Jose Ramirez, Matthew Garber, and Xinhao Wang
February 9th, 2016
CS 140 Natural Language Annotation for Machine Learning
SoccEval Task Description

The underlying goal for our project is to evaluate the ability and performance of a soccer player from both objective and subjective descriptions of them and their actions. More specifically, we hope to be able to use these evaluations to summarize the salient attributes of a player, to distinguish between good, mediocre, and poor players, and to use these judgments to compare the media's portrayal of a player to that player's objective statistics.

The corpus for this task will be drawn from internet articles about professional men's soccer written by professional sports journalists, from websites such as ESPNFC.com, Goal.com, and the Guardian. This may include review on players as well as report on games. In order to capture the variation in player ability, we plan to use a large number of documents, though the exact number has yet to be determined. Additionally, we intend to use articles that analyze performances over a variety of time frames, such as within a game, during a portion of a season, or for the entirety of a season. We plan to only use articles discussing the 2014-2015 season, although we may use other recent seasons if focusing on a single season proves insufficient. The documents selected will not necessarily be limited to news coverage and may include editorials and opinion pieces as well. At first, we may focus on players of specific teams, including top 4 Chelsea, Man U, Arsenal, Man C (or Liverpool) and bottom 3 Hull, Burnley, QP Rangers.

Initially, we will annotate the mentions of players as well as link instances of anaphora back to the tagged players. Information relevant to evaluating a player's performance, such as their position, will also be noted at this time. Coreference annotation will not take place between multiple documents; if a player is mentioned in multiple documents they will be (temporarily) treated as a new player in each one. Player references will be consolidated after all of the annotation has finished. Though this level of coreference annotation may not be strictly necessary to analyze players mentioned within the corpus, it will be useful when machine learning techniques are utilized to apply this annotation scheme onto new documents.

Following this we will annotate phrases that describe a player's performance. The annotation will distinguish between a player's objective actions and experiences (such as scored goals, injuries, and fouls) and more qualitative, subjective analyses of that player (such as “is a highly effective player” or “has superb technique”). We will tag each type of description with a positive or negative polarity depending on whether it indicates a player is performing well (positive) or poorly (negative). Each subjective phrase will also be given a category (such as *skill* or *movement*) that indicates what ability or attribute it refers to. The relevant phrase will be linked to the most appropriate explicit or implicit mention of the player.

We may choose to do a portion of the initial annotation automatically. If a sufficiently consistent coreference tagger can be found, it may be beneficial to have our annotators correct and improve upon the automatically tagged text rather than start from scratch. It should also be possible to highlight relevant phrases (or at least words in those phrases) before the annotation process begins. This should be especially true of the objective descriptions, which will include common soccer terms, such as *score*, *pass*, and *assist*. Pre-annotating the subjective phrases may prove more difficult, though there may be words that occur frequently enough that doing so would be valuable, if inconsistent.