SoccEval Annotation Project

Jose Ramirez
Matthew Garber
Xinhao Wang

Content

- Task goals
- Annotation Spec
- Characteristics of Dataset
- Difficulties
- Inter-Annotator Agreement
- Machine Learning
- Summary

Task Goals

Our task goal is...

- To rate soccer players, using news articles as a source, based on
  - Descriptions of their actions
  - Subjective evaluations of the players

- Extract the most salient attributes of a player in order to create a player summary

Annotation Specification

Five types of tag:

- Player
- Coref
- Fact
- Opinion
- TargetLink

Annotation Specification

- Player
  - Mentions of individual players by name (first, last, or both)
  - Annotators assign each distinct player a player ID
- Coref
  - Mentions of players by something other than name (ie. pronouns, descriptive noun phrases)

Annotation Specification

- Fact
  - Attributes: span, fact type, time
  - Covers actions concerning a player (goals, passes, substitutions, fouls, etc.)
  - Each fact also assigned a distinct ID (some facts mentioned more than once)

Annotation Specification

- Opinion
  - Tags all text where a subjective opinion of a player is expressed
  - Type Attributes: soccer skill, accomplishment, general attribute, impact on team, growth/decline, other
  - Annotators were asked to follow the following priority list in case of multiple possible tags:
    - soccer skill , accomplishment > general attribute > impact on team > growth/decline > other opinion

Annotation Specification

- Opinion (continued) -- other attributes
  - Sentiment polarity : + / -
  - Time: same as in Fact tags -- last match, current season, last season, distant past, present, future
  - Hypothetical:
    - Prediction
    - Counterfactual
  - Reported: for reported speech

Annotation Specification

- TargetLink
  - Links player or coref tags to Fact and Opinion tags associated with them

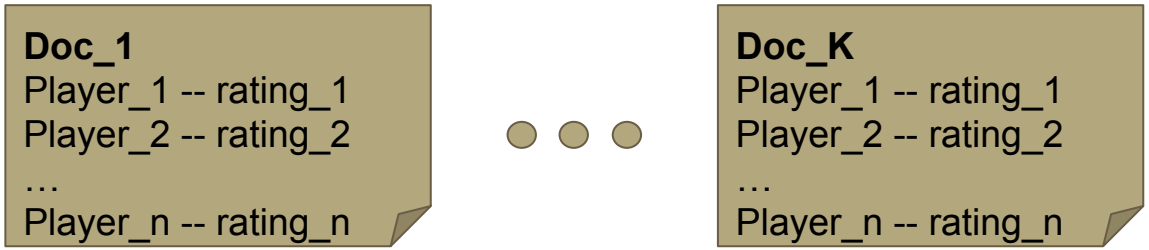
Annotation Specification

[Ward-Prowse]*player* almost levelled with a dangerous [free-kick]*fact:shot* to the far corner that drew a fine [save]*fact:save* from [Mignolet]*player*.

Blessed with [formidable speed]*opinion:particular_skill - positive* and [strength]*opiniongeneral_attribute - positive* to go with [his]*coref* [rare skill]*opinion:particular_skill - positive*, [the 25-year-old]*coref* was always [worth watching]*opinion: other opinion - positive*.

Characteristics of Dataset

- Articles taken from The Guardian and Goal.com (361 from the Guardian, 104 from Liverpool)
- Focused on 3 clubs from the English Premier League: Chelsea, Tottenham Hotspur, Liverpool
- Most of the articles are match reports, a few are end-of-season player or team reviews
- 34 articles were annotated
- Ratings were associated to each player



Difficulties

- For annotators, unfamiliar language was a problem in many documents.
  - British English
  - Soccer Terms
- For our scheme:
  - Not sufficient to catch everything
  - No easy solutions for ambiguity
- Span of the tag
  - Attempted to direct annotators to be conservative, by the guidelines were not clear enough on how to accomplish this.
- Ambiguity between the different categories
  - Some categories saw little intended use since the composition of our corpus was different than originally envisioned.

Difficulties

- Subjective evaluations of actions always run the risk of disagreement.
  - Limited range of polarity values likely helped these stay consistent.
- For adjudication:
  - Cautious when creating new tags.
  - However, sometimes things were clearly missed due to lack of soccer domain knowledge.
  - Developed conventions to stay consistent.

Inter-Annotator Agreement

To measure IAA, we used alpha-U:

- Our guidelines allowed annotators some freedom in determining the extent of relevant tags
- So, we needed a metric that could measure agreement about spans and allow partial credit for partial agreement on spans between annotators
- As well as measure agreement in labeling attributes of tags

Inter-Annotator Agreement -- Alpha-U

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

Machine Learning: Baseline

- Ratings for each player for each match were obtained from goal.com.
  - Ranged from 0.0 to 5.0 in increments of 0.5.
- A sub-document was created for each player mentioned in the match report.
  - Each sub-document contained every sentence explicitly mentioning a player's name.
  - 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.
- Trained baseline on boolean unigram features.
  - Stopwords were removed before extracting features.

