event linking, Article-Commentary linking, Mention type detection, Entity linking

1. INTRODUCTION

Cricket is one of the most popular sports with a large amount of related content published online. For each match, online portals publish ball-by-ball commentaries. These commentaries describe what happened when a ball was bowled including the name of the bowler, the batsman, number of runs scored, the type of the ball delivered, the type of shot, and sometimes comments on the form of the bowler or the batsman. Thus, every match has this detailed description of the match in the form of a maximum of 600 balls (plus the extras). Besides this, multiple experts write articles (or reports) describing the events that occurred during the match. The articles summarize the match in a concise way but miss out on interesting details due to lack of space. In this demo, we focus on providing an ability to the user to zoom in on particular event mentions from the match reports and read about a few ball commentaries most relevant to the event.

Figure 1 shows an example user interaction. When reading a match report, the user selected a phrase "The loss of Ahmed Shehzad followed by Sohaib Maqsood in the space of three balls from Umesh" and is shown ball commentary relevant to the event mentioned in the phrase. Given a match report and the corresponding ball commentaries, we focus on the problem of linking phrases from the match report with balls in the commentary. Linking could be done to a single ball, to a set of balls, or to the entire innings.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).
SIGIR'15, Aug 09-13, 2015, Santiago, Chile.
© 2015 ACM. ISBN 978-1-4503-3621-5/15/08 ...$15.00.
http://dx.doi.org/10.1145/2766462.2767865.

2. RELATED WORK

There is little work in the field of sports data analysis. To the best of our knowledge, this is the first work in the IR community focused on applying entity linking methods in the domain of sports. Shen et al. [3] provide a thorough overview of the main approaches to entity linking. A typical entity linking system consists of mention detection, candidate entity generation and candidate entity ranking. Entity linking has been found to be useful for information extraction, information retrieval, content analysis [1], question answering and knowledge base population [2]. We model the event mention linking problem to cricket commentary balls as an entity linking problem where entities are balls or sets of balls.

3. COMPONENTS OF CRICKETLINKING

Figure 2 shows a system diagram for `CricketLinking`. The system has three main components: Pre-processing, Mention Detection and Ball Linking.

3.1 Pre-processing Commentaries and Reports

The ball commentaries are available in a semi-structured form. For accurate linking, first we obtain a structured representation of each ball which contains the following fields: ball number, bowler name, bowler country, batsman name, batsman country, non-striker batsman name, runs scored, aggregate runs scored, current run rate, required run rate, previous bowler, event (out, four, six), striker statistics (runs, strike rate, fours, sixes), runner name, type of ball (short, yorker, etc.), type of shot, number of balls played by batsman till this ball, comment for batsman or bowler. Further, beyond the 600 ball entities, we also obtain linkable derived entities as follows: all balls faced by a batsman, balls that denote partnership between two batsmen, all balls bowled by a bowler, all balls bowled by a bowler and faced by a batsman, all balls on which a four was hit in the first/second innings, all balls on which a six was hit in

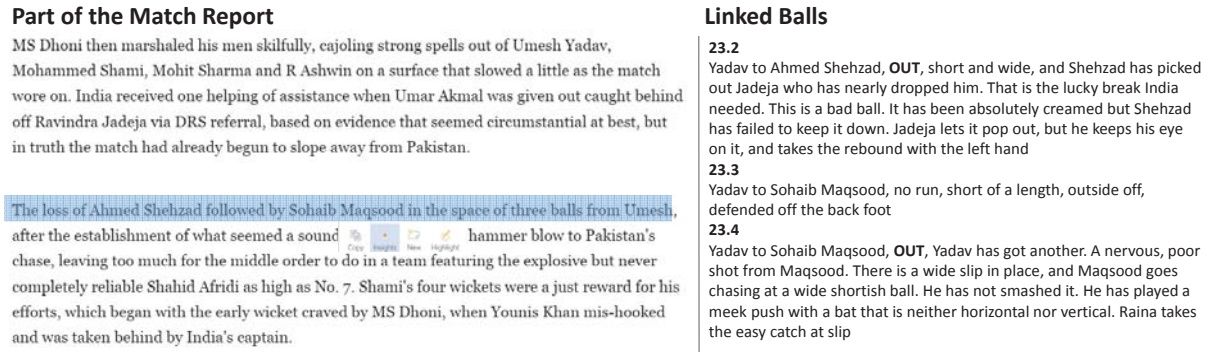

Figure 1: Screenshot of the *CricketLinking* System

the first/second innings, etc. Besides these, anomalous set of balls are also identified as derived entities. Anomalies in cricket include the following: very low runs conceded by a bowler, high number of wickets by a bowler, very high runs by a batsman, high strike rate, quick wickets, large number of fours in an over, etc.

