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

Manish Gupta (gmanish@microsoft.com)

Problem: Provide an ability to the user to zoom in on a particular event mention from a match report, and read the ball commentaries most relevant to the event.

Part of the Match Report

MS Dhoni then marshaled his men skilfully, cajoling strong spells out of Umesh Yadav, Mohammed Shami, Mohit Sharma and R Ashwin on a surface that slowed a little as the match wore on. India received one helping of assistance when Umar Akmal was given out caught behind off Ravindra Jadeja via DRS referral, based on evidence that seemed circumstantial at best, but in truth the match had already begun to slope away from Pakistan.

The loss of Ahmed Shehzad followed by Sohaib Maqsood in the space of three balls from Umesh, after the establishment of what seemed a sound hammer blow to Pakistan's chase, leaving too much for the middle order to do in a team featuring the explosive but never completely reliable Shahid Afridi as high as No. 7. Shami's four wickets were a just reward for his efforts, which began with the early wicket craved by MS Dhoni, when Younis Khan mis-hooked and was taken behind by India's captain.

Linked Balls

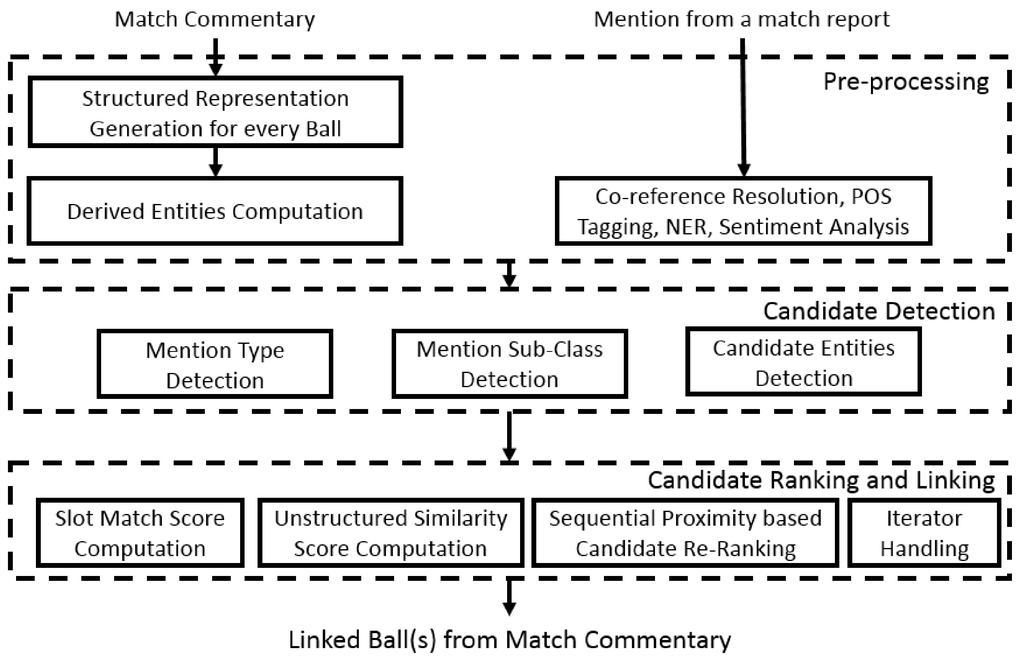
23.2
Yadav to Ahmed Shehzad, **OUT**, short and wide, and Shehzad has picked out Jadeja who has nearly dropped him. That is the lucky break India needed. This is a bad ball. It has been absolutely creamed but Shehzad has failed to keep it down. Jadeja lets it pop out, but he keeps his eye on it, and takes the rebound with the left hand
23.3
Yadav to Sohaib Maqsood, no run, short of a length, outside off, defended off the back foot
23.4
Yadav to Sohaib Maqsood, **OUT**, Yadav has got another. A nervous, poor shot from Maqsood. There is a wide slip in place, and Maqsood goes chasing at a wide shortish ball. He has not smashed it. He has played a meek push with a bat that is neither horizontal nor vertical. Raina takes the easy catch at slip

Examples

- "Ryan Doeschate produced a scintillating 119 from 110 balls" - this phrase should link to the 110 balls.
- "brilliant bowling figures of 2 for 47 in ten overs" - this phrase should link to the 60 balls bowled by the player.
- "a sparky cameo of 29 from 25 balls" - this phrase should link to the 25 balls.
- "Harbhajan and Munaf Patel put together a spell of 19 balls for just eight runs" - this phrase should link to the 19 balls.
- "The first 5 wickets fell pretty quickly" - this phrase should link to the 5 balls where the first 5 wickets got out.
- "The India innings" - this phrase should link to all the balls from the India innings.

Related Work

- Sports Data Mining
  - player performance analysis
  - player performance prediction
  - finding patterns and performing association rule mining
  - scouting or player selection
  - analyzing player dropouts
  - outcome prediction
  - retrieval of similar chess positions or similar movements from soccer game streams
  - predicting player recovery times.



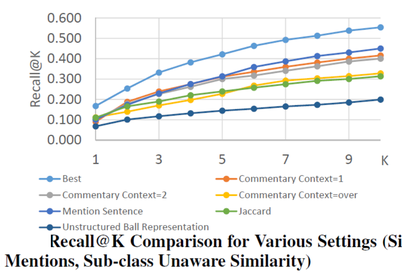
Schematic diagram showing the components of system.

