S.O.C.C.E.R - A novel framework for measuring sports facts relevance

Ionut Mihai
6264298

Master thesis submitted under the supervision of
Prof. Dr. Paweł W. Woźniak

the co-supervision of
Prof. Dr. Kees van Deemter

the daily supervision of
Casper Koning

in order to be awarded the Degree of
Master of Science in Information Science
major in Business Informatics

Academic year
2020 – 2021

Title: S.O.C.C.E.R - A novel framework of measuring sports facts relevance
Author: Ionut Mihai
Master of Science in Information Science – major in Business Informatics
Academic year: 2020 – 2021

Abstract

Sports stakeholders capture more details about all the facet of a sport due to economical reasons. Next, they have to select the relevant information to sports fans, gamblers, journalists and other stakeholders that can use this information. Sports has been a research subject of the Natural Language Generation (NLG) field due to the data it produces and the standard natural language at times. But, NLG system have struggled to deliver a suitable level of performance compared with journalists or other sports experts due to various issues. We established that one of these issues is due to the lack of a proper definition of relevance for sports content. We set out to fix this issue by defining a relevance measurement framework for the sports domain.

We introduce the term Sport fact (SF) to refer to natural language about sports. We defined seven general types of content that are part of a SF and can have an influence on relevance. We identified twelve properties that are used to measure relevance in other domains. We established a set a measuring guidelines that describe what and how a measurement should be done for a given property and content type. All of these form the Sports Facts Relevance Measurement Framework (SOCCER) framework, the main artifact of this research project.

Keywords: sports facts, relevance, natural language processing, content, interestingness, sports statistics, natural language generation

Preface

This report was written to fulfill the requirements for graduating the master in Business Informatics at Utrecht University. The report presents a new framework that can be used by researchers and practitioners to measure the relevance of sports content, in particular, sports facts.

I started this research project at the proposal of Gracenote Sports, a sports data company, at which I was an employee at the time. Gracenote Sports employ a template based NLG system to deliver sports facts to their clients in the news and the betting markets. But, the current relevance measurement don't deliver the required performance. I further refined the research problem and identified the literature gap with the assistance of my supervisors, Dr. Pawel W. Wozniak and Dr. Kees van Deemter. As with most research projects, the process was difficult as part of the research was executed during the COVID-19 crysis. But, I conducted a rigorous research process which allowed me to answer the main research question and reach the proposed goals.

Acknowledgements

I would like to express my sincere gratitude to my supervisors Dr. Pawel W. Wozniak and Dr. Kees van Deemter for their guidance, support and patience. Dr. Pawel W. Wozniak has guided me through the research process to assure its rigour and validity. I would've got to the end of the project without the technical assistance and moral support of Dr. Pawel W. Wozniak. Dr. Kees van Deemter guided the starting direction of the research project and further contributed to the design of the artifact through his insightful comments and questions. Secondly, I want to thank my daily supervisor, Casper Koning, for his on-going support and valuable advice. Casper has been a role model for me due to his person, experience and management skills. I offer my sincere gratitude to Arno de Ruiter who first proposed and supported this research project. I offer my special bows to the sports experts who shared their valuable knowledge related to sports. Thank you to all of the collective of Gracenote Sports for their assistance, feedback and support thorough out the research project.

The research project required dedication, time, motivation and hard work. At times, I lacked in some of them. I offer my gratitude to my girlfriend who supported me and encouraged me in the times I needed. I am extremely grateful towards the support and drive my family has offered.

Although, I am the one who wrote and executed this research project, it would've gotten to this point without the assistance of the mentioned persons. Thank you all.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

IPTC	The International Press Telecommunications Council
IR	Information Retrieval
IS	Information Science
MEA	Mean Absolute Error
ML	Machine Learning
NLG	Natural Language Generation
NLP	Natural Language Processing
OWL	Web Ontology Language
SA	Sports analytics
Seq2Seq	Sequence-to-Sequence
SF	Sport fact
SML	Sports Markup Language
SOCCER	Sports Facts Relevance Measurement Framework
TAR	Technical Action Research
TS	Text Summarization

Chapter 1
Introduction

Relevance, as a concept, has an essential role in Information Science (IS) (Mizzaro, 1997), but researchers have not concluded on the subject (Borlund, 2003). Researchers from the field of IS have debated on the notion of relevance in the areas of Information Retrieval (IR) (McCrudden & Schraw, 2007), Text Summarization (TS) (McCrudden & Schraw, 2007) or NLG (Nichols, Mahmud & Drews, February 14, 2012). Relevance theory, a framework for understanding utterance interpretation, was first proposed by Dan Sperber and Deirdre Wilson in 1986 in their book "Relevance: communication and cognition" (Sperber & Wilson, 1986). In the meantime, relevance has become an interdisciplinary concept which touches the fields of philosophy, psychology, logic, and information science (Greisdorf, 2000).

A fact is a piece of information that is known or proved to be true (*fact | Definition of fact in English by Lexico Dictionaries*, z. j.). For this thesis, we will use the term SF with the definition:

Definition 1. A sports fact is a piece of information that is known or proved to be truthful by sports data.

The sports field gathers attention from people all over the world. It provides entertainment but also encourages competitiveness. The oldest and most prestigious competition in the world, the Olympic Games started in ancient times and continues to this day. Another worldwide competition is the FIFA World Cup. In 2006, an estimated 715.1 people watched the final match of the 2006 FIFA World Cup between Italy and France. With soccer having more than 4 billion fans across the world, it is no wonder that companies and governments invest resources in the field (*FIFA World Cup™* , z. j.). Miroslav Klose won the Golden Boot for being the best scorer, with five goals in the 2006 FIFA World Cup. It is a SF that Miroslav Klose won the Golden Boot and that he scored five times in the 2006 FIFA World Cup.

SFs are used by fans to learn and gather insights about a sports event, an athlete, or any other related entity to a sport. With a limited amount of time and a multitude of facts available, this turns into a selection issue of the facts that are fit for the goal of the fan. The issue is similar to when someone wants to describe a famous person. Human speakers would choose facts that are *unexpected* to describe a famous person (Kutlak, van Deemter & Mellish, 2016). It could be unexpected for a sports fan that Miroslav Klose won the Golden Boot or that he scored five goals in the 2006 Fifa World Cup, but one would find interesting only the first part. We can argue that the fact is not relevant anymore taking into account that the event took place more than 13 years ago. *Famousness*, *unexpectedness* are properties that would influence the relevance of a SF in these situations. The *athlete*, the *sports action* and the *results* are the content that is relevant to a sports fan. As illustrated in Fig. 2.2 a SF contains content from which we derive relevance properties. For example, Miroslav Klose represents the content about *the Event Participant* with the relevance properties of *importance*, *significance* and *timelines*.

A SF is composed of different types of content types such as the event, the event participant, the role, the team, the location, the action and the statistic. Thus, a SF can

be described as event participant or team with a role performing an action at a location with a given statistical value. To measure the relevance of the SF we use *relevance properties* that are measured using *guidelines* for the enumerated content types. For example, we can identify popular SF by measuring the popularity of event participants using social media data. We use the term **metric** to describe a combination of a *content type*, *relevance property* and *measuring guideline*.

SF relevance is a multi-dimensional concept that captures subjective and objective properties that have an influence on the relevance measurement. Google has a similar approach where it identifies relevant search results by taking into account properties such as *freshness*, *quality of content* or *usability of webpages*.

In the following, we will establish the problem context in Sec. 1.1, then we define the research goal and questions in Sec. 1.2. In the final section we establish our contribution to the research field. The thesis report continues with Chapter 2 which describes the research methodology. Next, in Chapter 3, we discuss the related literature fields of Sports analytics (SA) and NLG. In Chapter 4 we describe the SOCCER framework, its metrics and its evolution during the research process. In Chapter 5 we answer the sub-research questions, define our contribution to the research field, discuss the limitations and threats to the validity of the research and elaborate on possible future research venues based on this research project. Finally, we end the thesis report with Chapter 6 where we answer the main research question.

1.1 Problem statement

Media companies provide fans with sports data in different formats, such as visual representations or text. In this thesis, we will focus on the textual presentation of SFs. For traditional media companies, SFs serve as content for news articles. Fans read sports news articles for entertainment and to learn about their favorite team or athlete. Betting companies represent another commercial aspect of SFs. The SFs can provide valuable information to a gambler and thus raising the chances that he will place the bet. Media companies can attract additional readers and keep the current ones engaged with relevant SFs.

Sports stakeholders are entities that benefit from tracking and redistributing SFs. They capture increasing amounts of data that are used to generate a plethora of sports statistics and SFs (Albert, 2010). Journalists and software systems have to select the data and the facts that are most relevant for their audiences. Journalists and sports experts use their experience to filter out irrelevant data, but software systems have had challenges in applying filters on the given content (Tagawa & Shimada, 2018). Software systems can use relevance measuring metrics to improve their filters and thus their applicability.

In this thesis, we will focus on NLG software systems. NLG systems use sports data to generate natural language. Data-to-text systems are a particular application in the field of NLG that are highly applicable to sports data. Such systems take non-linguistic data and generate text. Researchers divide data-to-text systems into two main categories in the sports field: (i) *commentary* style systems that produce text resembling the style of live commentary heard at sports events and (ii) *summary* style systems that produce text comparable to the one found in newspapers or web (Lee, Krahmer & Wubben, 2017). The systems create SFs using relevant data.

Content determination is a chore of NLG, which involves deciding which information to

deliver to the user. This task can be complicated for sports data-to-text systems. Template-based systems have increased variations in the outputs they generate (van Deemter, Theune & Krahmer, 2005), as such commentary style systems need to filter those variations to deliver the valuable information at the proper time. Summary style systems have to capture the exiting parts of the sports event. Thus they need to filter out the irrelevant SFs. For example, PASS is a system that generates tailored, short soccer reports of a match for the fans of a team or the other (Lee et al., 2017). The system used a limited amount of the available data to generate the reports. With more data available, tailoring options, and syntactic trees, the system would need to do sophisticated filtering on the content it would deliver to the user. Our research can be used in the *determination* stage of NLG systems to identify relevant SFs.

Besides sports analytics data, researchers have found new sources of data for NLG systems. Social networks like Twitter and Facebook are a rich source of text data. Fans, journalists, and other stakeholders post status updates on social networks during a sports event. Researchers have been using post updates from social networks to generate summaries of the given events. In this case, the content determination task involves selecting the status updates that provide an accurate and qualitative description of the current situation of the event. Such an undertaking is not straightforward when it involves more than 110000 tweets that were posted for the duration of the match between the US and Slovenia in the 2010 World Cup (Nichols et al., February 14, 2012). The tweets were used by Nichols et al. to detect the essential events of the match. Then the tweets were filtered and aggregated into the match summary. The selection was made by scoring tweets using the phrase method of Sharifi, Hutton en Kalita. The scoring technique provided flourishing results when the majority of the tweets described the event but led to mediocre results when a substantial part of them represented screams or applauses. Kubo, Sasano, Takamura en Okumura approached this issue by first identifying users on the social network whom they define as *good reporters*. Using this line of thought, Kubo et al. introduce in their paper *tweet score* and *user score*. Their system then uses the two scores to select the tweets that are describing a particular event in a match. Such systems, use metrics based on word count to identify the suitable output but do not leverage the relevance of the content. Such NLG systems can can extract the content types of the SFs. Then, it can measure the content relevance using relevance metrics. Thus, leading to increased quality of the output.

Kutlak et al. used heuristics based on the properties of the referent to select the facts that best describe famous persons to an unknown audience. Kutlak et al. relied on three heuristics to determine which fact to use: (i) the *Knowledge Heuristic* calculates the possibility of the property to be known to the user, (ii) the *Unexpectedness Heuristic* identifies rare properties in the population as a whole, (iii) the *Termination Heuristic* controlled how much information the description consists of. Some sports statistics are complicated and hard to understand by sports fans and gamblers. NLG systems can use a knowledge heuristic to identify suitable SFs for different sports fans and gamblers. In news articles, sports journalists try to present the *sensational*. As such, an NLG system can leverage an *unexpectedness heuristic* to identify the content that is out of the ordinary. Additionally, a news article has certain space limits. NLG systems need to use a *termination heuristic* to determine the limits and the relevant SFs that should be part of the final output.

Most of the identified research in the field of sports NLG focused on the produced text and the architecture of the NLG system and overlooked the relevance estimation

function. Researchers and practitioners could develop new relevance measuring heuristics by leveraging sports domain knowledge and meta-data. As such, we identified and defined a set of *metrics* that can be used to measure the relevance of SFs for software systems and processes that involve SFs.

1.2 Research Goal and Questions

The topic of this thesis is the measuremnt of the SFs relevance in the news and betting domains. By focusing on the given domains, we will capture metrics that picture an accurate measurement of relevancy. Based on the domain, a synonym for relevance would be *"newsworthiness"*, *"interestingness"*, *"betting worthiness"*, or others. A fact can take different forms with the same content. The form of a fact is closer to linguistics research. As such, in this thesis, we will focus on the relevancy of the content. By the content of a SF, we mean the explicit and implicit entities involved or described by a sports fact such as the athlete, the team, the match, or the competition, the statistics which we defined in Sect. 4.2.

One of the objectives of this thesis is to improve NLG systems that use sports data to generate content for the news and betting domains. To evaluate the applicability of the research, we executed a study at the sports data company Gracenote Sports. Omega is an NLG system that generates soccer facts that was developed by Gracenote Sports company. For the 2018 UEFA Champions League final between Liverpool and Tottenham, the project generated more than 100 facts. Some examples of those facts are:

SF1 Tottenham Hotspur have scored the opening goal in 7 of their last 10 home matches.