Match reports are completely unstructured. We annotate them by performing co-reference resolution at a paragraph level, parse tree construction, and dependency analysis.

3.2 Mention Detection

This stage consists of two main parts: Event Mention Detection and Mention Type Detection. The event mention detection part identifies linkable phrases (or 'mentions'). Parse tree annotations help us to identify phrases which are classified using a large number of domain specific features. We create cricket dictionaries for cricket terms, words to indicate specific events like a wicket, a boundary or a six, etc. The features used for mention detection include number of batsman names, number of bowler names, event words, maximum similarity score with any ball, number of batsman action words, number of bowler action words, number of partnership action words, number of words found in a list of words found frequently in mentions, position of mention in the report, contains score, number of words denoting extra balls, etc. The second part identifies the type of the mention: single ball, batsman innings, all balls by a bowler, partnership, spells, powerplay, inning, extras, anomaly, etc. This also uses a classifier trained using similar features as the ones used for mention detection.

3.3 Ball Linking

In this stage, we link the mentions discovered in the previous stage to a set of balls. The first step involves identification of temporal clues like: "in the 42nd over", "first 10 overs", "last 3 balls of the tenth over." Such clues help us to link the mention to the right set of balls accurately. If there are no temporal clues, we find all balls or derived entities whose structured representation contains words from the phrase. These words could include player names, number of wickets so far, runs, run rate, a particular event, etc. Such entities form candidates that can be linked to the mention. Further, the candidates are ranked based on the Jaccard Similarity between the candidate and the mention. The similarity is computed only with respect to words in the cricket terms dictionary (around 250 words). Given a paragraph, it is necessary to maintain consistency across all entities linked to mentions within the paragraph. Usually mentions in a paragraph follow a temporal order. Thus, the entities for every mention are re-ranked based on the sequential proximity of the linked set of balls. Finally, some derived entities (like all balls faced by a player) could contain a large number of balls. Such entities are then summarized by picking eventful balls

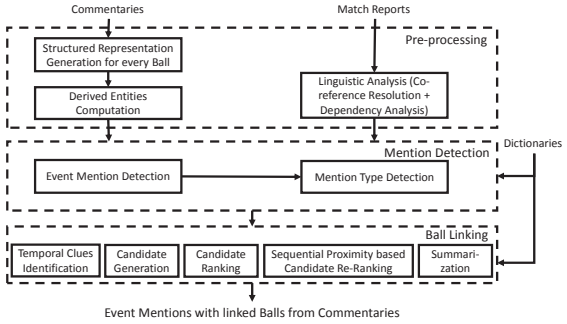

Figure 2: `CricketLinking`: System for Linking Event Mentions with Cricket Match Commentaries

and using them to show a concise summary. The user is also given an option to see all the related balls.

4. DEMONSTRATION

We will demonstrate the capabilities of `CricketLinking` as described in Section 3. We crawled match reports and commentaries for 421 matches from 2009 to 2014 from espncricinfo[1]. The demo will allow the users to select any commentary file and a match report file and see the mentions detected from the match report. The users can also select particular phrases and look for linked balls as well as a concise summary when the number of linked balls is large. There are no specific demo requirements.

5. CONCLUSION

We present the system `CricketLinking` which takes a cricket match report and commentary as input and outputs detected mentions from the report and links them to the commentary balls. The system goes beyond traditional entity linking by considering the sequential nature of both the reports as well as the entities (balls). This is the first system that applies the concept of entity linking to the sports domain. The system can be very useful for quick referencing of commentary balls when reading match reports on various cricket portals.

6. REFERENCES

[1] Abhishek Gattani, Digvijay S Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, and AnHai Doan. Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-based Approach. *PVLDB*, 6(11):1126–1137, 2013.

[2] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. In *IJCAI*, pages 3161–3165, 2013.

[3] Wei Shen, Jianyong Wang, and Jiawei Han. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *TKDE*, Jun 2014.

[1] http://www.espncricinfo.com/