SoccEval Annotation Project Review

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 scheme designed to capture factual information as well as subjective opinions regarding soccer players, in order to evaluate the performance of players from articles written about them. After explaining the annotation scheme and annotation process, we then describe a machine learning experiment that demonstrates both the promise of our scheme, as well as flaws in the complexity of its design.

1 Introduction

The underlying goal of the SoccEval Annotation Project is 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 plan to use these attributes to rank players based on their overall quality. After that, we hope to use these evaluations to summarize the salient attributes of a player and to distinguish between good, mediocre, and poor players.

Our annotation scheme is 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). SOBA was created with the intention of populating a knowledge base for use in a multimodal dialog system.

However, SOBA only includes factual information about events. We intend to create 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

This annotation project focuses on various descriptions and evaluations of soccer players. News articles can typically be divided into two types, fact and opinion. Fact is for all kinds of objective description toward an event or state while opinion is for any comment that contains subjective attitude. In this particular project, a players performance can also be categorized into these two groups[1]. Based on this, 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.

[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.

2.1 Player Tag

Player tag is an extent tag that is used to mark all mentions of a player directly by his name. For example, Player tags should be created for “Ward-Prowse”and “Mignolet”in sample sentence 1.

There are four attributes in Player tag. Just like all extent tags, Spans are the index ranges for this particular tag and Text is for the text itself being tagged. PlayerID is an important and unique ID that is assigned to each unique player. A player may be mentioned by either his first name or last name or full name or other names, but the PlayerID is all the same. 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.

Player Tag
Spans
Text
PlayerID
Name (optional)

Table 1: Player Tag

2.2 Coref Tag

Coref tag is an extent tag that is used to mark all instances of coreferences for players. Anytime a player is referred to by something other than his name, it is tagged as Coref. For example, Coref tags should be created for “his”and “the 25-year-old”in sample sentence 2.

Coref tag contains 3 attributes – Spans, Text and PlayerID. Spans and Text function exactly the same in Player tag. PlayerID is assigned the exact same ID as the player being referred to.

Coref Tag
Spans
Text
PlayerID

Table 2: Coref Tag

2.3 Fact Tag

Fact tag is an extent tag that is used to mark all phrases that describe events that happened during a match. These phrases are objective descriptions without carrying any sentiments with them. For example, Fact tags should be created for “free-kick ”and “save”in sample sentence 1.

There are six attributes associated with this tag. Spans and Text function exactly the same as in other extent tags. Type is an attribute that categorizes different types of an event, which includes goal, assist, pass, shot, movement, positioning, substitute out, substitute in, injury, tackle, save and foul. The Time attribute is for recording the time the event actually happened. 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.

Fact Tag
Spans
Text
Type
Time
FactID

Table 3: Fact Tag

2.4 Opinion Tag

Opinion tag is an extent tag that used to mark certain subjective attitude toward a player. For example, in sample sentence 2, “formidable speed”, “strength”, “rare skill”and “worth watching”should be tagged as opinion.

There are seven attributes associated with this tag. Spans and Text function exactly the same as in other extent tags. 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 it 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. For example, in sample sentence 1, the fact tag for “free-kick”is linked to

Opinion Tag
Spans
Text
Type
Polarity
Time
Hypothetical (optional)
Reported

Table 4: Opinion Tag

the player tag for "Ward-Prowse", and the fact tag for "save"to the player tag for "Mignolet".

The attributes for TargetLink tag simply takes the unique system IDs of those tags, as well as the texts.

TargetLink Tag
Fact or opinion system ID
Fact or opinion text
Target system ID
Target text

Table 5: TargetLink 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

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 premiere 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.

4 Annotation Process

34 articles were given to a team of 3 annotators. Most of the articles consisted of match reports from both the Guardian and Goal.com, with the exception of one end-of-season review consisting of opinions about the best player. The articles were given out over a period of five weeks, with one package of 6-10 articles given out each week. 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 (and avoid confusing them with manager or referee names).