SF2 The last time these teams met in this exact same fixture in Europe was on 24 April 1973.

Out of more than 100 facts, the company needs to select and deliver to their clients the most relevant ones. The fact **SF1** seems more suitable for the betting domain while fact **SF2** is more suitable for the news domain. The content of fact **SF1** is formed by the team, the competition and the match as the event, the template about scoring the opening goal and the result being 7 out of 10 matches. **SF1** contains an impressive result and is about a noteworthy action of the game, scoring the opening goal, which makes the fact exciting. Besides, one can bet that Tottenham Hotspur will score the opening goal again. Fact **SF1** has visible content such as the team but also tacit content such as the competition. The hidden content can affect the relevance of the fact when compared with facts about other competitions. Fact **SF2** can be ranked lower in terms of relevance as it does not contain an impressive result.

The artifact, problem context, and design problem are made explicit using the template of Wieringa (Wieringa, 2014):

> This research aims to improve relevance measurement of sports facts by devising and implementing a relevance measurement framework that is usable by researchers and practitioners in order to develop innovative relevance measurement heuristics for sports content in NLG software systems.

As part of this research, the framework will be evaluated and implemented in a case study at Gracenote Sports company:

Create a software system that use relevance measurement heuristics to identify suitable soccer facts for the news and the betting domain.

We envision a collection of content properties and measuring guidelines that affect the relevancy of a SFs, grouped in a framework that would facilitate as inspiration for researchers and practitioners to develop and create new heuristics that determine the relevance of sports . The collection forms the SOCCER framework. SOCCER will enable practitioners and researchers to improve software systems that involve sports content. Moreover, journalists could improve their workflow by using SOCCER for a smooth SF selection process. Ultimately, SOCCER could concentrate research efforts into comparable and evolvable heuristics for measuring the relevancy of SFs. Therefore, to address the objectives and evaluate our vision, we address the main research question of this thesis:

Main Research Question

How can a content relevance measurement framework for SFs be developed to aid practitioners and researchers in improving and constructing relevance measuring heuristics for NLG systems?

We follow several steps to answer the main research question. First, we need to develop the SOCCER framework by identifying and defining the content, the properties that have an influence on the relevance of SFs and measuring guidelines that quantify the impact of the content properties towards relevance. Then, we use the SOCCER framework to build heuristics and algorithms that can be used by NLG systems to determine the relevancy of SFs in a case study at Gracenote Sports company to determine whether the framework can be used for the purpose it was created. Finally, we evaluate the performance of the heuristics developed using the SOCCER framework to understand its utility. Each step provides specific information to answer the main research question. We defined several sub-questions that capture the steps mentioned.

SQ1: Which content properties affect the relevance of a sports fact and how can they be quantified?

SQ1.1 Which content properties affect the relevance of a sports fact?

SQ1.2 How should we quantify and measure the contribution of a content property to the relevance of a sports fact?

SQ1.3 How can content properties and measurement guidelines influence the relevance of a sports fact for the news and betting markets?

The first step of the thesis is to identify which content properties we can leverage for relevance measurement and quantify their contribution to the relevance of a SF. We do so by focusing on the fields of sports analytics and computational linguistics. We identified the statistics that are used in sports, especially their content and their data. We extracted the properties, algorithms and heuristics from NLG and data mining fields to measure the relevance of content. We used this information to define the content and the properties that have an influence on the relevance of a SF to answer **SQ1.1**. We defined guidelines that quantify the contribution of a content property to the relevance of the SF to answer **SQ1.2**. We executed a set of interviews to validate and extend the the set of content

properties that influence the relevance of a SF. We used data from the sports experts interviews and from the relevance measurement heuristics to answer **SQ1.3**.

SQ2: How can we use the SOCCER framework to build tailored heuristics for relevance measurement of SFs in NLG systems?

SQ2.1 How can we build relevance measuring heuristics that can be used in NLG systems?

SQ2.2 How could the relevance measuring heuristics support tailoring options for the news and the betting markets?

The goal of the SOCCER framework is to facilitate the development of relevance measurements heuristics for sports content. As such, we use the answer to SQ2 to elaborate on this part. Thus, we used the SOCCER framework to create new relevance measurement heuristics for soccer as part of a case study project at Gracenote Sports company. We implemented the heuristics in a software system that supports tailoring options for the news and betting markets. We use the data gathered during the development of the software component and the heuristics to answer SQ2.

SQ3: How do experts perceive the quality of the results delivered by the relevance measurement heuristics?

SQ3.1 How do experts perceive the usefulness of the results delivered by the relevance measurement heuristics?

SQ3.2 What is the accuracy of the results delivered by the relevance measurement heuristics compared with experts?

Finally, we evaluate the relevance measurement heuristics, and implicitly, the SOCCER framework ability to guide the development of such heuristics by answering SQ3. The software component recommends relevance categories for soccer facts. Therefore, we asked sports experts to determine the usefulness of the relevance measurements as the beneficiaries of the relevance measurements performed by the software component. As such, we can answer **SQ3.1**. Additionally, we evaluate if the system delivers comparable results to those of experts to answer **SQ3.2**. We elaborate on the validation methods in Sec. 2.4.

Finally, we aggregate the information from each sub-question to answer the main research question. We can state the design of relevance framework using the information from SQ1. The information from SQ2 provides us with an answer on how the relevance measuring heuristics could be implemented. Finally, SQ3 evaluates the ability of the framework to guide the development of heuristics that deliver the expected results by experts.

1.3 Research contribution

The thesis contribute to the scientific community by capturing domain knowledge of the sports field from experts. The thesis will turn the tacit knowledge of the sports experts about SF relevance in explicit knowledge available to both practitioners and researchers. SOCCER will synthesize that knowledge in a usable format by both communities.

In the NLG field, SOCCER will assist researchers and practitioners to improve the heuristics employed in the **content determination** stage to identify the relevant content

for a sports-related subject. In particular, this thesis will contribute to research that aims to summarize sports events (Nichols et al., February 14, 2012), automated journalism (ge Yao Jianmin Zhang Xiaojun Wan Jianguo Xiao Institute of Computer Science et al., 2017; Kubo et al., November 17, 2013) or commercial products such as the Omega project.

Chapter 2
Research Methods

In this chapter we discuss the research methods that we used in the thesis. First, we elaborate in Sec. 2.1 on the primary method which guides the thesis project. We used support methods to conduct the investigation, design the artifact and validate the results. We executed a literature review to expand the knowledge body on related research fields and to extract the necessary qualitative data for the artifact design. The process and the method are described in Sec. 2.2.1. We gathered knowledge from experts through interviews and we analysed the data using conceptual analysis to design the artifact. In Sec. 2.3.1 we describe the interview process and guidelines. Sec. 2.3.2 presents the guidelines we applied to analyse the data. Finally, we validated the research project using a case study. In Sec. 2.4 we describe the case study methodology and the evaluation metrics and guidelines.

2.1 Thesis research methodology

We structure this research project around the design science methodology. **Design Science** is an iterative problem-solving research method used in Information science research. It was first proposed by March en Smith to "insure that IT research is both relevant and effective". Wieringa describes design science as being the "the design and investigation of artifacts in context" and argues that design science problems are "improvement problems". In design science, the researcher investigates the interaction between the artifact (e.g., a method, model, prototype) and the problem context (e.g. social context, knowledge context) as it is this interaction that leads to the solving of the problem. The treatment is the interaction between the artifact and the problem context. Design science iterates over two primary activities: (i) designing an artifact that improves the problem context and (ii) answering knowledge questions about the artifact in context.

This research project is different from the normal design science projects. Design science is suitable for research projects that investigate processes and software systems that are used in practice but whom can be improved. In this case, we build the SOCCER framework from scratch and we are the ones who implemented it in a software project. Although our goal is different, we apply the design science methodology to organize the research project and guide its structure. We opted to use the design science methodology due to our experience with the method. Therefore, this increases the research rigour and the research value of the thesis. We investigated also the CRISP-DM methodology but we found that it is suitable for software systems that involve data mining or even knowledge discovery systems and not for NLG related research. In this research project, the primary artifact is the SOCCER framework. During the validation of the framework, we create secondary artifacts such as the measurement heuristics and the software component that calculates the relevance of soccer facts.

On a primary level, the thesis uses the Design Cycle activities to execute the research. We start the thesis with the *Problem investigation* phase in which we investigated the research fields related to the thesis using a literature review thus answering **SQ1.1** and

SQ1.2. In the *Treatment Design* phase we defined the SOCCER framework with its content properties and the measuring guidelines thus answering **SQ1.3**. Next, we evaluate the framework using a case study at Gracenote Sports in the *Treatment validation* phase thus answering SQ2 and SQ3. The process is iterative until we attain a respectable result.

2.2 Problem investigation

The purpose of the Problem investigation phase is to identify, describe, and evaluate the research problem in order to establish the necessity of the research and define the problem context. We determined the stakeholders of the research and their goals. Besides, we defined the theoretical framework by identifying the scientific theories that compose the knowledge context. The knowledge context is the starting point of a design science research project.

2.2.1 Literature review

We identify two topics that are primordial to SFs and this research: the sports analytics field and sports in computational linguistics. The sports analytics field deals with the historical sports data and the statistics that extract knowledge from it. A SF contains such knowledge in different formats. Secondly, we need to investigate how researchers treat sports in the field of computational linguistics in terms of relevance measurements, e.g. interestingness, newsworthiness. We performed a literature review that covers these two topics. At the same time, we identified methods, algorithms, and metrics used for sports facts relevance measurement to answer SQ1. We considered that a method or algorithm that is filtering, selecting or ordering sports facts is doing relevance measurement for their use case.

We executed a systematic literature review to identify the gaps in literature but also to extract relevant data to the research project. The systematic literature review follows the guidelines defined by Kitchenham en Charters. A systematic literature review identifies and synthesizes the existing body of knowledge in a fair manner that provides scientific value to the findings. Besides, the work is reproducible as we documented the literature review process in detail. Wohlin argues that a systematic literature review is as good as its search terms. That is the case here as *relevance* is an interdisciplinary concept that received a great deal of attention, which leads to noise in the query results. As such, we gave extensive attention to the search queries.

During the execution of the search strategy, we selected only the first ten results from each query. The approach allowed us to conduct a timely literature review. The risk is that we have omitted certain results that appeared on later positions. As such, we splited the literature review in two parts, sports analytics and computational linguistics. For each, we elaborated several search queries to reduce the risk of missed out papers. In sports analytics, we focused on the sports of cricket, soccer, football, basketball and baseball. In computational linguistics, we focused on *interestingness measures*, *Natural Language Processing (NLP)* and *NLG*. As such, we cover a broad range of papers.

The literature review started by identifying the research papers related to the two topics using the search strategy from Appendix A.1.2 and A.2.2 We noticed some duplicates early in the search results, so we removed them on the spot. The execution of the search strategy resulted in 59 papers related to sports analytics and 74 papers concerning computational linguistics. We trickled the results through the inclusion and the exclusion criteria

which resulted in 40 papers about sports analytics and 44 about sports in computational linguistics.

We used NVivo to identify important and relevant passages of text to this research project. At the same time, we used Google Docs and Google Sheets to extract data according to the forms in Appendix. A.1.5 and Appendix. A.2.7.

Data Extraction

Sports data sources. Sports data is needed as a factual source for NLG systems. Researchers need open-source data, or ways acquire it so that they can use it in their research. Secondly, commercial applications also need to acquire their sports data be it from commercial sources or open data sources. As such, one section in the data extraction form is dedicated to uncovering data sources. Several research projects involved building their own prototype which also acts as a data source (Halvorsen et al., 2013; Gowda et al., 2017; Stensland et al., 2014; Hamdad, Benatchba, Belkham & Cherairi, 2018). Opta Sports released a dataset with statistics for the 2011-2012 English Premier League season which was used by several research projects (Kumar, 2013; Sukumar, 2019; Hoeve, 2017; Asif, Zaheer, Haque & Hasan, 2016). Additionally, we focused on the time ranged of the available data. For soccer, the oldest data was from 1992 (Gangal, Talnikar, Dalvi, Zope & Kulkarni, 2015) while the newest was from the season of 2017-2018 (Sukumar, 2019). While soccer analytics data is scarcely available and spread across different sources, basketball, baseball and cricket have specialized sports data websites that proved to be of assistance to several research papers. The website `basketball-reference.com` was used as a data source for basketball related research(Arnold & Godbey, 2011; Yang, 2015) while for baseball it was `www.baseball-reference.com`. For cricket, an important data source for research papers is the website `http://espncricinfo.com`. Overall, researchers have several sources of data available for historical data but, for live data, researchers and companies will need a commercial data source.

