SoccEval: An Annotation Schema for Rating Soccer Players

Jose Ramirez **Matthew Garber** **Xinhao Wang**
Department of Computer Science
Brandeis University
{jramirez, mgarber, xinhao}@brandeis.edu

Abstract

This paper describes the SoccEval Annotation Project, an annotation schema designed to support machine-learning classification efforts to evaluate the performance of soccer players based on match reports taken from online news sources. In addition to factual information about player attributes and actions, the schema annotates subjective opinions about them. After explaining the annotation schema and annotation process, we describe a machine learning experiment. Classifiers trained on features derived from annotated data performed better than a baseline trained on unigram features. Initial results suggest that improvements can be made to the annotation scheme and guidelines as well as the amount of data annotated. We believe our schema could be potentially expanded to extract more information about soccer players and teams.

1 Introduction

The underlying goal of the SoccEval Annotation Project was to evaluate the ability and performance of a soccer player from both objective descriptions of their actions as well as subjective descriptions of the players themselves, using soccer news articles as a source. We used these attributes to rank players based on their overall quality.

Our annotation scheme was designed to support both these efforts by creating a corpus annotated with these descriptions in order to facilitate extraction of relevant features to rate players, as well as the most relevant attributes of individual players.

A previous soccer-related annotation scheme exists: the SmartWeb Ontology-based Annotation System (SOBA) which was designed to extract information on soccer-related entities, including players and events associated with them (Buitelaar et al., 2006).

However, SOBA only includes factual information about events. We created a player-specific annotation scheme that takes into account not only facts and events about a player, but also subjective evaluations, attaching a polarity value to these evaluations that can then be used not simply to extract information about a player, but to make judgments on the quality of the players.

2 Annotation Specification

To do the annotation task, our annotators used MAE (Multi-document Annotation Environment) (Rim, 2016), an open source, lightweight annotation tool which allows users to define their own annotation tasks and output annotations in standoff XML.

For annotation, MAE allows the creation of tags which define general categories. Tags then have attributes which serve as sub-categories from which a value can be selected. MAE supports the creation of two types of tags: extent tags and link tags. Extent tags mark a span of text, while link tags link two or more extent tags.

All extent tags have Spans and Text attributes. Spans refers to the range of indexes in the document for the text that an extent tag covers. Text contains the actual text.

This annotation project focuses on various descriptions and evaluations of soccer players. Descriptions from news articles can typically be divided into two types, facts and opinions[1]. Based

[1]This split between Fact and Opinion tags is inspired in part by the example of the MPQA Corpus (Wilson et al., 2016), which has separate Objective Speech Event Frames and Subjective Frames. The MPQA Corpus also inspired the use of Player IDs, as well as the decision not to impose strict rules for text span lengths.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop, pages 89–94
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-3015

on these categories, four extent tags and one link tag were created to capture the performance of a player.

The following 2 sample sentences will be used in explaining the tags in detail:

Sample sentence 1: Ward-Prowse almost levelled with a dangerous free-kick to the far corner that drew a fine save from Mignolet.

Sample sentence 2: Blessed with formidable speed and strength to go with his rare skill, the 25-year-old was always worth watching.

2.1 Player Tag

The Player tag is used to mark all mentions of a player directly by his name.

There are two attributes in the Player tag in addition to the default Spans and Text attributes. PlayerID is an ID that is assigned to each unique player. Name is an optional attribute created solely for the purpose of helping annotators distinguish players by entering any comments or notes they want for this Player tag.

2.2 Coref Tag

The Coref tag is an extent tag that is used to mark all references to a player by something other than his name. The Coref tag contains 3 attributes – Spans, Text and PlayerID. PlayerID is assigned the exact same ID as the player being referred to.

2.3 Fact Tag

The Fact tag is used to mark all text spans that describe events within a match that are connected to a player.

There are three attributes associated with this tag in addition to Spans and Text: Type, Time, and FactID. Type includes goal, assist, pass, shot, movement, positioning, substitute out, substitute in, injury, tackle, save and foul. The Time attribute is for represents the time of the event with relation to the match. Its possible values are: distance past, last season, current season, last match, present or future. FactID is generally unique. However, in certain cases where the same event is mentioned multiple times, the same FactID is assigned.

2.4 Opinion Tag

The Opinion tag is used to mark subjective attitudes toward a player.

There are five attributes associated with this tag besides Spans and Text: Type, Polarity, Time, Hypothetical, and Reported. Type groups different opinions into the following categories: soccer skill, accomplishment, general attribute, impact on team, growth or decline and other opinion. Polarity is the sentiment toward a player in this opinion tag, which can either positive or negative. The Time attribute is the same as that in Fact tag. The Hypothetical attribute is used only when the Opinion is either a prediction or counterfactive. The Reported attribute is a Boolean to distinguish if the Opinion is being reported by someone within the article, such as a secondary source who is not the writer of the article himself.

