Football Whispers

Transfer Rumour Detection

Fabio Ciravegna¹, Neil Ireson¹, Vivion Cox² and Kieren Beltrame²

¹ Department of Computer Science, The University of Sheffield
² FootballWhispers.com
The FW Solution

Millions of Messages

- Filtering
- Entity Extraction
- Entity Linking
- Prediction
- Rumour Identification
- Event Extraction
- Human Analysis
- Customer Feedback
Initial Knowledge Base

- **OptaSports**
  - 41,238 active players from 3,266 clubs
  - 46,631 player name variations
    - 17,489 (42%) common last name (Smith, da Silva, Rodríguez, González, …)
    - 785 (2%) have identical names

- **Misspellings and variations**
  - Fábio Pereira da Silva -> Fábio
  - Jay Dasilva -> Jay Da Silva
  - Santiago Cazorla -> Santi Cazorla
  - Patrick McNair -> Paddy McNair

- **Language/Transcription/Spelling variations**
  - Liverpool (FC) -> Ливерпуль (футбольный клуб)
  - Andriy Yarmolenko -> Андрей Ярмоленко
  - Henrikh Mkhitaryan -> Henrich/Henrik Mchitarjan/Mkhitarian
Integrating the Web of Data

• Wikidata
  • 210,375 (all) players from 30,710 clubs
  • 272,291 player name (language) variations

• DBpedia
  • 126,790 (all) players from 18,351 clubs
  • 353,001 player name (language & nickname) variations
Mapping Knowledge Bases

Name/Surname

Jaro Winkler
distance metrics P>0.9

Birth date

+/- 3 days

Birth place

Semantic inclusion

Nationality

80% of Opta’s
32,754 players
95,535 variations

Google Knowledge API Search

99% of the remaining 20%

• Wikidata/DBpedia do not often give first/last names)
• Google KG provides limited player features
  (with no name variations)

Weight

conversion from imperial

Height

conversion from imperial
Final KB

- Named Entity Recognition of entities:
  - 30,710 clubs
  - 264,437 people
  - 365 (multilingual) transfer terms
    - e.g. transfer, move, join, sign
    - To maintain the correct meaning translated terms must be observed for known transfers, e.g.
      - PlayerX transfer TeamY -> PlayerX transfert TeamY
      - PlayerX offer TeamY -> PlayerX offre TeamY

- players and non-player football people
  - e.g. managers, owners, pundits,
Entity Linking

• Candidate entities recognised, in parallel, using efficient Deterministic Finite Automaton (Aho-Corasick)
  • Matching occurs on normalised text:
    • removing punctuation, space variation and folding characters to ASCII
  • Players are matched on their last, first/last and pseudonyms
  • Club names and pseudonyms, which are too ambiguous for the domain filter, are also included:
    • Wolverhampton Wanders - > Wolves
    • West Bromwich Albion - > WBA
Dealing with Anomalies

- **Ambiguous player names:**
  - Full names are required for entity match
  - Players are only matched in context (i.e. when co-occurring with their club)

- **Ambiguous club names:**
  - Club are only matched in context (i.e. when co-occurring with their players)

- **Player names with low co-occurrences with their club:**
  - Examine name to identify potential issues:
    - e.g. Names which include punctuation required variations which do not include that punctuation: Georges-Kévin N’Koudou -> Georges-Kévin NKoudou
Entropy for Anomaly Detection

- **Player co-occurrence with clubs**
  - High entropy indicates that player name co-occurs equally with a number of clubs:
    - Highly ambiguous last name: e.g. Max Power, David Ball, Felix Clever
    - Common name: e.g. Smith, da Silva, Rodríguez, González
    - Super player: e.g. Messi, Ronaldo

- **Club co-occurrence with players**
  - High entropy indicates that club name co-occurs equally with a number of players:
    - Highly ambiguous name: e.g. Nice, Anger
Evidence for a Rumour

• A piece of evidence for a transfer rumour is given by:
  • a message containing a player, club and transfer term, where the player in not a current or previous member of the club

• Combining evidence considers four factors
  • Consensus
    • Evidence repeated often and by multiple independent sources
  • Authority
    • Reputation of sources providing the evidence
  • Time
    • Evidence is recent and repeated over time; in different time frames
  • Coherence/Consistency
    • Evidence is not contradictory - this is not considered for transfers as players can be rumoured to transfer to multiple clubs
• Scores for a rumour are calculated for a scoring window:
  • Consensus is measured by summing the evidence for a rumour
    • each user can provide only one piece of evidence in a scoring window
  • The evidence from a user is weighted by their authority
    • which is determined by their history of providing “true” rumours
• An exponential weighted moving average is applied to the scores from previous scoring windows
  • Which considers previous scores with a strong bias towards more recent scores
Top 5 Leagues: 2016

- 840M messages filtered in 2016
- Total rumours: 6,645
- Accepted by experts: 3,169 (P: 47.7%)
- Accepted ending in transfers: 137 (P: 4.3%)
Conclusions

• Football Whispers has >2.5M unique monthly users after 2 years from launch
  • Cost of operations is a fraction of that of traditional press methods
  • Major customers of FW are now Sky Sports and ESPN
  • Expanding to other sports
    • NFL
• System was set up and became operational in 1 1/2 months
  • We analyse about 70M messages/month
    • 56% in non-English language
• Semantic Technologies have been key:
  • Existing knowledge bases (DBpedia/Wikidata/Google KGraph)
  • Fast entity linking algorithms
  • Context modelling
Thank You!

f.ciravegna@shef.ac.uk