0.1 The problem

Input file="input1.txt" source="The Hindu" dated="Monday, May 31, 2004"
Bangladesh posts highest ever total, Rafique emulates Bashar

ST. LUCIA, MAY 30. Mohammad Rafique hit an unbeaten, maiden Test century as Bangladesh notched up its highest ever total on a rainy second day of the opening Test against the West Indies at Beausejour Stadium.

The visitor, overnight 278 for seven, used Rafique's unbeaten 103 to power its first innings on Saturday to 406 for nine on a day shortened to just 41.1 overs ...

[Output]

<output input-document="input1.txt">
<sport-name>Cricket, Test</sport-name>
<event-start-weekday>Friday</event-start-weekday>
<site-name>Beausejour Stadium</site-name>
<team>
<team-name>Bangladesh</team-name>
<team-country>Bangladesh</team-country>
<team-stats-descr>Bangladesh notched up its highest ever total</team-stats-descr>
<team-stats-descr>The visitor, overnight 278 for seven</team-stats-descr>
<team-stats-descr>406 for nine</team-stats-descr>
</team>
<team>
<team-name>West Indies</team-name>
<team-country>West Indies</team-country>
</team>
<player>
<team-name>Bangladesh</team-name>
<player-name>Mohammad Rafique</player-name>
<player-country>Bangladesh</player-country>
<player-stats-descr>Mohammad Rafique hit an unbeaten, maiden Test century</player-stats-descr>
<player-stats-descr>unbeaten 103</player-stats-descr>
</player>
</output>

0.2 Possible tags

- person/player
- place– as a general place like Bombay /how do u distinguish between a person's name and a place (probably the context "in <place>" or "from <place> "
- stadium or site
- country – can do it with a gazette
- date - of every kind but mostly Jun 30 kind not 10-2-2003 BUT they actually want an event-start-day
- other stuff goes as description

- team-name
- sport-name - with a gazette ..should be able to gather this from the related words..in the above eg cricket is not explicitly mentioned anywhere
- number – to know scores etc,

Roughly these are the steps

1. De-HTMLtag the documents
2. Sentencify and named-entity tag the sentences
3. Train a "classifier" for each sentence

The idea :
After the html-cleaned sentences are tagged with our tags, we would be in a position to know which are 'persons','places','teams', 'dates' etc. in the sentence. Then the classifier shud depending on the entities and the context in which it occurs should be able to distinguish between what sentence talks abt the player, abt the team, abt the event, abt the statistics of the game etc. But if 2 teams are involved we should be able to relate the corresponding statistics. depending on the output of the classifier fill out the slots of the xml template

Some things we need

- web pages with sports articles
- walk through for few examples to check how things are working out

Some issues
I didn't follow much abt the Bayes classifier or EM algorithm, could gather that we might think abt using them while building the classifier.
Might need lots of manually-tagged examples while building the classifier