Relevance metrics, algorithms and requirements. Another goal of the literature review was to identify the methods researchers use to measure the interestingness of the content. Additionally, we collected the algorithms that are used in NLG projects, heuristics that perform relevance measurements, and requirements for NLG projects. As such, from the literature review we extracted 22 algorithms that were employed in NLG related tasks. We can use the algorithms to identify requierments and limits for relevance measurement from the NLG systems that implement them. Also, we gathered requirements and challenges that were encountered by the researchers. Thus, we found that an NLG system that uses neural networks can begin to „hallucinate facts" (Kanerva, Rönnqvist, Kekki, Salakoski & Ginter, 2019). Also, NLG systems needs to follow the same journalistic requirements as journalists (Leppänen, Munezero, Granroth-Wilding & Toivonen, 2017). From the literature review we identified which content properties the researchers use to measure the relevance or interestingness of they data. The search strategy uncovered papers from the NLG field but also from data mining. In data mining, researchers use statistics to measure the interestingness of the content but also subjective measures. As such, we gathered several relevance properties and measuring guidelines from this field. In total, we identified 44 content properties that we can use as inspiration to design SOCCER, some of which were duplicated. Some of the metrics are straightforward while for some the authors didn't mentioned the computational method. Thus, we extracted 26 relevance

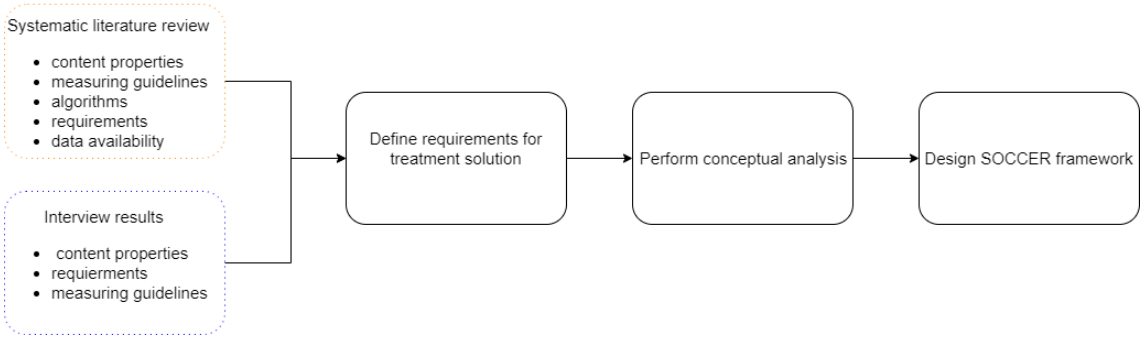

Figure 2.1 – The processed followed for data analysis and the creation of the artifact

measuring guidelines for relevance properties. This data can acted as a basis for the design of the SOCCER framework.

2.3 Treatment design

In the treatment design phase, the researcher takes the necessary steps to elaborate a solution treatment. The primary artifact of this research is the SOCCER framework, which aims to provide guidelines for measuring the relevance of sports facts. Fig. 2.1 illustrates the process that we followed to create the artifact.

We started by designing an initial version of the framework that takes into account the underlying limitations of sports analytics. We further develop the framework by leveraging the techniques, properties, and requirements uncovered during the literature review on the topic of computational linguistics. The framework is refined further using conceptual analysis on the results from the literature review and the interviews with sport experts.

2.3.1 Expert interviews

Besides the research community, practitioners carry knowledge as stakeholders working on NLP systems for editorial activity that involves SFs. The knowledge of practitioners is captured through several interviews. We used semi-structured interviews as this is an exploratory problem, and we can adapt to the distinctive specialties of the interviewees.

An interview can be organized as *drama*, with its own props, stage, actors and script (Myers & Newman, 2007). We used the guidelines proposed by Myers en Newman to mitigate the threats to the validity of the research when using experts interviews. In addition, we used a a mix between the appreciative and soft-laddering interview format, as described by Schultze en Avital. As such, the interviewee was able to formulate judgements related to his favourite sport besides soccer.

We interviewed eight sports experts with different areas of expertise that fit the target beneficiaries of this research project. Thus, we first focused on employees that were involved in the Omega project or are using the project actively for identifying relevant SFs. As such, we interviewed the software engineer that developed the Omega project, the product owner that manages the project and 3 content editors that actively use the the software to identify relevant SFs for journalists or the betting market. Additionally, we interviewed two other employees from the company which have extensive knowledge about the sports industry. Appendix B.1 contains the interviewees pseudonyms, their job role and the years of experience they have in the industry. Besides the practitioners we also interviewed

a researcher involved in the development of sports NLG systems for reporting amateur leagues.

The goal of the interview was to identify which content types contain a relevance property and affect the relevance of the SF. Thus, the main question for the expert was „Can the *relevance property* be measured for *content type* and influence the relevance of a SF?". Due to the coronavirus outbreak, the interview processed was executed using online communication tools. The interviewee received instructions on how to set up the online environment such as the communication tool and the documents. Besides the instructions, the interviewee received a document with examples and the definition of the content types to facilitate the critical thinking of the expert. To gather the experts judgements, we used a third document which contained each property of the SOCCER framework in the format of a survey as illustrated Sec. B.3. We recorded the screen of the interviewee so that we identify the options he selected. Each relevance properly implicates two steps in the interview process. First, the expert would see the description of a relevance property. He is asked to select and identify the content types for which an NLG would have to measure the given property as it has an influence on the relevance of SFs. In the second part, the expert was confronted with our selections. The differences were discussed until we reached a conclusion. The expert could agree or disagree with us as illustrated in the results of the interview in B.4. We measured the inter-rater reliability of the answers using Krippendorff's Gracenote Sports coefficient to identify the level of subjectiveness involved in the answers. Each property was evaluated by four of the eight sports experts. Thus, the data collected from the survey is inconsistent across properties. We can measure the inter-rater reliability for such cases using Krippendorff's Gracenote Sports coefficient (Gwet, 2011). The coefficient value for the first part of the debate is around 0.35 and 0.72 after the second part. A value of 1 means that the interviewees were in total agreement while 0 signifies total disagreement. The coefficient value allows us to draw tentative conclusions as it is higher than the suggested value of 0.667 by Krippendorff but lower than 0.8 (Hayes & Krippendorff, 2007).

We analysed the interview using several software tools. We transcribed the interviews and analysed them using the NVIVO software and Google Speech to Text service. For each interview, we viewed the video recording and took notes of the interviewee choices in an Excel sheet. At the same time, we took notes on the arguments of the interviewee and how they influenced his choices. In NVIVO, we created a data coding tree based on the content types and relevance properties of the SOCCER framework. The qualitative data generated insights into the experts selections. We evaluated six relevance properties per expert. We assigned the six properties based on the experience of the expert. The quantitative data is outputted in Sect. B.4. The process followed the guidelines of Blandford, Furniss en Makri thus, we used a „grounded" qualitative data analysis in which we first started extracting the interviewee choices and took notes and then moved to coding the data in NVIVO. Thus, we got an in depth knowledge on the content of the interviews. To analyse the quantitative data, we use a Jupyter notebook and several data processing libraries such as Pandas, Numpy and others. The notebook allowed for quick iterations and flexibility in transforming it in suitable structures.

2.3.2 Conceptual analysis

Conceptual analysis is a method for investigating and interpreting qualitative data. In our case, there are two qualitative data sources: the literature review and the interviews. In

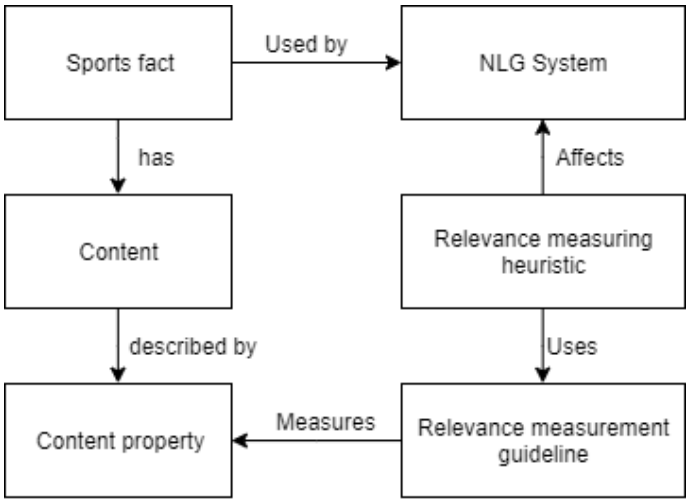

Figure 2.2 – Conceptual model

the conceptual analysis, researchers analyze allegorical data to identify instances of the concepts that are part of the conceptual model of the research at hand (Wieringa, 2014). A SF has *content* that is described by *content properties*, which can be *measured*. An *NLP system* uses SFs, which are selected or filtered using *content relevance algorithms* that depended on *measurements* of *content properties*.

The qualitative data sources are analyzed using NVivo software. A coding tree is created based on the conceptual model described in Fig. 2.2. The goal is to identify the type of content that forms a SF, the properties of the content, and measurements that can quantify the presence of these properties and their importance for relevance measurement.

The measurement guidelines, content types and the content properties are an interpretation of the researcher based on the analysis of the data. The validity of the interpretation is supported by applying methodological triangulation of the data sources: the literature review and the interviews. We constructed a preliminary framework using data from the literature review. We refined the framework design using the data elicited from the expert interviews.

Using the data from the literature review we constructed the shape of the SOCCER framework by identifying the content types that form a SF, properties that influence the relevance of data and measuring guidelines to reach the relevance target. We filled in the blank spots of the SOCCER framework with the help of sports experts which assisted us in identifying and validating the relevance metrics.

2.4 Treatment validation

The Treatment Validation entails justifying if the designed treatment solves the stakeholders' goal when applied in a real context (Wieringa, 2014). We apply the SOCCER framework in a case study at Gracenote Sports, a sports data company. We used the SOCCER framework to develop a software component that determines the relevance of the soccer facts for the betting and the news domains. We applied the process described in Fig. 2.3 which follows the guidelines of the Technical Action Research (TAR) methodology.

We started with the investigation of the problem context for the case study. We established the goals and issues of the Omega project in terms of relevance measurement. Then, we defined the requirements for the content determination stage. We decided on the

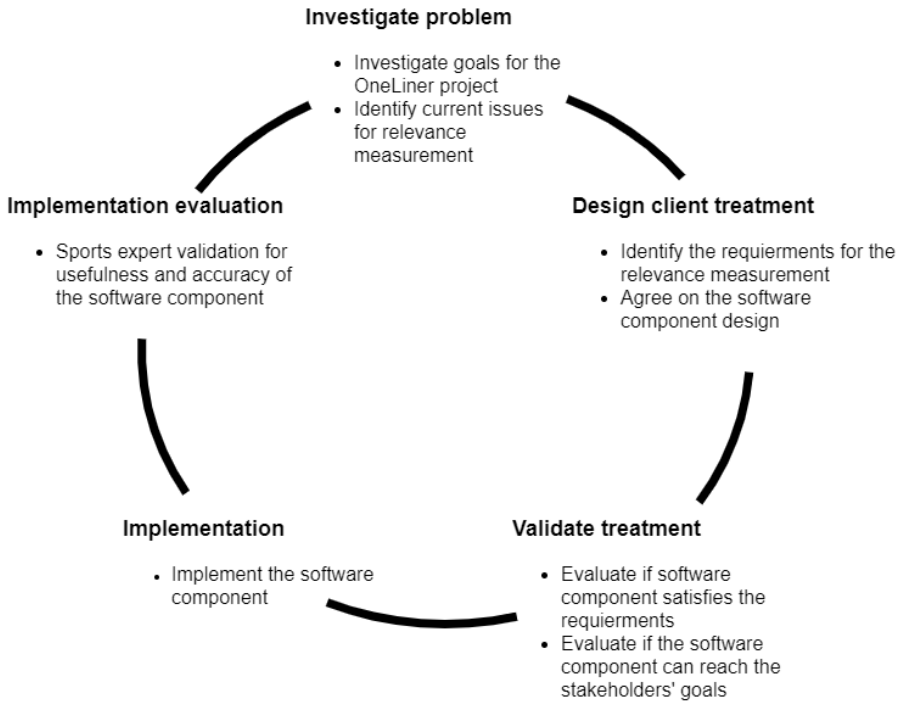

Figure 2.3 – Client engineering cycle

software component architecture such as the programming language, data sources, output methods, tailoring options. Then, we evaluated if the software component satisfies the requirements of the stakeholders. After this step, we implement the software component and then we validated it.

Omega is a template based NLG system which produces SF for the news and the betting domains. Omega produces a wide range of SFs for a soccer match in different languages and several soccer competitions. We focused on the English language for the case study as it is a widely used language in NLG field. We used SFs from the English Premier League due to the high coverage in terms of data by the company. Currently, the system would compute the relevance of a SF based on the statistical significance of the result part of the SF and the template involved. But, the current relevance measurements don't deliver the necessary versatility and it requires the assistance of a sports expert to select the most relevant SFs. As such, Gracenote Sports company wants to improve the current relevance measurements to ease and automate the task of the sports experts. Currently, the sports experts who filter out the facts have the experience and the tacit knowledge to identify the relevant facts. As such we based the case study around ML techniques which can capture and automate human tasks.

A critical part of a ML model is the data you use to train it. In our case, we need to grasp the subjectiveness of relevance for the news and the betting domains. CE1 and CE2 are employees of Gracenote Sports company which use the Omega project to select relevant facts for the news and the betting domains. As such, we collaborated in the case study to gather a golden data set to train the ML algorithms. CE1 and CE2 participated in the interview process where they got acquainted to the SOCCER framework. Thus, we found them as a reliable source for creating a training data set. Labelling data can be a dull and demanding task as the sports experts also had their daily job responsibilities. Therefore, we implemented several safeguards to reduce the workload of the sports experts, and to

Likert Scale				
Not at all	Slightly	Moderately	Very	Extremely
Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree

Table 2.1 – Likert scale variations

ensure the quality of the data. Firstly, we used a 5 point asymetric Likert scale to evaluate the relevance of a SF. A SF can be relevant to a certain degree or not relevant at all for the news or the betting domains therefore the asymmetry of the Likert scale. The choices for the Likert scale questions are illustrated in the first row of Table 2.1. Relevance is a subjective measure and it has no definitive measurement. Thus, the Likert scale allows the sports experts to pick a relevance category for a SF that is closer to their opinion (Joshi, Kale, Chandel & Pal, 2015). Additionally, we provided guidance and instructions for the Likert scale categories. Secondly, we reduced the workload for the sports experts by limiting the number of SF they had to label per day. Thirdly, we facilitated the contextual information for the sports experts by selecting facts from successive events of three teams. Therefore, CE1 and CE2 could build a timeline for the teams and evaluate the relevance of the SFs based on the progress the team had in the competition.

