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.

**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, OTHERS.
  - Multi-ball mentions: BAT, BOWL, BATBOWL, FOUR, SIX, PARTNERSHIP, WICKETS, OVERS, POWERPLAY, REFERRAL-DROPPED, EXTRAS, 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**
  - 4 ways of combining scores
    - MLE with Ball Filter
    - MLE with no Ball Filter
    - Bayesian with Ball Filter
    - Bayesian with no Ball Filter

**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.

**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.

**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.

**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, OTHERS.
  - Multi-ball mentions: BAT, BOWL, BATBOWL, FOUR, SIX, PARTNERSHIP, WICKETS, OVERS, POWERPLAY, REFERRAL-DROPPED, EXTRAS, 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**
  - 4 ways of combining scores
    - MLE with Ball Filter
    - MLE with no Ball Filter
    - Bayesian with Ball Filter
    - Bayesian with no Ball Filter

**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.