2.5 TargetLink Tag

TargetLink is a link tag that links a fact or opinion to a player or coreference tag.

2.6 Sample Annotation

Below is a simplified annotated version of the two sample sentences:

Annotated sample sentence 1:
[Ward-Prowse]$_{Player1}$ almost levelled with a dangerous [free-kick]$_{Fact:shot}$ to the far corner that drew a fine [save]$_{Fact:save}$ from [Mignolet]$_{Player2}$.

TargetLink:
T1: [free-kick] – [Ward-Prowse]
T2: [save] – [Mignolet]

Annotated sample sentence 2:
Blessed with
[formidable speed]$_{opinion:particularskill\_positive}$
and
[strength]$_{opinion:generalattribute\_positive}$
to go with
[his]$_{coref1}$
[rare skill]$_{opinion:particularskill\_positive}$,
[the 25-year-old]$_{coref2}$
was always
[worth watching]$_{opinion:otheropinion\_positive}$.

TargetLink:
T1: [formidable speed] – [his]
T2: [strength] – [his]
T3: [rare skill] – [his]
T4: [worth watching] – [the 25-year-old]

3 Corpus Selection and Annotation

Documents were taken from two sources, Goal.com[2] and The Guardian[3]. Initially, a total of 465 documents were collected, 361 of which were taken from The Guardian, while the rest were taken from Goal.com.

The articles focused on three clubs from the English Premier League: Chelsea, Tottenham Hotspur, and Liverpool. The majority of the articles were match reports, though there were also a few end-of-season player and team reviews as well. The final corpus included 34 documents taken from both sources, almost all of which were match reports covering games in which Chelsea had played (there was also one end-of-season player review).

While not part of the corpus per se, player ratings for the corresponding matches were retrieved from Goal.com. Each rating document measured the performance of each player during that match on a scale from 0.0 to 5.0, in increments of 0.5.

All the articles given out were connected to one team, Chelsea. This was done with the intention of making it easier for annotators to keep track of player names.

4 Annotation Guidelines

There are a few aspects of our annotation guidelines which are worth noting.

First, we gave annotators free choice in determining the length of the text span worth annotating. Since descriptions of players, especially subjective ones, come in many forms, we thought it would be best to leave that unspecified. We believed that nonetheless, annotators would generally agree on a rough span of text to be annotated, even if their spans were not exactly the same. We did note in the guidelines that Fact spans were likely to be noun phrases, while Opinion spans would most often either be noun phrases or verb phrases.

We recognized that our team of annotators was generally unfamiliar with soccer, though we assumed a basic knowledge. When dealing with unfamiliar terms, we instructed our annotators to research the unfamiliar terminology using Wikipedia, Google, or other online sources.

In practice, we realized that some of our Opinion attributes were more general than others, and some of the categories were likely to overlap: for example, an accomplishment could also serve as an example of a player's growth. In these cases, we instructed our annotators to follow a priority system from more specific attributes to more general ones. So in the example here, we would instruct our annotators to prioritize the less vague "accomplishment" attribute instead of the "growth/decline" one.

[2]http://www.goal.com/en-us

[3]http://www.theguardian.com/

5 Inter-Annotator Agreement

To evaluate inter-annotator agreement on our annotated corpus, we used Krippendorff's alpha (Krippendorff, 2004)

Tag	*IAA score*
Player	0.9728
Coref	0.5828
Fact	0.4735
Opinion	0.4041

Table 1: IAA scores for tags (Krippendorff's alpha)

Attribute	*IAA score*
Player::playerID	0.9197
Fact:: time	0.8971
Opinion::reported	0.7639
Opinion::polarity	0.6747
Fact:: type	0.6366
Opinion::time	0.6031
Fact::FactID	0.4991
Opinion::type	0.4997
Coref::playerID	0.4989
Opinion::hypothetical	0.4122
Player::name	NaN

Table 2: IAA scores for tags and their attributes (Krippendorff's alpha)

Regarding attributes for Fact tags, we had relatively good agreement on Fact type, which was important, as well as strong agreement on time, which was relatively easy for annotators to detect. Agreement in attributes for Opinion tags was lower compared to that in attributes of Fact tags, reflecting the wider degree of subjectivity, but perhaps also the higher degree of ambiguity in our annotation guidelines. However, we did obtain good agreement for polarity values, as well as reported speech attributes.The agreement in polarity

values was particularly important, since our machine learning experiments made use of polarities in creating features from the opinion tags.

Finally, the score for the Hypothetical attribute is misleading, simply because one of our annotators seems to have marked every Opinion tag with this attribute. Otherwise, we observed during adjudication that annotators were relatively consistent in marking Hypothetical attributes.