It was important for the case study to use a software language and ML library which can use the SOCCER framework metrics to predict the relevance of a SF. Therefore, we used the Kaggle platform to guide our choices. Kaggle is a platform where researchers, practitioners and enthusiats participate in work related to ML. We opted to use the Python language in the case study as it is supported by most libraries for ML. In the prototype, we had to predict a relevance category for a SF. The task in the case study falls in the category of multinomial classification as we have five relevance categories. The users on the Kaggle platform perform a similar task for the Titanic dataset (Cicoria, Sherlock, Muniswamaiah & Clarke, 2014). Therefore, we used a similar suite of ML algorithms in the case study. We implemented the ML algorithms in Jupyter notebooks which allowed us to iterate and prototype fast. We used the *scikit-learn* library which provides a suite of ML algorithms. Scikit-learn is a straightforward Python based library that exposes a large range of ML algorithms (Pedregosa et al., 2011). The library is widely used on the Kaggle platform and is well documented. We will not elaborate on the scopes and details of the algorithms as they are the means of the case study and not the goal of it.

It is not always straightforward to evaluate an NLG system. Researchers can evaluate the validity of the generated facts, the size of the generated content, the appropriateness of the output to the context (Dale & Mellish, 1998). But, we implemented several ML algorithms that rate the relevance of a SF for the news and the betting domains. Therefore, we first evaluated the reliability of the relevance ratings using metrics from the ML field. We measured the mean accuracy, precision, recall and F score for the predictions (Hossin & Sulaiman, 2015). The mean accuracy metric is easy to use and simple but it won't always identify the optimal model. Therefore, we used precision, recall and the F score metrics to identify the model which could predict the relevance of SFs as accurate and complete as possible. Hossin en Sulaiman presents additional accuracy metrics for classification tasks such as mean square error, optimised precision and others. We did experimented with other metrics implemented in the scikit-learn library but they didn't improve the models. The goal of the case study was to prove and identify how the SOCCER framework can be used in practice and not to identify the best ML model that predicts the relevance of SFs.

Category	Question
Quality	The ratings for the news|betting market match my interests.
	The ratings identify *most*|*least* relevant sports facts accurately.
	The ratings identify sports facts that are useful for gamblers|journalists.
	The relevance ratings of sports facts take into consideration the requirements of the gamblers|journalists.
Ease of use	I became familiar with the relevance ratings very quickly.
	I easily found the relevant sports facts.
	I would quickly become productive with the relevance ratings.
	Finding relevant soccer facts, even with the help of the ratings, consumes too much time.
Usefulness	The sports facts ratings could effectively help me find the ideal facts.
	The ratings would influence my selection of sports facts.
	I feel supported in selecting the sports facts that I need with the help of the relevance ratings.
	Overall, I am satisfied with the relevance ratings of the sports facts.
	The relevance ratings made me more confident about my selection/decision.
	The sports facts relevance ratings can be trusted.
Intentions	If a system that rates sports facts on relevance as this exists, I will use it to find relevant sports facts.
	I would use these relevance ratings frequently.
	I prefer to use this type of relevance ratings in the future.

Table 2.2 – Likert survey based on ResQue framework

Thus, we opted for accuracy metrics that are well known by researchers and practitioners.

The ML models learned to predict the relevance of SFs based on the input from the SOCCER metrics and the relevance ratings given by the sports experts. Next, the ML models could be used in practice to assist the sports experts for rating the relevance of large amounts of SFs or in a NLG system. For the first case, we can state that the ML models recommend a relevance rating for a SF to the sport expert. The task is similar to those of recommender systems. As such, we used the ResQue framework to measured the usefulness of the predicted ratings by the ML models (Pu, Chen & Hu, 2011). The ResQue framework is a user centric evaluation model that can be used to measure the perceived usefulness and quality of recommender systems. We picked relevant questions from the ResQue framework which are illustrated in Table 2.2 to create a Likert based survey. CE1 and CE2 are sports experts which can evaluate the relevance of a SF using their knowledge and gut feeling. Also, we produced the training data for the ML models with their help. Additionally, the software component should facilitate the identification of relevant SF for CE1 and CE2 to ease their daily job. Therefore, they were also the ones who evaluated the usefulness of the software component.

The validation process is two-fold. First, we evaluate the accuracy of the relevance measurements. Secondly, we use the ResQue framework to determine the usefulness of of the relevance ratings for the sports experts. The second method, measures the capability of the system on a subjective level which can be influenced by the bias of the sports experts. In the first method, we will use data to determine the accuracy of ML models. Accuracy metrics can measure the performance of the ML models but it has its disadvantages. Overall, the metrics can measure the prediction performance but not its usefulness. Therefore, we asked the sports experts to state their perceived usefulness of the relevance ratings. Secondly, precision metrics perform well on large and balanced datasets. But, we used a limited and imbalanced dataset. Another risk is that we can overfit the data and the application won't be able to generalize well. We limited this risk by evaluating the ML models on data outside the range of the training set.

We evaluated the usefulness of the software component using the ResQue framework and answered **SQ3.1**. Then, we used the accuracy metrics to answer **SQ3.2**. The implementation of the case study produced valuable insights on the SOCCER framework which we will discuss in Sect. 4.4. The treatment validation process validates the applicability

of SOCCER in practice but also the limits and areas of improvement for the framework. We have discussed a few of the risks and issues of the research method and the way we mitigated them in this chapter. We will debate them further in Sect. 5.3.

Chapter 3
Related literature

This chapter will explore the topics of *sports analytics* and *natural language generation*. Section 3.1 explores the domain of sports analytics for the sports of soccer, baseball, basketball, and cricket. We end the section about the sports analytics with the conclusions we drew in relation to this research project in Sec. 3.1.5. The second part of the chapter elaborates on the field of NLG in Sec. 3.2. Most NLG systems that generate natural language related to sports are considered data-to-text systems and they are templated based. We describe the architecture, functionality of such systems in Sec. 3.2.1. Also, there are NLG end-to-end systems which employ novel algorithms from the fields of ML to generate text. Sec. 3.2.2 elaborates on the usage of end-to-end systems sports related tasks.

3.1 Sports analytics

SA research has a history of more than 50 years and recently took a dramatic rise in the level of scholarly interest (Coleman, 2012). Coleman examined 140 journals and identified 1147 articles that tackle the application of analytics in sports. The *Journal of Sports Economics* and *Journal of Quantitative Analysis in Sports* published the most articles. The institutions that contributed to most articles were British: Lancaster University and the University of Salford. Research in SA is not without its challenges. Passfield en Hopker argues that practical data analysis in SA requires the fusion of heterogeneous expert knowledge, such as in sports psychology, data handling and analysis, statistics, training theory, mathematical modeling, determinants of performance. Gathering such experts represents a legitimate challenge as not many individuals are sufficiently proficient in such disparate areas.

SA provides an objective view on the performance of athletes and teams and can be used to gain a competitive advantage. SA has two components: on-field and off-field analytics. The on-field analytics deal with the measurement of the performance of the athletes and teams. The off-field analytics deal with the business side of sports. In this section, the focus will be on the on-field analytics as it relates to the topic of this research and the interests of fans and journalists.

In most sports, we can divide the statistics involved in traditional or simple statistics and modern or complex statistics. All fans can understand traditional statistics as they are simple and straightforward. Due to their simplicity, such statistics are a poor indicator of performance as they can not capture sports strategies. Pioneers of sports analytics noticed this and asked themselves if these statistics measure what people think they measure. Pioneers as Bill James and Dean Oliver began to offer new statistics and insights as they asked themselves this question (Schumaker, Solieman & Chen, 2010). Their work led to the development of modern statistics that are focused on measuring specific performances of athletes and teams. Traditional statistics are often discussed and presented on television to provide necessary information for the viewers, such as who is winning or loosing, who is over-performing in the sports event. Modern statistics require a broader knowledge of the

sport than a regular viewer has. As such, they are suited for fans who are involved with the sport or try to gather knowledge for other goals, e.g., betting.

Traditional statistics are easy to present to the average viewer. As researchers create new statistics, they also develop new methods of visualizing their insights. The new visualizations methods need to deliver complex information in a simple manner that is accessible to the average viewer. In basketball, the heatmap combines different statistics with the location on the field. Drazan, Loya, Horne en Eglash used a heatmap to illustrated pass accuracy in an experiment that involved under-represented youths in STEM (Science, Technology, Engineering and Mathematics). The youths and the coaches then used the heatmaps to identify areas in which they can improve. Complex situations and statistics require an innovative approach. Stein et al. illustrated collective team movement by abstracting its features, such as the covered distance of the ball, number of passes, distance of movement, number of overcome players. The viewer could see in a top-down view the soccer field with movement trajectories of the players and the ball, and color based visuals of the abstracted features. The researchers validated the visualization tool using sports experts who rated the tool as „very helpful for detecting, exploring, analyzing, and comparing interesting game situations" (Stein et al., 2015). Hoeve also tackles visualizations in soccer. In his thesis, he combines a top down view of the soccer field with statistics about the player position and movement-related statistics such as acceleration, velocity, and more. Hoeve created a conceptual design of a pie chart that combines passes per 90 minutes and accuracy for a player but also a radar chart that merges several statistics. He combines shots per match and shots taken per match into a scatter plot where he denotes teams with a quiet or busy attack or defense. Fans and journalists can use such visualizations to identify facts such as the team with the best attack or the player with the best pass accuracy.

A picture can convey a thousand words, but text summaries of a sports event have to describe the full event in a thousand words. With the limited space available, NLP systems need to extract crucial information from the sea of statistics.

3.1.1 Soccer analytics

Soccer is one of the sports that is increasingly hard to be gauged by sports analytics. A player has the possession of the ball only three minutes on average out of 90 minutes in a soccer game (Bornn, Cervone & Fernandez, 2018). Low ball possession limits the amount of data that can be collected and have steered researchers in a different direction, that of player position and tactics. Bornn et al. compared several pitch control models that are used by analysts to identify how a player position can affect the game. The Voronoi tesselation model partitions the field into „Voronoi dominant regions". These regions represent a player sphere of influence in which a player can arrive faster in that area than others. There are several variations of the model, such as the classical Voronoi model, the weighted Voronoi model. The weighted Voronoi model uses a weighting function to account for the level of influence the player has over the location. Physics-based models of pitch control measure the probability that a given player controls the ball if it passed to a given location.

Innovative statistics are appearing in the analytics field of soccer, but that does not mean that traditional statistics are losing importance. Kumar identified traditional statistics that are part of expert ratings of a player. Several of the traditional statistics that affect a player rating are the number of goals scored, red cards received, assists, errors that lead to a goal and others. While traditional statistics have their shortcomings, the data

available for them has the most extended history. Most data sources contain traditional statistics. Modern statistics such as pitch control models require data on the position of the players which only recently became available. Traditional statistics are based on the essential events of the match and are easier to track using manual annotation.

Researchers use analytics to explain shocking performances. An astonishing performance is that of Leicester City winning the English Premier League season of 2015/16 and then battling relegation nine months later. Ruiz, Power, Wei en Lucey used a mixture of traditional and modern performance measuring statistics to compare the team across the two seasons. The number of shots against, goal conceded, and missed shots were compared against the league average and to calculate the expected save value. Expected save value measures the prospect of a goal-keeper making a save from a shot and would fit in the category of modern statistics. Ruiz et al. applied the same approach when appraising the ability of a shot turning into a goal, which led to the expected goal value statistic. Ruiz et al. used the models to compare the expected performance of the players and the teams to their actual performance. Another bewildering event is the loss of Brazil against Germany at the 2014 FIFA World Cup with a score of seven to one. Cotta, de Melo, Benevenuto en Loureiro used a peculiar data set to explain this result, the data from the video game FIFA. The company that develops the video game hires scouts all over the world to rate player attributes as realistically as possible. They identified that Brazilian defenders evolved mostly in non-defensive attributes, such as *Crossing* and *Dribbling*. German players evolved more in offensive attributes such as *Finishing* and *Shot Power*. A strong offensive can explain why the German team scored seven goals.

A goal of analytics is to predict future performance or results of events. The prediction provides value for scouts but also the fans who want to place a bet. van Wijk tried to develop models that can help fans put the winning bet. He developed several prediction models using traditional statistics, such as the number of goals, fouls points. The models had an accuracy of 53%.

It is without a doubt that soccer is one of the most challenging fields for sports analytics, but this means that one has the opportunity to make significant contributions to the field. As more data is captured, researchers have new opportunities, such as identifying new statistics, improve or create existing prediction models.

3.1.2 Baseball analytics or sabermetrics

One of the most successful stories of sports analytics happened in baseball. The term sabermetrics was coined as the definition of on-field sports analytics. Bill James, a security guard with a passion for baseball analytics, wrote the popular *Baseball Abstract* books in the late 1970 and coined the word *sabermetrics* (Beneventano, Berger & Weinberg, 2012). The field became mainstream with the success story of the book Moneyball. The story is about Oakland Athletics, a baseball team. Oakland Athletics identified undervalued players with sabermetrics. The club hired the players and then won 20 consecutive matches. In the present day, baseball clubs can consider sabermetrics a trade secret (Frankel, 2012) as it affects their business operations, such as scouting or salaries (Chang & Zenilman, 2013).

Sabermetrics combines traditional statistics in meticulous combinations that measure a player's performance. There is no clear distinction between traditional statistics and sabermetrics. Traditional statistics appear in box scores, have an established reputation and are easy to understand by the viewers. Some sabermetrics get adopted by media

and Major League Baseball. An example would be on-base plus slugging (OPS), which measures the ability of a player both to get on base and to hit for power (Albert, 2010).

