FootballWhispers: Transfer Rumour Detection

Neil Ireson[1] and Fabio Ciravegna[1]

Sheffield University, Sheffield, UK

Abstract. Social media has been shown to have potential to predict various real world events, such as movements in the stock market and the outcomes of political elections. In this paper we present the Football Whispers (FW), a website dedicated to fans discussing transfer rumours. The unique selling point of the site is that it provides a crowdsourced assessment of those rumours, measuring the relative likelihood of a player's movements from social media chatter. This talk will focus on the rumour identification process, highlighting the role of open knowledge graphs and linked data to augment a domain knowledge-based to enable effective Named Entity Linking in noisy, informal social media messages.

1 Introduction

Football is the world's most popular sport, it is responsible for generating the majority of sporting revenue, in excess of $20 billion per annum, and it is estimated that over a billion people can be counted as football fans. Social media has become an important tool for maintaining the relationship between fans and their teams. With Twitter activity assured by both the substantial global fan base and dedicated localised fans. Football Whispers (FW) (www.footballwhispers.com) provides a service focused on the rumoured transfer of players between teams. A fundamental feature of FW is the estimation of the veracity of the rumours; providing a relative likelihood of a player moving to a given team. The likelihood is determined using the Twitter conversations concerning transfers, this involves two processes: Named-Entity Linking (NEL); and the assessment and combination of rumours to determine their relative likelihood.

2 Named Entity Linking & Rumour Detection

The NEL process involves the identification of entities (people and teams) mentioned in tweets and the disambiguation of those mentions; linking them to specific instances in a domain knowledge base (KB). The inherent issues when dealing with such short text with a high degree of abbreviated, informal language use is accentuated by the international, and thus multilingual, conversations concerning football. The initial domain KB is provided by Opta Sports (http://www.optasports.com/), while Opta is arguably the de facto source of football related information it only provides the official and possibly a single more common name. The 41,238 active players, from 3,266 teams, have only

46,631 name variations; with 17,489 (42%) players sharing a last name and 785 (2%) have identical names. In order to increase the number of alternative names the Opta entities are mapped to Wikidata and DBpedia entities. Wikidata contains 210,375 players and 30,710 teams, although a large number of the players are inactive, in order to identify potentially ambiguity it is necessary to include all names which may be mentioned. In addition to players, 11,439 managers, pundits, referees, etc. are also extracted. The talk will describe the entity mapping process, which considers the similarity of entities' available features, e.g. string similarity of names and numerical distance of dates, with the importance of a features being weighted according to the degree of variation in its values. The mapping process was also applied to DBpedia, which contains 126,790 players; this resulted in a slight increase in alternative names, primarily due to the DBpedia extraction of nicknames. In total 32,754 (80%) of Opta players are mapped, and for these players name variations are tripled to 95,535.

The team and player names are then used to generate a Deterministic Finite Automata to efficiently extract candidate entity mentions from the message text. The talk will describe how the contextual disambiguation processes are used to link mentions to an entity instance, where other entity candidates in the message provide the context. A name which occurs frequently in small number of (expected) contexts (e.g. player name mentioned only with their team) is deemed to maintain its meaning outside those contexts, while a name which occurs in multiple (unexpected) contexts is deemed too ambiguous to be used for entity linking when not contextualised. In addition, the message language is also considered, as names can be ambiguous within a given language context.

Evidence for a rumour is given by a message containing player and team entities, and at least one transfer term. The talk will briefly outline the four determinants of the veracity of a rumour: consensus (amount of evidence), recency/constancy (evidence time decay), authority (evidence sources) and coherence/consistency (evidence is not contradictory).

3 Football Whispers

In order to select tweets, which belong to the football domain, team names are used to filter the messages, this results in between 1-2 million tweets per day, and despite only English team names being used in the filter almost 56% of messages are in other languages. The rumour detection system processes the messages in real-time and the resultant rumours and likelihoods are validated by FW experts before appearing on the website. The talk will present the evaluation of the rumour detection and show how the use of knowledge graph (KG) data has led to significantly increased player NEL performance, and identification of the rumours concerning actual 2016 football transfers. Primarily this is due to the availability of multilingual data in the KGs and the use of language agnostic statistical disambiguation techniques in NEL. The success of FW (which currently has two million users) has now led to the developed of Sports Whispers and the application of this approach to other sporting domains.