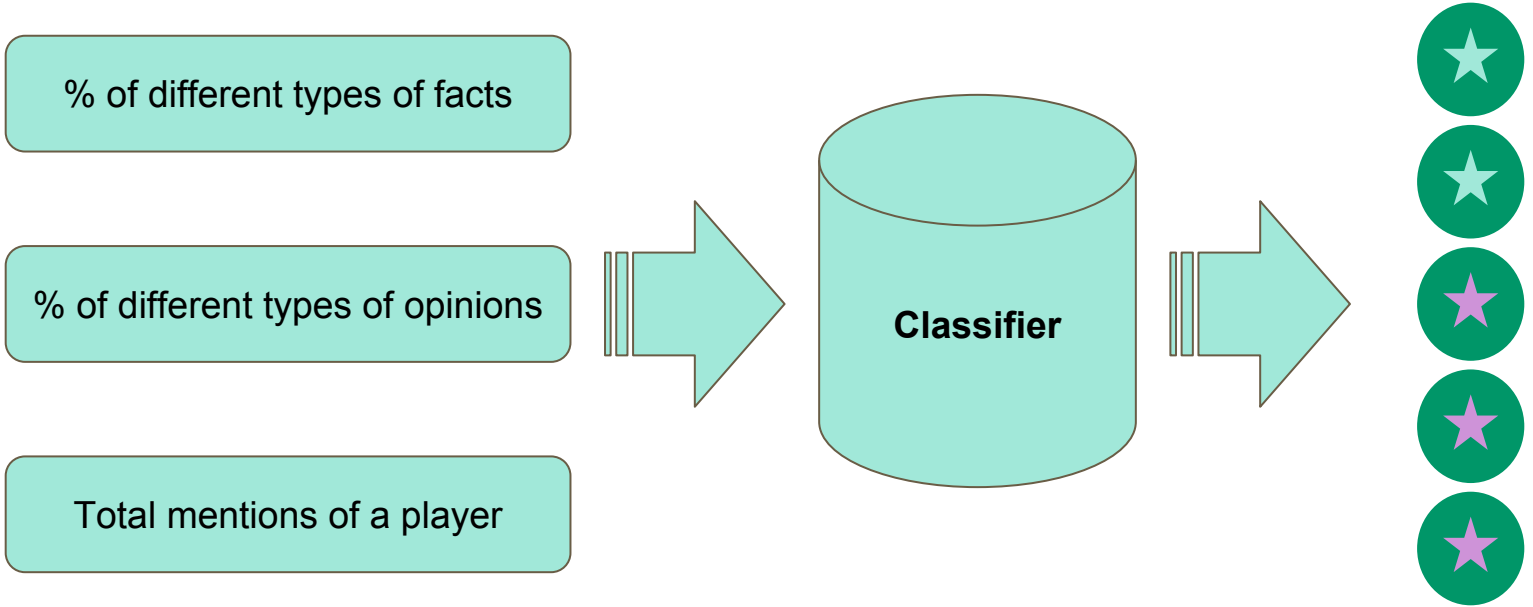
Machine Learning: Baseline Results

- Trained multiple baseline models, tested using k-fold cross-validation with k=10.
- MaxEnt was superior to both SVM and Decision Tree models.

MaxEnt Baseline Classifier			
Rating*	P	R	F
2.0	0.085	0.049	0.06
2.5	0.27	0.28	0.27
3.0	0.37	0.51	0.43
3.5	0.27	0.3	0.28
4.0	0.2	0.15	0.14
Macro Average	0.14	0.15	0.14
Micro Average	0.31	0.31	0.31

* Scores for ratings not shown were 0.

Machine Learning: Adding Features



Machine Learning: Adding Features

		Precision	Recall	F1
Linear Regression	Micro	0.362	0.362	0.362
	Macro	0.212	0.242	0.201
SVR	Micro	0.360	0.360	0.360
	Macro	0.196	0.223	0.187
Maximum Entropy	Micro	0.364	0.364	0.364
	Macro	0.140	0.209	0.153
Random Forest	Micro	0.296	0.296	0.296
	Macro	0.163	0.191	0.164

Note: regression results converted into nearest rating point

Machine Learning: Adding Features

		Precision	Recall	F1
Linear Regression	Micro	0.362	0.362	0.362
	Macro	0.212	0.242	0.201
Linear Regression with SVD	Micro	0.309	0.309	0.309
	Macro	0.170	0.191	0.157

Note: regression results converted into nearest rating point

Machine Learning Challenges

- Quality
  - A few ratings not properly linked
  - Potential errors in gold standard
- Quantity
  - Trained on 31 docs, tested on 2
- Relevance
  - Features not a perfect predictor
  - Ratings from one source, articles from two
- Classifiers
  - MaxEnt, Linear regression > Decision Tree
  - Nominal vs Interval
  - Correlation
  - Feature importance ranking unstable

Summary

- Though our model performed appreciably better than baseline, it still did not perform particularly well.
- However, it is worth noting that our model was trained with relatively few features on much less data.
- At the very least, our experiment shows that this path is worth exploring further.

Summary: Possible Future Changes

Annotation specification and guidelines

- Narrowing the number of attributes
- Eliminating time category, reported category, Fact IDs -- replace with linktag
  - Most of these occurred far too rarely to be significant.
- Including a glossary for terms -- an introduction to soccer
  - Terms were difficult to determine ahead of time.
  - It would be much easier to compile a glossary now.
- Stricter rules on spans -- minimal span (include noun, modifier)
- For match reports, opinion tag might not be necessary, and can be added as an attribute in fact tag, just like a sentiment

Summary: Possible Future Changes

Data and Modelling

- More variety of teams/players
  - Only focused on a single team (and their opponents)
    - This was likely helpful for annotators
    - However, data was obviously not independent.
- Find other sources of ratings.
  - goal.com ratings were useful, but split into too many categories.
  - Samples for the high and low ends of the rating system were too infrequent.
    - Simply not enough data to extract meaningful features.

Summary: Further Applications

- Extract salient characteristics of players to determine what makes a good player
- Team-level rating
- Using application to determine transfer value