Beneventano et al. evaluated traditional metrics and sabermetrics in the matter of predicting *run production* and *run prevention*. They used stepwise-multiple-regression models where traditional metrics and sabermetrics are the independent variables. The sabermetric *walks plus hits per inning pitched* (WHIP) was the best predictor of runs prevented. Some traditional metrics such as fielding percentage made it into the final model for runs prevented. It was surprising to find that other sabermetrics as *ultimate zone rating* (UZR) did not make it into the final model. The result was similar for the run production model where some sabermetrics made it into the final model, and some did not. Beneventano et al. illustrate how researchers should not overlook traditional statistics when building prediction models.

3.1.3 Basketball analytics or APBRmetrics

Every sport has different strategies and variables that contribute to the success of the team. In basketball, *possessions* form the basis for analyzing a team's performance. A possession starts when one team holds the ball and ends when loses it. Throughout a match, the number of possessions is bound to be almost equal. Kubatko, Oliver, Pelton en Rosenbaum wrote their paper *„A Starting Point for Analyzing Basketball Statistics"* to serve as a base for future basketball research assuring a common terminology. In their work, they introduce basketball statistics such as *Offensive* and *Defensive Ratings*, *True Shooting Percentage*, *Effective Field Goal Percentage*, *Plus/Minus Statistics*, and several others. Analysists identified that the per-minute statistics tend to be consistent even when the on field time of the player varies. The game lasts for 40 minutes in international matches but 48 minutes in NBA. For statistics, analysts use data from the first 40 minutes to keep the consistency between the leagues.

Researchers and basketball enthusiasts try to develop new statistics and models that can predict the outcome of a game. Bhattacharjee en Talukdar use the statistic Player Efficiency Rating (PER) in a linear regression model to predict the outcome of an NBA season by accounting the contribution of player performance to the team's success. In an entirely mathematical world, such a model would excel, but life is full of unexpected events. Nevertheless, one would ask if such a model would predict the outcome of a match, why should the match even be played? Bhattacharjee en Talukdar identified that the model had a few drawbacks such as the team roster changing, low accuracy when the team had a close win ratio, and not lastly, luck.

3.1.4 Cricket analytics

Cricket is a popular game with a rich history. In the classical format, a match would last five days. As the attention span of fans does not last that long, cricket federations sought alternatives. The federations introduced the One Day International and then the *Twenty20* format. The *Twenty20* format has gathered popularity as the game would last 3 hours in normal circumstances. Although the game is similar to baseball, it has complicated rules and several formats. Also, the cricket analytics field is under-developed in comparison with baseball analytics.

Like in baseball, researchers try to identify new statistics that measure the performance of a player. Shah en Shah created a new statistic called FORM, which measures the

form of a player. The new metric takes into account every score of the player, with older ones contributing less to the final result. The development of new metrics that gauge the performance of a player can lead to the creation of better models of prediction of future performance. Verma en Izadi used ML on a multitude of historical statistical data and match state indicators to predict the outcome of a match. The prediction model could pick the winner of a cricket match in 70% of the cases. Bhattacharjee en Talukdar used a different approach at predicting the outcome of a cricket match in the Twenty20 format. Bhattacharjee en Talukdar used the Pressure Index metric, with predictive discriminant analysis to forecast the outcome of the match at different phases of the game. The prediction method had an accuracy of over 70% with the best results at the 12th and 14th over with an accuracy of 78.38%.

Cricket poses a challenge to the analytics field due to the different formats of the game and the multitude of rules that are part of the game. The cricket analytics field is bound to take off as more data becomes available.

3.1.5 Conclusions

Analytics affect sports at different levels, be it scouting for new talent or keeping the audience engaged. During this literature review, we identified the size of the sports analytics research field and its exponential growth (Coleman, 2012) but also how the field touches others such as education (Drazan et al., 2017; Arnold & Godbey, 2011) or law (Frankel, 2012). The sports analytics field is advancing with the development of tools that capture more data (Halvorsen et al., 2013; Gowda et al., 2017) or tools that improve the visualization capabilities (Stein et al., 2015; Hoeve, 2017; Ogus, 2014).

Through this literature review, we gained insights into the requirements and limits of this research project. *Data* is a necessity for any statistical analysis. We identified that researchers have a rich selection of data sources. We can use historical sports data to identify which facts provide valuable information to the reader. For example, we can rank a fact about the number of goals a team scored in the last ten matches by comparing it with the performances of the league average or the teams' historical performance. Using historical data, we can identify when the performance stands out and rank its corresponding fact higher in terms of relevance.

We identified that researchers focus on old and new statistical measures of performances and algorithms to predict future performance (Beneventano et al., 2012) or to detect anomalies (Vinué & Epifanio, 2017). Statistics are about the best performers but can also provide insights into the underperformers. Facts capture the good and the bad, so we need to take such situations into account. Like statistics, facts are about star athletes and the bottom echelon. When e capture the relevance of a SF, we need to take into account that the audience is interested in both situations.

The on-field sports analytics field has evolved with new statistics and data collection systems. Nevertheless, some of the new statistics are obscure, and mainstream media do not use them. The SOCCER framework needs to leverage the popularity of different statistics when identifying relevant facts. A baseball fan will understand facts about simple stats, such as on-base percentage, that have a long history in the sport. However, the fan will not understand facts about obscure statistics, such as Ultimate Zone Rating.

Using the data extractor form illustrated in Annex A.1.5, we identified the statistics used by different sports, the data sources of the given statistics data, the algorithms that employ them, and the research goals when employing sports analytics. We now have an

overview to serve as the foundation for employing relevance measurement of SFs. We can determine properties of relevance that affect the statistics, such as the statistical significance, the complexity, the popularity. We can create measuring guidelines, such as how to establish the complexity of a statistic or its popularity. When we define SOCCER, we take into consideration the requirements of the algorithms that are employed by the given statistics, the available data, and how researchers employ them.

3.2 Natural Language Generation

NLG systems aim to deliver cohesive text that is similar to human authored ones by employing different architectures, algorithms and data sources in a logical process. The NLG system needs to employ a sophisticated process that selects the crucial information to generate text that is of interest to the user, looks natural, and has variety. NLG systems have different classifications based on the system architecture and data source, such as concept-to-text, data-to-text or end-to-end.

An NLG system that generates sports facts or news related to sports needs to meet certain journalistic requirements. Leppänen et al. identified six journalistic qualities for an NLG system. An important quality in the news market is **transparency**. The NLG application can gain the trust of its stakeholders by being transparent in the way it selects and transforms the data into text. The produced text must be **accurate** by using factual data and shouldn't mislead the readers. The application must be **easy to modify** so that it can support a variety of use cases that appear in the newsroom. This requirement is important for media companies that have a limited budget and resources. The generated text needs to be **fluent and coherent** so that the readers are satisfied. The **availability of the data** affects the speed of content production and its newsworthiness. Finally, the NLG application needs to produce **topical content** by discussing events that are relevant to the reader.

The SOCCER framework needs to meet the same journalistic requirements so that it is usable by researchers and practitioners in their processes. Additionally, the framework should improve the enumerated qualities of an application that employs it.

3.2.1 Templated based concept-to-text and data-to-text systems

Concept-to-text and data-to-text applications produce text from non-linguistic data. Data-to-text systems use raw data such as sensor logs to generate the text. In comparison, concept-to-text systems generate text from formal knowledge representations such as ontologies (Lampouras & Androutsopoulos, 2018) . Both systems deal with information in different forms such as databases, ontologies, raw text which we identify using the term *fact*.

In most cases, the systems follow a pipeline architecture based on the stages illustrated in Fig. 3.1. First, the system identifies the facts that should be part of the document and orders them. In this stage, the NLG system could do simple operations such as fact ordering or more complex ones such as deciding the rhetorical structure of the text. At this phase, the document is structured and it contains sets of facts for each entity. Some of the sets can be aggregated. For example, we could have two sentences about a soccer match. A sentence states the number of scored goals. The other sentences could be about the number of missed shots. The two sentences can be aggregated in a phrase that captures both facts. Now, the application can improve lexicality of the text. For example, it can decide to

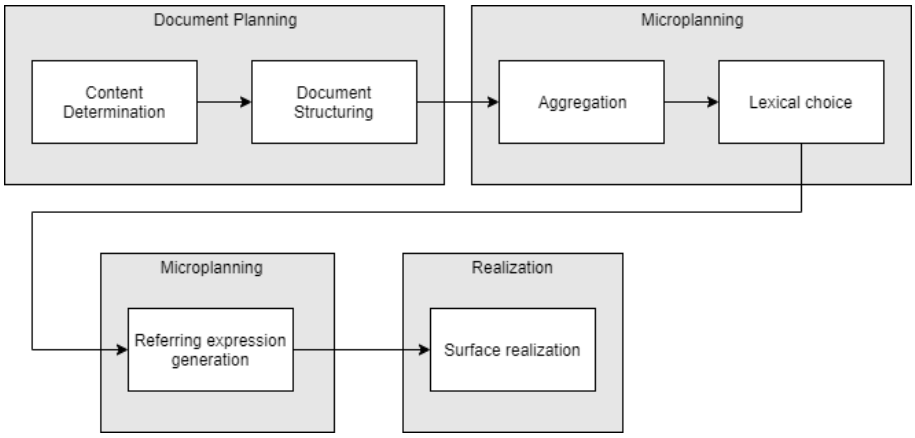

Figure 3.1 – NLG stages (Reiter & Dale, 2000)

replace the 50% ball possession with *average* or *outstanding* ball possession. Afterwards, the application decides which expressions reflect the entities in the facts. For example, it can decide wherever to use the phrase „ *The match between Barcelona and Real Madrid*" or the expression „El classico". Finally, the application produces the actual text that fulfills the rules of syntax, morphology, and orthography. The SOCCER framework can be used in the content determination stage to determine the relevant facts that should be part of the output.

Ontologies serve as an ideal source of semantic representations for the sports domain (Q. Nguyen, Cao, Nguyen & Hagino, August 2012). An ontology describes a conceptualization of a knowledge domain such as sports, consumer electronics. For example, Fig.3.2 illustrates a part of the ontology BBC created for the 2012 Olympics. From the BBC ontology, an NLG application could generate the sentece „Men's sprint is a division of the 2012 Olympics." The ontology can have instances of the each class, text templates for the facts, referring expressions for the classes and much more. Web Ontology Language (OWL) is a semantic web language that can be used to describe ontologies.

Lampouras en Androutsopoulos built NaturalOWL, a concept-to-text system, that generates text from ontology axioms. NaturalOWL converts OWL statements in message triples <S, R, O>, in which S represents an individual or a class, O a different individual class, or data type, and R is the relation or property that connects the two. For example, the class *SoccerPlayer* with the individual *LionelMessit* and the OWL statement *ObjectHasValue(:playsFor :Barcelona* can be converted to the message triple <:LionelMessi, :playsFor, :Barcelona>. A message triple is a fact. Each fact can be expresed through one or more sentence plans (text template). A sentence plan with an OWL schema about Lionel Messi can generate facs such as:

> Lionell Messi was born on 24 June 1987. Lionell Messi players for Barcelona. Lionel Messi scored 681 goals. Lionel Messi has a total of 277 assists.

Lampouras en Androutsopoulos modified the NaturalOWL system to avoid greedy choices that are made individually at each stage. Instead of local choices, the system jointly considers all the stages when selecting the facts. As such, NaturalOWL can generate shorter sentences that contain a larger amount of facts. In the content selection phase, the system decides which facts should be part of the final results based on the *importance score*. In NaturalOWL, all facts have an importance score of 1. The SOCCER framework can

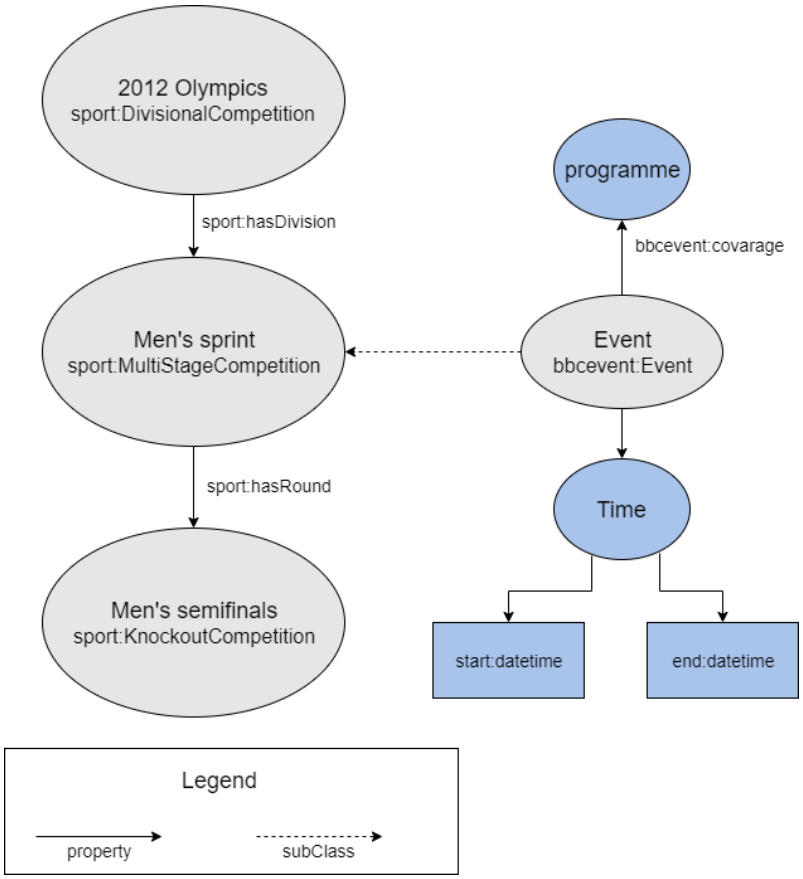

Figure 3.2 – BBC Olympics ontology

improve the content selection phase when dealing with sports data by assigning different importance scores to the facts.