6 Adjudication Process

We included Fact tags in our gold standard if at least one annotator tagged it. Occasionally, if a span of text should obviously have been marked as a Fact but had not been tagged by any annotators, we nonetheless tagged it as a Fact in our gold standard. In many cases this involved relatively obvious readings of events such as goals, saves, and other facts which we believe the annotators should easily have caught according to our guidelines. We attempted to do this very sparingly, though. On the other hand, we only included Opinion tags if at least two annotators tagged a span. With regard to attributes, we generally opted for "majority rules". If there was complete disagreement about the attribute, we selected the one that to us seemed most appropriate.

We usually selected the span that the majority of annotators agreed on, which usually was the minimal relevant span.

7 Experiments

An experiment was performed using the previously mentioned player ratings. Players that were explicitly mentioned in a document were classified by the rating obtained from Goal.com.

7.1 Baseline

Three types of baseline models were trained utilizing Scikit-learn (Pedregosa et al., 2011) embedded in NLTK (Bird et al., 2009) wrappers: a support vector machine (SVM) model, a maximum entropy (MaxEnt) model, and a decision tree (DT) model. All baseline models were trained with boolean unigram features, though stopwords were removed before feature extraction. No dimension reduction was performed other than what inherently occurred in each type of model.

For each match report, a sub-document was created for each player mentioned in the match report. Each player's sub-document included every sentence explicitly mentioning that player's name. In a naive model of coreference, sentences containing anaphora were added to the sub-document of the most recently mentioned player. Each sub-document was paired with the rating for that player for that match.

Micro-precision was high for all models, though this was largely due to the fact that they tended to predict a score of 3.0, which was by far the most common player rating. The MaxEnt and Decision Tree models performed roughly equally well, though neither could be considered a successful model.

It is worth noting that no model was able to predict ratings at the high and low extremes due to a sparsity of data for the ratings.

Classifier	Precision	Recall	F1
SVM (Micro)	0.327	0.327	0.327
SVM (Macro)	0.0764	0.169	0.0968
MaxEnt (Micro)	0.297	0.297	0.297
MaxEnt (Macro)	0.121	0.163	0.127
DT (Micro)	0.281	0.281	0.281
DT (Macro)	0.15	0.166	0.148

Table 3: Scores for different baseline classifiers

Rating	Precision	Recall	F1
2.0	0.0294	0.0294	0.0294
2.5	0.121	0.154	0.128
3.0	0.345	0.464	0.375
3.5	0.324	0.327	0.307
4.0	0.159	0.115	0.126

Table 4: Scores for Decision Tree baseline by rating [5]

7.2 Classifiers

Different types of classifiers were applied to the annotated corpus, including maximum entropy (MaxEnt), linear regression (LR), support vector machine (SVM) and random forest (RF). Precision, recall and F1 score were calculated for each classifier, with 17-fold cross-validation, which tested 2 files each time. Since regression predicts a continuous scaling measure instead of a discrete 5 point scale, the prediction of a regression was converted to the nearest rating point. For example,

[5] Scores for ratings not shown were all 0.0.

if linear regression output 3.33, it was converted into 3.5.

7.3 Feature Extraction

Multiple attempts were made to achieve a better score. In the initial attempt, the following features were used:

- Normalized percentage of different types of facts in a single article
- Normalized percentage of different types of opinions in a single article
- Total mentions of each player in a single article

The following issues have also been taken into consideration and the model is slighted adjusted accordingly.

Correlation: There were certain degrees of correlation between some features, though due to the limited amount of data these correlations were unstable. However, removing one of two significantly correlated features made no notable improvement in the accuracy of the classifiers.

Dimension reduction: In order to remove redundancies in the features, singular vector decomposition was applied to the feature matrix before doing linear regression. However, linear regression with SVD actually performed slightly worse than linear regression without SVD.

LR, SVM and MaxtEnt performed equally well in terms of their micro-averages, although MaxEnt achieved the best score, 0.367, by a very small margin. While this was only slightly better than baseline, the macro F1-score for the LR model was 0.204, which was a more notable improvement.

Classifier	Precision	Recall	F1
LR (Micro)	0.364	0.364	0.364
LR (Macro)	0.219	0.252	0.204
LR-SVD (Micro)	0.328	0.328	0.328
LR-SVD (Macro)	0.216	0.216	0.187
SVM (Micro)	0.363	0.363	0.363
SVM (Macro)	0.206	0.233	0.194
MaxEnt (Micro)	0.367	0.367	0.367
MaxEnt (Macro)	0.147	0.219	0.160
RF (Micro)	0.283	0.283	0.283
RF (Macro)	0.176	0.188	0.171

Table 5: Scores for different classifiers

8 Challenges

8.1 Challenges in Annotation

One issue with the annotation process was the use of British English and soccer jargon in match reports. Annotators who are not familiar with British English vocabulary and soccer terms reported difficulties in understanding some of the match reports.