System Components

- **Pre-processing Commentaries and Reports**
  - Commentary: Structured representation + text
  - Reports: POS, NER, Coreference Resolution, Sentiment Analysis
  - Scorecards: Player names, Powerplays
  - Derived entities: Semantic group of balls, e.g., all balls faced by a batsman, partnership between two batsman.
- **Detecting Candidate Entities**
  - Mention Type Detection
    - Dictionary features, Entity Features, Features capturing similarity with any ball, Other features
  - Mention Sub-class Detection
    - Single-ball mentions: OUT, LASTBALL, BALL, DROPPED, SIX, FOUR, REFERRAL, and OTHERS.
    - Multi-ball mentions: BAT, BOWL, BATBOWL, FOUR, SIX, PARTNERSHIP, WICKETS, OVERS, POWERPLAY, REFERRAL-DROPPED, EXTRAS, and OTHERS.
  - Candidate Entities Detection: Hard vs soft assignment.
- **Ranking Candidate Entities and Linking**
  - Sub-class Unaware Similarity
    - Jaccard vs Cosine-TFIDF, Co-reference Resolution or not, Commentary context, Mention Context, Ball representation
    - Multi-ball mention: Knee and Max Average Sub-array
  - Sub-class Aware Slot-based Similarity
    - $Score(m,b) = \lambda\ UnstructuredSim(m,b) + (1-\lambda) SlotMatchScore(m,b,m_t)$
    - $Score(m,b) = \sum_{t \in T}\ P(t|m)(\lambda\ UnstructuredSim(m,b) + (1-\lambda) SlotMatchScore(m,b,t)\ )$
  - 4 ways of combining scores
    - MLE with Ball Filter
    - MLE with no Ball Filter
    - Bayesian with Ball Filter
    - Bayesian with no Ball Filter
- **Iterators**
- **Sequential Proximity**
  - minDiff, minRankDiff, minScoreReciprocalDiff

Results

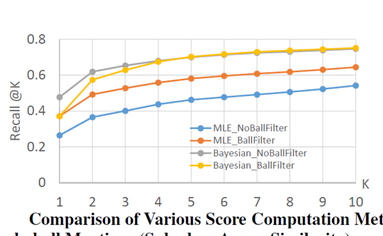
- 30 matches from 2011 Cricket World Cup,187 articles, 14 countries, 207 players, 15718 balls, 2828 labelled mentions
- Mention Type classifier accuracy=85%
- Mention Sub-class classifier accuracy=65% (multi-ball), 74% (single-ball), Dictionary features are most important.



Recall@K Comparison for Various Settings (Single-ball Mentions, Sub-class Unaware Similarity)

	Precision (P)	Recall (R)	F1
Best	0.269	0.369	0.311
Commentary Context=0	0.231	0.456	0.307
Commentary Context=1	0.230	0.444	0.303
Commentary Context=2	0.229	0.442	0.302
Mention Sentence	0.179	0.542	0.269
Unstructured Ball Representation	0.231	0.456	0.307
maxSubArrayAvg_1.5	0.242	0.393	0.300
maxSubArrayAvg_2	0.197	0.653	0.302
maxSubArrayAvg_3	0.231	0.456	0.307
maxSubArrayAvg_4	0.245	0.339	0.285

Precision, Recall and F1 Comparison for Various Settings (Multi-ball Mentions, Sub-class Unaware Similarity)



Comparison of Various Score Computation Methods for Single-ball Mentions (Sub-class Aware Similarity)

	Sub-class Aware + Iterator			Sub-class Aware + Iterator + Sequential Proximity		
	P	R	F1	P	R	F1
All	0.570	0.527	0.548	0.572	0.539	0.555
BAT	0.746	0.819	0.781	0.746	0.819	0.781
BATBOWL	0.398	0.700	0.507	0.398	0.700	0.507
BOWL	0.600	0.220	0.322	0.600	0.220	0.322
EXTRAS	0.101	0.125	0.111	0.101	0.125	0.111
FOUR	0.493	0.546	0.518	0.493	0.546	0.518
OTHERS	0.257	0.344	0.294	0.257	0.344	0.294
OVERS	0.370	0.722	0.489	0.379	0.714	0.495
PARTNERSHIP	0.710	0.645	0.676	0.710	0.645	0.676
POWERPLAY	0.677	0.162	0.261	0.677	0.162	0.261
REFERRAL-DROPPED	0.079	0.066	0.072	0.138	0.162	0.149
SIX	0.509	0.485	0.497	0.509	0.485	0.497
WICKETS	0.432	0.465	0.448	0.427	0.533	0.475

Accuracy Comparison between Sub-class Aware Method and Sequential Proximity for Multi-ball Mentions

	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
SA	0.466	0.568	0.619	0.652	0.672	0.693	0.709	0.719	0.728	0.737
SP	0.477	0.580	0.630	0.664	0.687	0.708	0.727	0.737	0.745	0.758

Recall Comparison between Sub-class Aware Method (SA) and Sequential Proximity (SP) for Single-ball Mentions

Method	Precision (P)	Recall (R)	F1
MLE_NoBallFilter	0.511	0.427	0.465
MLE_BallFilter	0.519	0.507	0.510
Bayesian_NoBallFilter	0.527	0.587	0.556
Bayesian_BallFilter	0.522	0.558	0.539
Bayesian_NoBallFilter+Iterator	0.568	0.566	0.567

Paper ID: 12

Microsoft