Most data-to-text and concept-to-text systems are template based. The data or the concepts fill a slot in a text template. System that use text templates to generate the text can score low in fluency. In sports, the language can be sensitive. The sports journalists select the performances that stand out. For example, the journalist can use the term „whup" instead of the normal word „beat" if the team won by a large margin. Li, Su, Qi en Xiao developed a model based on the Generative Adversarial Networks architecture to generate sports news from basketball stats. The application uses a neural network to establish which score is suited for a phrase. The approach is used for sentences that describe the general summary of the game. Another neural network determines which player performance is worth reporting on. The model can determine if a player had an outstanding performance on different indexes such as scores or assists and then fills in the details in a template. The player performance model employs similar content properties as those of SOCCER.

3.2.2 End-to-end systems

Pipeline NLG system leverage text templates or other methods to generate text which can have it's disadvantages. The templates are hard to transfer across domains and miss insights, background knowledge and interpretation to a certain level (Kanerva et al., 2019). The improvements in ML algorithms and techniques are influencing the NLP domain. Researchers are employing ML algorithms to build „end-to-end" NLG systems. ML algorithms need a text corpus as a training data which limits the capabilities of applications that em-

ploy them due to the the scarcity of the data. Additionally, end-to-end systems do not respect the journalistic transparency requirements. End-to-end systems output depends on the training data and can not be controlled or explained by the creators.

Among the multitude of ML algorithms, Sequence-to-Sequence (Seq2Seq) models represent the state of the art for data-to-text generation. Kanerva et al. build an end-to-end NLP system using the Seq2Seq algorithm. The text corpus comprised of news articles covering more than 3400 ice hockey games. Additionally, the news articles were linked to game statistics that were stored in text files. The dataset required manual work to link text spans to the corresponding game events detailed in the statistics. Furthermore, the system required manual annotation of the game events to the corresponding text in the news article. The performance and the output quality of an ML model is highly dependent on the input data, as such it is mandatory to clean and align the data. Otherwise, the system will not bring the desired results. The reports generated by the Seq2Seq model were evaluated by two news reporters. They deemed that 75% to 90% of the generated text can be directly used in a news agency. Kanerva et al. mention that it was challenging to decide on which aspects and particular features of an event the system should verbalize. For example, journalists sometimes mention who assisted a goal but in other cases it doesn't. SOCCER could guide on which actions of the match the researchers or an algorithm as Seq2Seq should focus on.

End-to-end systems can also fall in the area of data-to-text systems. The neural network can learn which statistics can be linked to certain phrases (Kanerva et al., 2019). The issue is that the neural network can "hallucinate" and generate false facts. A different architecture is to precompute certain statistical features of the sports data which then acts as a training feature for a Seq2Seq network (Nie, Wang, Yao, Pan & Lin, 2018). The pre-executed operations quantify the result and the operation but also the embedding for the operation and for the result. They are then used in the Seq2Seq model to decide certain lexical choices for the output. The neural network with the pre-executed layer outperforms other similar algorithms and architectures for automatic corpus measures. It illustrates how end-to-end systems can use different precomputed metrics as a training feature to improve their performance for certain tasks of the NLG process.

3.3 Summary

A multitude of actions take place in a sport event. In soccer, the players are situated on a field, pass the ball around and use different strategies to score goals. This leads to normal statistics such as ball possession, the number of goals scored on average and others. But, as more data is captured, researchers and fans develop advanced statistics and models such as the „Voronoi dominant regions" (Bornn et al., 2018). Socccer is a sport that is lacking in terms of statistics compared with baseball, where sabermetrics plays a crucial role such that is considered a trade secret by clubs (Frankel, 2012). The successful story of sabermetrics is influencing the other sports. Such is the case of APBRmetrics in baskeball, a field that captures the development of advanced baskeball statistics. Sports statistics are the backbone of sports facts. Sports statistics are used in NLG systems to generate sports facts that explain the event. The statistics of the match dictate if a team had a close win or „whupped" the other one (Li et al., 2019). Statistics evolve and become increasingly complex, thus they are hard to understand by the average viewer. Such is the case with soccer, where player movement and events affect the game in a grander sense. Researchers

develop viewing tools for analysing interesting situations (Stein et al., 2015) or to illustrate different complex statistics straightforwardly (Hoeve, 2017). Besides visual tools, natural language is a perfect tool to explain such complex statistics and situations, thus a job for NLG systems. As sports statistics evolve, change and multiply, NLG systems need to do the same. The content determination phase needs to be adaptable to current situations but also future ones. SOCCER framework abstracts relevance properties and measuring guidelines as such can be used to build evolvable and adaptable heuristics for relevance measurement.

Sports is often the subject of NLG research as there are different open-source and commercial datasets with sports statistics available. Data-to-text NLG systems will generate the output based on a predefined multi-step process illustrated in Fig. 3.1. In the *content determination* phase, the NLG application will select and filter the content. In sports, this is the step in which the NLG application will identify which statistics, events and actions will use to generate the output (Lee et al., 2017).

NLG systems need to fulfill certain journalistic requirements such as *fluency*. Data-to-text systems will fill in a template with their data which can limit their fluency. But, template based NLG systems have evolved over time that can be compared to the so called end-to-end NLG systems (van Deemter et al., 2005). End-to-end systems will use ML algorithms to train models that will generate natural language. They bypass the NLG steps from Fig. 3.1 and produce the output in one stage. End-to-end systems are highly reliant on a text corpus to train the ML models, but, such data is scarce. Kanerva et al. used a text corpus with more than 3400 ice hockey games to train an ML model. The training data was formed of text phrases that described sports facts and their corresponding statistics or events. The dataset required cleaning and aligning before it could be used for the ML model. Each type of system has its weaknesses and strong points in terms of fulfilling journalistic requirements. As mentioned, data-to-text systems are often template based. Thus, it is hard to adapt them to different domains and sports as they require new templates van Deemter et al.. Data-to-text system follow a systematic process to generate they output as such they can be transparent on how they produced certain natural language. Not the same can be said about end-to-end NLG systems. End-to-end systems are based on certain ML models which are considered to have a „black-box" architecture (Leppänen et al., 2017). SOCCER will assist both types of NLG systems in improving their weaknesses in terms of adaptability, respectively transparency. The relevance properties and measuring guidelines are abstract and can be applied to different sports with small modifications. Thus, the content determination heuristics could be used for different sports. At the same time, the properties can be part of the training data of the ML algorithm that is used by end-to-end systems. The properties should affect which content the end-to-end system uses to generate the output, thus it will be transparent. Additionally, researchers could use the technique *transfer learning* to share the content selection ML model to other sports or sport related ML algorithms.

Relevance has been approach in different ways in sports related NLG systems. Zhang, ge Yao en Wan uses a mix of content relevance properties and word count based metrics to identify sports sentences from live commentaries that can be used to construct sports news articles. Words in the sports facts are split in two categories, stop words and highlight markers. The content relevance heuristic treats sentences which contains a higher counts of stopwords as less important. Sentences which contain a high number of highlight markers have a higher relevance value. At the same time, the system calculates the similarity

between neighboring sentences to reduce the redundancy of the text. The content selection heuristic also took into account several relevance properties when filtering sports facts. The author used 25 keywords to detect several *important actions* such as cards, goals, free kicks and others. Such actions are often taken by different players, some who are well known by the viewers. The author used the relevance property *player popularity* to identify sentences about the stars players. Zhang et al. identifies the number of players in a sentences and then calculates a sum of their popularity based on how often they appear in the news. In a followup research paper, the system was adapted to do the news generation in a real-time fashion. The new system included a new metrics, *phrase newsworthiness* which measured if a certain phrase appeared often in the news (ge Yao Jianmin Zhang Xiaojun Wan Jianguo Xiao Institute of Computer Science et al., 2017). A similar property was used by Renjun et al. where they constructed a dictionary with soccer news and a neural network would identify which sentences were similar to professional news content. Relevance content properties such as *player popularity*, *action importance* are used end-to-end systems that involve ML algorithms too (Li et al., 2019; Kanerva et al., 2019). Such systems illustrate how SOCCER relevance properties and measuring guidelines can be integrated with different ML models and end-to-end NLG systems.

Sports involve different statistics and databases filled with data and knowledge about them. Thus they form a knowledge database which is limited to sports. Association rule mining is an operation that is often executed for knowledge bases to identify interesting patterns in the data. For example, an association rule would be if athlete X plays in the league NBA then he plays the sport basketball. Association rule mining algorithms will identify extremely large number of rules, thus researchers investigated different interestingness measures to order or filter them (Lenca, Vaillant, Meyer & Lallich, 2007). Lenca et al. identifies two types of interestingness measures for association rules: (i) objectives and (ii) subjective measures. Objective measures are data-driven while subjective measures are user-driven such that they take into account the user's a priori experience and objectives. SOCCER framework relevance properties fall in the category of subjective metrics. *Novelty, surprisingness, pecularity, unexpectedness* are several subjective measures of interestingness for association rules (Hrovat, Jr, Yermak, Stiglic & Fister, 2015; Shaharanee, 2012; Kontonasios, Spyropoulou & Bie, 2012). Subjective measures of interestingness are abstract as they need to be applied to different association rules that are related to the domains part of the knowledge database. The properties have measuring guidelines that are clear and which allow the user to adapt it to their needs with different algorithms. For example, unexpectedness identifies patterns which contradicts what the user knows or believes about the data or it contradicts common domain knowledge (Kontonasios et al., 2012). Thus, we need to quantify the degree of contradiction to measure the unexpectedness of a pattern. Then, the metric and the measuring guidelines fits different domains. SOCCER uses the same abstraction to increase its adaptability to different sports.

In this chapter we introduced the field of sports analytics with a focus on on-field analytics in Sect. 3.1. The section introduces how sports statistics are used to analyse and predict outcomes of sports events. Researchers develop new, complex statistics but also tools to visualize or capture the data. In Sect. 3.2 we present different architectures and types of NLG systems. Additionally, we show how sports based NLG system perform the content selection task and present ideas for how SOCCER could improve them. Furthermore, we identified the journalistic requirements for NLG systems need to fulfill and showcased the weakness and strong points in terms of requirements of different NLG

systems architectures. Finally, in this section we briefly presented the facts above with the addition of how relevance is approached in association rule data mining.

Chapter 4
The Sports Facts Relevance Measurement Framework

SOCCER is the main artifact of this research project. In this chapter we will present the design process and its results. SOCCER is designed to be used in the news and betting market by NLG systems. The markets and the NLG systems have certain requirements and limitations that the framework needs to consider. Thus, in Sect. 4.1 we elaborate on the said requirements.

4.1 Framework requirements

Researchers and practitioners need to develop NLG systems under certain limitations. For example, NLG systems developed for the news market need to follow the same journalistic requirements as the reporters (Leppänen et al., 2017). Also, the algorithms have limitations due to the availability of the data. Thus, we will elaborate on how these requirements affect the framework.

NLG systems target at the sports domain will use the SOCCER framework to determine their output. The given NLG system need to fulfill the same requirements as journalists (Leppänen et al., 2017), as in :

- Transparency
- Accuracy
- Modifiability and transferability of the system
- Fluency of output
- Data availability
- Topicality of news

Transparency. These requirements are derived from the ethical challenges that algorithmic journalism encountered (Dörr & Hollnbuchner, 2017). As part of a research paper, the SOCCER framework was designed in a transparent and scientific way. For NLG systems, Leppänen et al. states that transparency might include making certain steps of the process public such as data sources, data selection methods, the source code. An NLG system that use the SOCCER framework increases its transparency by using a public available framework. The users of the framework can publicize which metrics and measuring guidelines they used to increase the transparency of their NLG system. As the users will build relevance measurement heuristics they can choose to make those public too or keep them as a proprietary secret for commercial systems. Transparency is a simple to present for data-to-text systems as they can provide the relevance heuristics used for the content determination stage. End-to-end NLG systems have certain issues with transparency as they are „black boxes" which are depended on the labeled trained data or its text corpus.

End-to-end systems can integrate metrics from the SOCCER framework in its model to increase transparency.

Accuracy The output of an NLG system needs to be based on factual data and shouldn't provide misleading affirmations. Relevance measurement heuristics that are based on SOCCER metrics need to identify factual data that matters and that won't be miss interpreted by the final user. This is especially important for gamblers who wage their money but also for the betting company which is regulated by law.

Modifiability and transferability of the system This requirement is important due to economic reasons. Media companies have a limited budget thus the NLG system needs to produce content for several domains. The sport domain is vast, for example, only the Tokyo 2020 Olympics will feature 33 different sports One of the strong points of the SOCCER framework is that it is based on the content of a sports fact and its properties that affect the relevance. As such, most relevance heuristics that are based on it will be suited for most sports.

Fluency of output We mention that **fluency** is motly related to the form of a sports fact thus not related to the topic of this research thesis. But, the SOCCER framework can be extended in this direction.

Data availability Journalists and NLG systems need data to produce content. In sports, data is used to also measure the performances of a team or athlete. All types of NLG systems need data in one format or another. Data-to-text system need the sports data for which to generate the content. End-to-end NLG systems are based on machine learning model thus they need labeled data to trained the models. The users of the SOCCER framework should first run an analysis of the data available before developing the relevance measurement heuristics or the NLG system. Most metrics from the SOCCER framework need certain data to determine the relevancy of SFs. The users of the framework should use the metrics for which they have available data. Also, an NLG system needs to use data that is newsworthy and should ignore noise. Sport events produce huge amounts of data which makes the process of separating noise and newsworthy data a challenge. We designed SOCCER framework taking this challenge into account. Thus, we threat all types of content and content properties that can be derived from data.

