Extraction of Goals from Premier League Match Reports using Natural Language Processing Techniques

Student Project

presented by
Jochen Hülß & Matthias Rabus
Matriculation Number 1376749 & 1207834

submitted to the
Chair of Information Systems V
Prof. Dr. Christian Bizer
University Mannheim

Mai 2013

Contents

Chapter 1

Project Summary

1.1 Application Domain and Goals

Our interest in looking into text mining and, particulary, into sport report mining emerges from the overwhelming availability, omnipresence, and ubiquitousness of information on sports in media.
The project originates from a classical text mining problem. Generally speaking, we would like to deal with the question whether or not Natural Language Processing (NLP) tools are able to extract and correctly classify certain events in a given text. More specifically, our intention is machine-reading written reports on Premier League football matches. The particular event we are looking into is the scoring of a goal. At first sight, this appears rather straightforward, but by taking into account that there are numerous or even infinite ways to express that a team marks a goal, the difficulty to find as many of them as possible is challenging. Thus, a classification of text snippets based on manually found patterns indicating a goal is needed. Furthermore, we would like to enhance our search result by using learning algorithms.

Domains:

- Information Extraction w/ help of web crawling, NLP, POS Tagging, NER, manually designed patterns matching football goals - Text Mining: Learning a Classification model on automatically-generated labeled data seeds, bootstrapped negative seeds

part-of-speech taggers are computer programs for assigning contextually appropriate grammatical descriptors to words in texts; these approaches are: taggers based on handwritten local rules, taggers based on n-grams automatically derived from text corpora n-gram are subsets of a text with length n (**?**, p.219)

Due to the explosion of available textual data, text mining and Information Extraction (IE) from texts have become important topics of study in recent years. In particular, detection of relations between named entities is a challenging task to automatically discover new relationships in texts. (**?**, p.1)

Some previous works use hand- crafted linguistic IE rules for that task which is time consuming [7,5]. Other methods based on Machine Learning (ML) techniques [10] give good results but run as a black box (**?**, p.1)

Sequential pattern mining [1] is a data mining technique that aims at discovering correlations between events through their order of appearance. Sequential pat- tern mining is an important eld of data mining with broad applications (e.g., biology, marketing, security) (**?**, p.1)

Information extraction (IE) is the automatic identification of selected types of entities, relations, or events in free text. (**?**, p.545)

The project originates from a classical text mining problem. Generally speaking, we would like to deal with the question whether or not Natural Language Processing (NLP) tools are able to extract and correctly classify certain events in a given text. More specifically, our intention is machine-reading written reports on Premier League football matches. The particular event we are looking into is the scoring of a goal. At first sight, this appears rather straightforward, but by taking into account that there are numerous or even infinite ways to express that a team marks a goal, the difficulty to find as many of them as possible is challenging. Thus, a classification of text snippets based on manually found patterns indicating a goal is needed. Furthermore, we would like to enhance our search result by using learning algorithms.

While looking for goals in football match reports, we are omitting the detection of team that scored as well as the unique recognition of a goal as many reports refer multiple times to the same goal.

Jochen. Einleitung, Uebertragbarkeit auf andere Text Mining Tasks. Steps to goal

1.2 Structure and size of the data set

Since many NLP tools work best with English language, we would like to increase our chance of success by using football reports of the English Premier League. The BBC sports department offers match reports for every game of the currently ongoing season 2012/2013. The amount of reports is sufficient for the purpose of this project. A Premier League season consists of 38 matchdays with 10 games each. BBC provides one report per game. BBC is a goog datasource because the reports are neutrally written and have no preference for any team. Furthermore, the used language has a higher level and should be, compared to spoken word, easier

to analyse.
The first step of the project will be collecting the data from the BBC homepage.

Matze. siehe outline, punkt 1.2 Struktur input und struktur fuer web mining (satzweise getrennt)

1.3 Preprocessing

Matze: Wie wurde struktur aus vorherigem Abschnitt erreicht? Jochen: Gate, Jape-Rules, ngrams

1.4 Actual Web Mining

Matze. Verbindung zur Vorlesung. Web mining techniken. Rapidminer prozess. classifier. Suche nach perfect settings. apply model

1.5 Evaluation

Jochen. Resultate von Precision und Recall der Classification. Diskussion. Positive und negative Einflsse auf ergebnisse. Analog fuer ungelabelte Daten

1.6 Conclusion

Wrap-up.