To do the annotation task, our annotators used MAE (Multi-document Annotation Environment) (Rim 2016). MAE is an open source, lightweight annotation tool which allows users to define their own annotation tasks, markup arbitrary text spans, and use non-consuming tags, as well allowing them to easily create links between annotations and output annotations in stand-off XML.

We remained in contact with our annotators through Facebook Messenger and through email, taking questions from annotators and if necessary, making small adjustments to the annotation guidelines in order to clarify confusing points.

[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.

The general formula for Krippendorffs alpha is as follows:

$$\alpha = 1 - (D_o/D_e)$$

where D_o is the observed disagreement and D_e is the expected disagreement between annotators.

Krippendorff's alpha can measure agreement between more than 2 annotators and can measure agreement regarding the length of a tagged text span. It allows partial credit to be given for partial agreement between annotators on the length of a span. This makes Krippendorff's alpha a suitable IAA metric for our annotation task, since our task gives annotators freedom in determining the length of text that constitutes a fact or an opinion.

Tag (:: attribute)	*IAA score*
Player	0.9728
Player::name	NaN
Player::playerID	0.9197
Coref	0.5828
Coref::playerID	0.4989
Fact	0.4735
Fact::FactID	0.4991
Fact:: time	0.8971
Fact:: type	0.6366
Opinion	0.4041
Opinion::hypothetical	0.4122
Opinion::polarity	0.6747
Opinion::reported	0.7639
Opinion::time	0.6031
Opinion::type	0.4997

Table 6: IAA scores for tags and their attributes (Krippendorff's alpha)

Most player tags had been created during preprocessing, so it is unsurprising to see such a high score for them. In the case of Coref tags, the lower level of agreement is most likely due to the fact that the annotation guidelines did not explicitly specify how to handle possessive pronouns ('his'/'their').

Agreement on Fact and Opinion tags was lower than ideal, but not terrible, considering the greater subjectivity involved in making judgments about both. With Fact tags, a great deal of the disagreement was most likely due to the difference in span lengths tagged. With Opinion tags, in addition, there were more cases where spans of text were judged to be an opinion by one annotator and not by others.

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. In the case of the Time attribute, the lower score when compared to the Fact tags might be connected to the difficulty in interpreting certain phrases, or the ambiguity within the documents themselves regarding the timeframe of a certain opinion. 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

Our group faced a few challenges during the adjudication process.

The first challenge was the relative lack of agreement on what constituted an opinion, as well as frequent annotator disagreement over fact and opinion attribute types. We also faced situations in which most or all of our annotators did not tag text spans that we believed should have clearly been tagged according to our annotation guidelines.

Here we faced a conflict in our gold standard between trying to obtain an accurate text that reflected the annotation guidelines according to us, and obtaining an accurate impression of agreement between the annotators themselves that reflects their perception of the annotation guidelines. Ultimately, we opted to prioritize the latter.

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. Because opinions are by definition more subjective, we assumed that there would be far less agreement on what constitutes an opinion, even with our annotation guidelines. As a result, we were much more cautious about including opinions in our gold standard unless there was substantial agreement between annotators.

With regard to attributes, we generally opted for "majority rules", choosing the attribute most frequently marked. If there was complete disagreement about the attribute, we selected the one that to us seemed most appropriate.

Our next challenge was in the choice of span length to include in our gold standard. Since our instructions gave annotators a high amount of freedom to choose the span length for facts and opinions, the spans were of varying lengths, with some annotators choosing the minimal span, and others the maximal one.

In this case, we usually selected the span that the majority of annotators agreed on, which usually was the minimal relevant span.

7 Machine Learning

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. From features extracted from that document, models were trained to classify players'performances during other matches.

7.1 Baseline

Three types of baseline models were trained utlizing Scikit-learn (Pedregosa et al., 2011) embedded in NLTK wrappers: a supported 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 extemes 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 7: 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 8: Scores for Decision Tree baseline by rating [4]

7.2 Classifiers

With the same wrapper packages as in baseline model, 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). These classifiers were chosen due to the differing features and capabilities they offer. Precision, recall and F1 score were calculated for each classi-

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

fier, with 17-fold cross-validation, which makes 2 files be tested 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, if linear regression output 3.33, it was converted into 3.5. Hence, the measurement on accuracy can be compared across different models.