Topicality of news An NLG system needs to generate content that is of interest to its users such as recent events, local events or others. Topicality is one of the challenges tackled in this research project as we try to identify the differences between the news and the betting markets with **SQ1.3**. Topicality is covered by SOCCER using several relevance metrics.

4.2 Sports facts content types

A SF is formed by content which is expressed in natural language in an expressive way by journalists to gather the attention of their readers. For example, in the Champions League final of 2019, the captain of Totteham had the lowest amount of ball touches from his team which is described using the fact **CF9**. At a first glance we identify that the SF content

is formed by the athlete with his role, the action as the ball touches and the result as his number of touches. But, behind the simple action and the result there is a data query and the data related to the result. As such, we created the content type Template and Instance which capture the full context. We will describe the two content types and the others in this section.

In the following we will use the Champions League final of 2019 to provide examples for the concepts we will introduce. We picked this event as it is a well known soccer event which can break viewership records (Slater, 2019). In his article, Raynor arguments the importance of full-backs for winning the match, as such, **CF1** contains a fact related to this position. **CF2**, **CF3**, **CF4** were selected from an article published on the official website of Liverpool. Raynor presents in the article the possibility of Sadio Mane breaking the record of being the first Senegal internation player of Liverpool to score back-to-back in the competition. Additionally, the article contains several other facts. Money plays an important role in soccer. Clubs can buy experienced sports players to strengthen their team. In his article, Manoli arguments how money plays an important role in winning. **CF5** and **CF6** are facts that are linked to the economics of the sport. McMahon presents the event from an historical perspective by describing the teams road to event final and several historical facts from which we selected **CF7** and **CF8**. The previous facts are mostly from pre-match articles. Grez does a post-match analysis from which we selected **CF9**.

Champions League Final of 2019 facts

CF1 „No defender has ever registered more assists in the top division of the English league than Trent Alexander-Arnold's 12 this season." (Wilson, 2019)

CF2 „In nine games against Tottenham, Liverpool manager Jürgen Klopp has lost once, with four victories and four draws." (Raynor, 2019)

CF3 „The German is the first manager to take an English club to a trio of European finals in his first three seasons of European competition." (Raynor, 2019)

CF4 „A goal at Estadio Metropolitano in Madrid would make the Senegal international the first ever Reds player to score in back-to-back finals in the competition." (Raynor, 2019)

CF5 „According to the latest Deloitte Annual Review of Football Finance, the English top league had a total revenue of €5.44 billion in the 2017-18 season. This is almost the same as the next two leagues in the rankings put together, Germany's Bundesliga (€3.17 billion) and Spain's La Liga (€3.1 billion)." (Manoli, 2019)

CF6 „In 2018-19, only Chelsea outspent Liverpool in terms of transfer market activity." (Manoli, 2019)

CF7 „Often overlooked is the fact that Spurs was the first British team to win a major European trophy. Back in 1963, Spurs beat Atletico Madrid 5-1 to win the now-defunct European Cup Winners' Cup." (McMahon, 2019)

CF8 „170 – The number of times, to date, Liverpool and Spurs have played each other in all competitions. Liverpool has won 79, Spurs' 48 and there have been 43 draws." (McMahon, 2019)

CF9 „The Spurs captain mustered just 11 touches of the ball in the first half, less than every other player in white." (Grez, 2019)

CF10 „The battle between Tottenham and Liverpool at the Wanda Metropolitano will go down as one of the cleanest Champions League final in history, with not a single card shown." (Ángel García, 2019)

CF11 „The last time the final of the continent's elite competition finished without a card was in 1988 when it was known as the European Cup, in 1988's showdown between PSV Eindhoven and Benfica." (Ángel García, 2019)

First, we will define the type of content that form a sport fact and can influence the relevance of a SF through its nature. The content types are a core part of the SOCCER framework as they influence how relevance is measured for a given relevance property. Companies will use different schema and methods to share the metadata. The International Press Telecommunications Council (IPTC) provides a group of standard markup languages that can be used to share metadata about certain media topics, including sports. Sports Markup Language (SML) schema provides a standard XML structure for sports data interchange (*SportsML - IPTC*, z. j.) . The SML schema is open source and free to use

by companies. Also, the schema is designed to be adaptable to all sports. As such, we extracted several content types based SML schema and the structure of SF. By relation, we coop the same properties in SOCCER and fulfill the enumerated requirements of transparency and transferability from Sect. 4.1. In the following we will present each content type with a few examples based on Champions League facts enumerated previously.

Event

Sports facts are about different events, such as a soccer match. But, we use the term Event to refer to also different stages of a tournament and even the tournament itself. For example, fact **CF10** captures three Events: the Champions League, the final phase and the match itself. Thus, we identify the following content of type Event in Sect. 4.2:

- Premier League in **CF1**, **CF5**;
- Champions League in **CF3**,**CF4**,**CF9**,**CF10**,**CF11**
- Final of Champions League in **CF3**,**CF4**,**CF9**,**CF10**,**CF11**
- Bundesliga in **CF5**
- La Liga in **CF5**
- European Cup Winners' Cup in **CF7**

IPTC uses the term „sports-event" to describe the structure that contains a set of teams or players with optional data about the event such as actions, highlights and so on. At the same time, the sports-event can contain sub events that describe the schedule or the tournament structure. As such, we use the term **Event** to describe events of different granularities. As such, the **Event** can mean a certain moment in a soccer match, the match itself or the league that the match is part of.

Event participant

With the content type Event Participant we try to capture all persons which participate in a sport at a given event and not only such as athletes, coaches, referees and others. We can measure the influence of an Event Participant on relevance using the same relevance properties when he is a coach, a player or another person that is involved in the sport. We identify the following content of type Event Participant in Sect. 4.2:

- Trent Alexander-Arnold in **CF1**
- Jürgen Klopp in **CF2**, **CF3**
- Sadio Mane in **CF4**
- Harry Kane in **CF9**

IPTC uses several terms to describe the Event Participants of an event. For example, it uses the term *player* for athletes that participate in a sport and *officials* for the referees. In SOCCER framework, we define them as Event Participants has we can measure the influence on relevance in the same way.

Role

Event Participants denominate different persons who are involved in the event. The Event Participants have different roles. They can be coaches or athletes. Additionally, in some sports the roles are specific. For example, Grez accentuates that fact **CF9** is about the Tottenham captain. Additionally, the Role content type can refer the nationality of an Event Participant. In general, the Role can identify a particularity of an Event Participant. We identify the following content of type role in Sect. 4.2:

- Defender in **CF1**
- Manager in **CF2**, **CF3**
- German in **CF3**
- Senegalese in **CF4**
- Forward in **CF4**
- British in **CF7**
- Captain in **CF9**
- Forward in **CF10**

In soccer, IPTC uses the role as metadata for actions such as that the player was a striker, the corner kicker and so on. In SOCCER, the content type Event Participant takes a generic approach which abstracts it role. As such, we miss information on how the Role of the Event Participant influences his relevance to the event and the sport itself.

We assumed that the Role content would extensively impact the relevance of a sports fact for certain metrics. But, most sports experts had the opinion that the impact of the Role content type to the relevance of a sports fact is captured mostly by the Event Participant. For example, CE2 stated „If you make a statistic about Lionel Messi or Cristiano Ronaldo it is not about their role, it is irrelevant at that moment". Thus, we need to mention, that in some cases of extremely relevant Event Participants, the impact of Role content type to the relevance of a sports fact would be reduced. But, it can have influence the relevance in some cases, such as when discussing about the nationality of Event Participants as stated by PO3 „ ... I never cared about the location basically, and I definitely do not care about the role. But I care about the nationality of the event participant ...". Thus, the influence on relevance of the Role content type is situational.

Team

The Team content type entitles a group of Event Participants that act in a sport under the same banner. For example, in soccer, the players play for a club. Sports fans are usually extremely loyal to their teams in sports where they appear. Thus, the influence on relevance of the Event Participants is reduced in such cases. As such, although a team could also be an Event Participant, we created it as a separated content type. Additionally, the relevance metrics for the two content types can have different sources of data and implementations. We identify the following content of type team for the SFs in Sect. 4.2:

- Liverpool in **CF1**,**CF2**,**CF3**,**CF4**,**CF6**,**CF8**,**CF10**

- Tottenham Hotspur in **CF7**, **CF9**, **CF10**
- Chelsea in **CF6**
- Atletico Madrid in **CF7**
- PSV Eindhoven in **CF11**
- Benfica in **CF11**

IPTC uses the *team* element to define the teams that participate in an event. The team element contains the aggregated information for it such as the players and is part of the top level structures from the SportsML standard.

During the interviews, PO3 stated that the Team content type could be included in the Event Participant. PO3 went further and stated that the Team and the Event Participant could capture the same relevance metrics „They (the team) can change where they play, you can relegate, you can promote. There are so many things that apply to the team as they apply to the event participant". In SOCCER, we choose to keep the team as a separated entity as the data sources and the methods to calculate different relevance properties differ. Also, by keeping the two separate the framework and its content types would be easier to understand by researchers and practitioners. But, the idea can be included in further research if the there is a need for such a change.

Location

Every event takes place in a certain location such as a venue, city. Facts can mention the location such as **CF4** and some might be linked to different statistics. Also, the location can describe the position where certain events take place. For example, in soccer, the player could score from inside the penalty area or outside. Such details matter as they could make a fact more interesting and thus relevant to the audience. We identify the following content of type Location in the SFs in Sect. 4.2:

- Estadio Metropolitano in CF4,CF10
- Madrid in CF4

IPTC links the location to the sport event as the physical place in which the event takes place. But, the location entity is linked to entities such as the player as their birth date or nationality, or even to the team. In SOCCER, the location could also identify the position where a certain action took place in an event.

During the interviews, the sports experts identified that the Location content type, in particular the venues, as having a lower weight on the overall importance of a SF compared to other content types. As quoted before, PO3 stated that he doesn't care about the location. Also, if we take a look at the results from the interview in Appendix B.4, we can see that there were quite a few cases where the sports experts didn't consider that the Location content type influences the relevance of a sports fact in multiple cases for different relevance properties.

Template

The **Template** represents the core of a sports fact and its underlying data. SF are created from data which is identified and analysed using certain data queries. The Template,

in essence, captures the form of those data queries. The data queries can be general or specific depending on the architecture or the goal of the NLG system. For example, a general Template content type would be „*goals as **Role** for last **X** matches*" while a specific Template would be „*goals as defender for last **X** matches*". In essence, we define the Template content type similarly to a template in the NLG field.

Taking the SFs listed in Sect. 4.2 we identify the following content of type **Template**:

- *assists as **role*** in CF1
- ***Event Participant** results against opposing team* in CF2
- ***Event Participant** to win trophies in **event** list during his first **X** time* in CF3
- *scoring back-to-back at **Event** for **role*** in CF4
- *financial results of league* in CF5
- *transfer value of **team*** in CF6
- ***team** to win **event*** in CF7
- ***teams** head to head fixture results* in CF8
- ***role** ball touches in **event*** in CF9
- *penalties record in **event***
- **penalties record history in event**

In the SML schema, IPTC defines *sports-content* as a top level object to share news related content. It can contain meta data about the content itself and data about scores, standings, statistics, etc as illustrated in it's schema in Listing 4.1. The *sports-content* entity from SML schema can be used to capture the structure related to a SF. Based on the structure, we can apply relevance metrics from the SOCCER framework to measure the relevance of that SF.

Listing 4.1 – SportsML source schema for the sports-content entity

```
<xs:element name="sports-content">
  <xs:complexType>
    <xs:sequence>
      <xs:choice minOccurs="0" maxOccurs="unbounded">
        <xs:element name="sports-metadata"
          type="sportsMetadataComplexType">
        </xs:element>
        <xs:element name="sports-event" type="sportsEventComplexType"/>
        <xs:element name="tournament" type="tournamentComplexType"/>
        <xs:element name="schedule" type="scheduleComplexType"/>
        <xs:element name="standing" type="standingComplexType"/>
        <xs:element name="statistic" type="statisticComplexType"/>
        <xs:element name="article" type="articleComplexType"/>
      </xs:choice>
    </xs:sequence>
    <xs:attributeGroup ref="commonAttributes"/>
```

```
  </xs:complexType>
</xs:element>
```

Initially, we used the name *Action* for the Template content type. The Template, at its core, captures a certain action that happens in a sport, but it also captures the context around it. For example, when the SF involves a certain role, the related event and so on. But, the name *Action* implies that it is only about the actions that occur in sports which lead to miss understandings and confusion for the sports experts during the interview. The term *Template* captures the broad meaning of the content type. Also, the term is already used in the NLG community, thus it will be easier understood by the beneficiaries of this research project.

Instance

The **Instance** content type identifies the data and the results that fill in the content type *Template*. To generate a SF, you need the data to back it up, either historical or from a source that contains historical data. For example, **CF1** refers to historical data on the assist of all defenders for the Premier League. Additionally, the data query that is describe by the Template will produce a result, such as *most* for **CF1** or the revenue in **CF5**.

We can debate that the data and the result could be designed as two different content types. But, they are close intertwined and their relevance is interdependent. Data without a result doesn't hold much relevance as it is hard to interpret and understand. Results without data don't have a factual basis and are limited on their value to a sports fan. Thus, we find that the *Instance* content type is a solution that links the two and provides a way to measure the relevance of a sports fact.