Another issue was the ambiguity between certain categories in the annotation scheme. For example, in Fact tags, type "assist" and type "goal" are a subsets of "pass" and "shot" respectively. In Opinion tags, "accomplishment" overlaps "growth/decline", since accomplishments are often indicative of a player's improvement.

The lack of precision in the annotation guidelines regarding the span of the text to be tagged resulted in wide disagreements over spans.

Finally, some of the categories were not often used by the annotators. This mainly resulted from the fact that we initially designed our DTD based on the categories found in match reports and player reviews from the Guardian, which include more opinions and subjective judgments. However, the Goal.com match reports focused more heavily on reporting facts, with few subjective judgments on the part of the writer. However, if we were to expand the corpus to include a more diverse range of sources, we might see cases where Opinion tags would be useful.

8.2 Challenges in Machine Learning

One issue was the limited amount of annotated files. This directly led to unstable results where in some cases, certain features are strongly correlated or the F1 score exceeds 0.6, while in other cases, the features have no correlation at all or the F1 score is lower than the baseline.

The second issue was whether the features being extracted are fundamentally a good predictor for a players rating. Since the rating is based on the actual performance of a player, and the match reports will not cover every detail happened in a match, this incomplete description may or may not be sufficient to predict the rating accurately. In addition, the ratings were collected from one of the sources from which the corpus was built, which may contain its own bias.

Furthermore, as the ratings themselves are determined by sports writers, they are themselves inherently subjective and problematic as a gold

standard label, since two different writers might disagree on a rating for a specific player. The Goal.com ratings that we used as a reference label are themselves created by the Goal.com staff and factor in sub-ratings in subjective traits such as 'vision', 'work rate', and 'killer instinct'. Unless we use hard data only as a criterion for determining ratings (ie. counts of specific actions like appearances, goals, saves, etc.), the ratings themselves which we are evaluating will be unreliable as a gold standard label. One possible solution to obtain more agreement on labels might be to restrict the number of labels to two or three instead, instead of going by increments of 0.5. That might help obtain a more reliable gold standard for labels, since there would likely be more agreement on star players vs. terrible players, as opposed to the difference between a 3.0 and a 3.5. We might lose a certain level of granularity, but our labels would likely be more grounded in reality.

Another issue is the methodologies of the classifiers. Discriminant classifiers or decision trees treat ratings as a nominal measure. Therefore, the interval information of ratings will be lost. Although regression keeps such information, it has a stricter requirement for the relationships among features and the target in order to get a better result.

9 Conclusion

This annotation project focuses on a player's performance as described by soccer news articles. By capturing the actions of a particular player as well as subjective evaluations about them, a rating prediction can be made. Models based on the current scheme performed appreciably better than the baseline. However, they still did not perform particularly well, due to the factors mentioned above.

Increasing the corpus size and variety on players performances and ratings are two changes that can be made in the future which would potentially give a more stable result. We might potentially change the rating system to restrict the number of labels, as mentioned above.

We can also improve the current annotation scheme by narrowing the number of fact or opinion types and eliminating redundant attributes. We can select annotators who are knowledgeable enough about soccer to easily understand match reports. Alternatively, in order to lower the cognitive load caused by unfamiliarity with the sport and its jargon, we can create an appendix within the guidelines introducing annotators to the basic rules and vocabulary of soccer.

In terms of further applications, this project can be expanded to include a model for rating teams. If we apply syntactic parsing, we could also extract salient characteristics of players to determine what makes a good player. Finally, in addition to ratings, external statistics of a player, such as transfer value, salary, growth/decline, etc., could also be incorporated into the model to provide a more comprehensive summary of a player.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Paul Buitelaar, Thomas Eigner, Greg Gulrajani, Alexander Schutz, Melanie Siegel, Nicolas Weber, Philipp Cimiano, Günther Ladwig, Matthias Mantel, and Honggang Zhu. 2006. Generating and visualizing a soccer knowledge base. In *EACL '06 Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*. pages 123–126. http://www.aclweb.org/anthology/E06-2010.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality & quantity* 38:787–800.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12:2825–2830.

Kyeongmin Rim. 2016. Mae2: Portable annotation tool for general natural language use. In *Proceedings of 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. pages 75–80.

Theresa Wilson, Janyce Wiebe, and Claire Cardie. 2016. Mpqa opinion corpus. In James Pustejovsky and Nancy Ide, editors, *Handbook of Linguistic Annotation*, Springer, New York.