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 limited amount of data these correlations were unstable. However, removing one of two significanly correlated features, such such as the percentage of shots or total mentions of a player, 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.

8 Challenges

There were some difficulties and challenges over the course of the project, both during the annotation process as well as the machine learning experiments, which may partly account for the results obtained from our models.

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 9: Scores for different classifiers

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. Our annotation guideline recommended searching for unfamiliar terms on Google or Wikipedia, but this was not enough in some cases. Furthermore, some terms were misunderstood by our annotators who assumed they had a different meaning, leading them to select an incorrect attribute. For example, the term "centreback", which is a soccer player's position and often serves as a descriptive coreference for the player, was misinterpreted by at least one annotator as a Fact tag with a Movement attribute.

Another issue was the ambiguity between certain categories in the annotation scheme. For example, in Fact tags, type "assist" and type "goal" is a subset of "pass" and "shot" respectively. In Opinion tags, "accomplishment" overlaps "growth/decline", since accomplishments are often indicative of a player's improvement. The priority list to resolve the cases where multiple tags are applicable also creates additional cognitive load for annotators.

The lack of precision in the annotation guidelines regarding the span of the text to be tagged resulted in wide disagreements over spans. Our instructions were to "tag as much text as is connected to a specific event" , which encouraged a maximal interpretation. In practice though, annotators often tagged only the minimal spans, especially for Fact tags. Furthermore, our instructions imposed a greater cognitive load on annotators by asking them to annotate differ-

ent lengths depending on the tag and attribute type. For example, we usually encouraged annotators to tag whole verb phrases for most Opinion tags. However, for the "soccer skills" and "general attribute" attributes we asked them to restrict themselves to the noun phrase containing the key vocabulary for the attribute (i.e. "the $[\text{agility}]_{opinion:general\_attribute\_positive}$ of their goalkeeper.").

Finally, some of the categories were not often used by the annotators. This mainly results 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 used a different writing style, which focused more heavily on reporting facts, with few subjective judgments on the part of the writer. As a result, Opinion tags were rarer in the Goal.com articles. However, if we were to expand the corpus to include a more diverse range of sources, we might see cases where Opinion tags are useful.

8.2 Challenges in Machine Learning

One issue was the limited amount of annotated files. This directly results in 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 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 the one of the sources from which the corpus was built on, which may contain its own bias.

Another issue lies in 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 Summary

This annotation project focuses on a players performance in 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, it 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 can also improve the current annotation scheme by narrowing the number of fact or opinion types and eliminating redundant attributes, or even only focus on either fact or opinion and make the other as an attribute. In addition, in order to lower the cognitive load caused by unfamiliarity with the sport and its jargon, we would create an appendix within the guidelines introducing annotators to the basic rules and vocabulary of soccer. Our first iteration of the annotation cycle provided us with valuable feedback on the terms and aspects of soccer that are not intuitive to a non-fan, which will make it easier to create an effective guide to the sport.

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.

Acknowledgments

This project was part of the coursework for COSI140B: Natural Language Annotation for Machine Learning. We are grateful to Professor James Pustejovsky, Teaching Assistant Keigh Rim, and our three annotators Justin Su, Kelley Lynch and Yuanyuan Ma, for their help on this project.

References

Fabian Pedregosa, Gael 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 Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, vol 12, pages 2825–2830.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality & quantity*, vol 38, pages 114–133.

Kyeongmin Rim. 2016. MAE2: Portable Annotation Tool for General Natural Language Use. *Proceedings of 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 75-80.

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

Theresa Wilson, Janyce Wiebe and Claire Cardie. 2016. MPQA Opinion Corpus. To appear in James Pustejovsky and Nancy Ide (eds.), *Handbook of Linguistic Annotation*, New York: Springer.