We identify the following content of type **Instance** for the Champions League SF listed in Sect. 4.2:

- For **CF1**, the number of assists per season in EPL per **role** represents the data part of the instance type. The result is captured by the *most*.
- For **CF2**, the previous head to head results between the team and the manager represents the data. The number of losses, victories and draws represents the result.
- For **CF3**, the teams who went to finals of European competition and their managers represents the data. The result is that the German manager would be the first to reach three finals during his first three seasons.
- For **CF4**, the people who scored in Champions League finals from Liverpool represent the data. The result is that there wasn't a person of Senegalese nationality from Liverpool who scored in consecutive finals.
- For **CF5**, the revenue data for soccer leagues represent the data. The result is that the Premier League revenue is more that the next two combined.
- For **CF6**, the transfer market activity represents the data. The result is the two teams who spent the most.
- For **CF7**, the teams who won European trophies and the score represents the data. The result is the first team that won and the score.

- For **CF8**, the previous head to head events between the two teams and their results represents the data. The number of losses, victories and draws represents the result.
- For **CF9**, the number of ball touches of the players in the first half of the final represents the data. The result is the player with the least ball touches.
- For **CF10**, the number of cards in each Champions League final acts as the data. The result is that the match between Tottenham and Liverpool is one of the few with no cards.
- **CF11** has the same data as **CF10**. The result is the last time the same situation happened.

In SML, the Template resents the source and the schema of a data transfer. The Instance content type represents the values that fill in the schema. For example, Listing 4.2 contains an instanced schema for a soccer team stats but also a player. The listing contains several different statistics and results for Liverpool and the player Virgil van Dijk.

Listing 4.2 – SportsML data instance for team and player soccer stats

```
<sports-content>
    <statistic>
        <statistic-metadata team-coverage="single-team"
            temporal-unit-value="2020"/>
        <team>
            <team-metadata>
                <name>Liverpool</name>
            </team-metadata>
            <team-stats>
              <stats>
                <stat value="58"/>
                <stat stat-type="shots-total" value="90"/>
                <stat stat-type="goals-total" value="14"/>
              </stats>
            </team-stats>
            <player id="p.41328">
                <player-metadata position-regular="defender">
                    <name>VIRGIL VAN DIJK</name>
                </player-metadata>
                <player-stats>
                  <stats temporal-unit-type="season">
                    <stat value="788"/>
                    <stat stat-type="passes-total" value="478"/>
                    <stat stat-type="passes-complete" value="378"/>
                  </stats>
                </player-stats>
            </player>
        </team>
    </statistic>
</sports-content>
```

Initially, we named this content type as *Statistic* due to its common use in sports to describe the results and the data that is captured in a sports statistic. But, the sports experts had trouble to understand that the content type describes the data and the result. This lead to challenges in evaluating the relevance property of *Timeliness* which we will discuss in the following section. Also, the current identifier fits in the current naming scheme of the content types and it is a direct correlation to the SML schema data instances.

4.3 Relevance metrics

A relevance metric is a set of guidelines on what and how to measure the relevance related to a property for a given content type. During the literature review, we extracted the content types and the relevance properties researchers leveraged to do content filtering for their use case. We consulted sports experts to identify which properties can and should be measured for the content types. The process resulted in 26 relevance metrics as illustrated in Table 4.1. We didn't perform all changes to the metrics suggested by the sports experts due to two reasons. Firstly, the sports experts misunderstood some of the metrics, thus their feedback was misplaced in such a case. Secondly, the inter-rater reliability coefficient for their answers was 0.72 where a value of 1 means total agreement. Therefore, they had diverging opinions in some cases. As such, we performed the changes that the sports experts motivated with convincing arguments and were in agreement.

SOCCER	Event Participant	Role	Team	Location	Event	Template	Instance
Popularity	Metric 1						
Newsworthiness	Metric 2						
Importance	Metric 3	Metric 4	Metric 5		Metric 6	Metric 7	Metric 8
Significance					Metric 9	Metric 10	Metric 11
Unexpectedness						Metric 12	Metric 13
Complexity						Metric 14	Metric 15
Sentiment	Metric 16						
Timeliness	Metric 17				Metric 18		Metric 19
Novelty						Metric 20	Metric 21
Utility	Metric 22						
Peculiarity						Metric 23	Metric 24
Predictability						Metric 25	Metric 26

Table 4.1 – SOCCER relevance metrics

Each relevance metric targets relevance from the perspective of its corresponding content type and relevance property. The metrics treat general and particular cases in which the property influences the relevance of a SF for a given content type. SFs about a popular athlete has a higher relevance compared with unpopular ones (Li et al., 2019). NLG system can cover such cases by implementing Metric 1 in the content determination stage. Each relevance metric can be used in an NLG system to determine relevant content that should be part of the output.

In the following we will discuss the relevance metrics that can influence the relevance of a SF. Each relevance metric contains a general description of what it should measure and how it can be computed. The metrics have broad definitions as they are designed to be used in all sports and different NLG systems. For each metric we will present a series of cases where it can be applied and the opinion of the sports experts.

Popularity

Popularity is a relevance property that should measure how well known is a content type to sports fans compared with content types of the same category. The property would then identify content and SF that would engage the sports fans as they understand it and have knowledge about it.

Most of the times, we see news related reports to famous players or athletes as noted by Li et al.. Due to the same reasons, popularity was one of the first properties that we came up intuitively. Famous athletes gather fans who want to learn and hear more about them. Thus, SF about popular content types become relevant to them. Zhang et al. leverage this idea in their NLG system to select those sentences which mention a popular athlete. In addition to the athletes, we assumed that the other content types could also be popular and influence relevance. Therefore, we used sports experts interviews to validate the assumption.

Metric 1: Content type popularity

It should measure how well known a content type is to the sports fans compared with other content types of the same type.

Possible data sources are social media websites, commercial products that measure brand popularity, sports experts or other sources that can give a value on the popularity of a content type.

The sports experts had diverging opinions for which content types we should measure Metric 1 as illustrated in Sect. B.4.1. All sports experts stated that we should measure Metric 1 for the Event Participant content type. RE1 stated that he considers that the Event Participant popularity influences the relevance of a sports fact less than the popularity of the Team or Event. PO1 considered that the popularity of the Team has the greatest impact on relevance followed by that of the Event Participant. When asked which content type can be popular, CE1 stated „Definitely the participants because sports fans always want to know about the players and the team. They follow the players and the teams". Overall, all sports experts identified that the Event Participant popularity can influence relevance and was on the top of their choices.

The sports experts identified that the popularity of Role content type has the least impact on relevance when compared to other content types. PO1 disagreed with the fact that the popularity of Role and the Location could influence the relevance of a SF. PO1 stated that „There is no popularity for the role in a team. And, it's also because the roles in the team can be arbitrary. So when José Mourinho is a trainer he is probably the most known in the team. So now, if José Mourinho is the trainer of Tottenham then he will be the most well known." which exemplifies how the certain metrics have a huge influence in relevance that would make other metrics absolute. But, when the Event Participant is less well knowm, his role could make a difference. When presented with this situation, SE1 choose to agree to the fact the popularity of the Role could influence the relevance of a SF. In fact, even PO1 identified that the content type could have a certain influence on relevance. But, there is too little influence for him to consider it. RE1 had a similar opinion. He stated that the Role wouldn't add much value when the fact is about a famous athlete such as Lionel Messi. But, he agreed that strikers are more popular than other Roles.

Most sports experts considered the popularity of the Team content type as having one of the highest influences on the relevance of a SF. The teams play an important part in sports that are team based. Athletes might move to a different team or retire, but the team can be present for a long time. For example, in soccer, you will have 11 players on the field and a few on the bench. It might be that most fans don't know all the players. RE1 stated that the preference is similar for news reports about amateur matches. There, journalists and fans are more interested in facts about the team. But, RE1 stated that the situation changes where there is a famous player like LeBron James. Then, fans and journalists are interested in facts about the athletes too. As with the Event Participant content type, all sports experts considered that the popularity of a Team influences the relevance of a sports fact as illustrated by their choices in Appendix B.4.1.

The popularity of the Location influences the relevance of a SF destined for particular sports and specific location types. During the interviews, the sports experts found that for soccer, if a venue is part of a SF it wouldn't influence its relevance and it is mostly additional information. SE1 stated that the popularity of the Location becomes relevant only when the stat is about the venue „I mean, if you have a statistical fact about matches in a certain stadium such as this is the 1000th match in X stadium, then I would check that mark (*it would influence relevance*)". RE1 considered that the Location content type, in particular venues, only matters when it is the home stadium for the fans of that team. But, when discussing other sports, such as auto racing where you have certain sections of the race track that have a reputation, or cycling road where different stages are more well known and gather more interest than others, RE1 agreed that the popularity of the Location could have an influence on relevance. PO1 didn't consider that the popularity of a Location has a large impact on relevance of a SF. CE1 selects relevat SF that are then delivered to clients from the news and the betting domains. Thus, when I asked if she received any fact request about certain Locations, CE1 included the content type Location in Metric 1. Overall, the sports experts did identify situations where the popularity of the Location could impact the relevance despite their choices. Thus, we decided to keep the content type Location part of Metric 1. But, we mention that Metric 1 should be measured for the content type Location only in sports or situations where it can make a difference. The suggestion stands for all the relevance metrics in SOCCER. An NLG system might not need all the relevance metrics to deliver acceptable results.

The sports experts considered that popularity of the Event has a large impact on the relevance of a SF together with popularity of Teams and Event Participants as it is illustrated in their choices in Appendix B.4.1. The popularity of the Event can be used to identify facts that are relevant to a local or global audience. PO1 gave the following example „So when Ajax play in Champions League is of more importance for people around the world then when they play FC Twente in the dutch league. Because, nobody in Germany, France or England would follow that match. But, if they play in Champions League against Chelsea for example, than it is important for fans". In his statement, we get a glimpse on how the importance and the popularity of an event influence the relevance of a SF. The match would be important since it is in Champions League, but it is also popular to the fans due to similar reasons. The popularity of an Event is particularly helpful when an NLG system needs to generate and select facts from different Events. NLG system can use Metric 1 for the content type Event to deliver content that is relevant to a local audience.

An NLG system can generate a large variety of SFs. Additionally, the NLG system can be extended during it lifetime to generate new SFs. We can identify a Template content

type for each of the SFs. Some of those Templates might be less known to the end-users of the NLG system. Therefore, the popularity of a Template would differentiate in terms of relevance between them. Also, some Template can involve sports actions that are just popular to the fans and they want to read about it. In soccer, for example, sports fans might want to know about who scored the most hat tricks. Thus, an NLG system can use the popularity of a Template to identify SF that involve popular actions. PO1 opinion was that popularity should be measure only for the Event Participant, Team and Event content types and not for the others. But, the other sports experts identified that the popularity of a Template can have an influence on the relevance of a SF when it comes about certain sports actions. During the literature review we identified different possible Templates for statistical types such as the Voronoi model for soccer or sabermetrics for baseball. SF that involve recently developed statistical types might be less known to the general audience, therefore less relevant.

The Instance content type describes the data behind the Template but also its result. A multitude of results could be computed for a given Template depending on the data. For example, an NLG system can generate a SF every event for the template *player X participated in Y event Z times*. Participating two, three times in a given event isn't necessarily popular with the stakeholders. But, if the participant participates for the first time or his 100th time in an event, it is a milestone, therefore popular and relevant. Other popular instances would be when something happens for the first time, a record is broken, a milestone is reached and so on. NLG system can implement Metric 1 for the content type Instance to identify such cases.

The sports experts had mixed responses regarding the applicability of Metric 1 for the Instance content type. As mentioned before, PO1 found Metric 1 applicable only for the content types Event Participant, Team and Event. SE1 stated that it is difficult to conceptualise the popularity of Instances. Thus, he didn't identified it as influencing the relevance initially. But, after discussing Instances such as milestones he agreed that the metric could influence the relevance of a sports fact. RE1 and CE1 found Metric 1 applicable for Instances only from the definition of the popularity property.

During the interviews we also noticed that sports experts confused popularity with importance in a few cases. The same situation repeated for other relevance properties. We identify three possible causes for the confusion between properties. First, during the interviewees, we initially provided the sports experts with a simple definition for popularity. Secondly, we evaluated one relevance property at a time and each sport expert reviewed six properties during the interviews. Lastly, properties cover relevance of sports facts from overlapping viewpoints. For example, a Template that is important to a sport it is also popular most of the times and the same happens for other content types. Even though relevance properties overlap, they cover their own edge cases.

Newsworthiness

In automated journalism, NLG systems try to replicate the performance of journalists as dictated by the requirements of the domain (Leppänen et al., 2017). Also, besides automatic metrics for NLG, researchers have also used professionals to evaluate the output of their NLG systems (ge Yao Jianmin Zhang Xiaojun Wan Jianguo Xiao Institute of Computer Science et al., 2017). As such, it is important to identify what journalists write about and what they consider to be of value to the reader. The property was used by ge Yao Jianmin Zhang Xiaojun Wan Jianguo Xiao Institute of Computer Science et al.

who developed an NLG system that selected those phrase and words that were often used in the news reports.

Newsworthiness is a property that is related to the news domain which is focused on identifying content that is worth reporting on by journalists. To a certain degree, it is similar to the property *popularity*. But, *popularity* is focused on popular content for the sports fans, and *newsworthiness* is measured in regards to journalists.

Metric 2: Content newsworthiness

It should measure the value of the content type to the news marked compared with other content types of the same type.

It can be measured by counting how often a content type is mentioned in news articles, with the help of sports experts or other data sources that could derive the news market value of the content type.

Newsworthiness is a property that could be updated regularly, even in real time, in an NLG system so that it keeps up with the trends. By doing so, the NLG system would generate facts that are actual and relevant. Event Participants can become relevant to a journalist due to various reasons. RE1 gave the example of a the national Dutch team player that got injured and stated that „to report about him is definitely more important than about some other player". The example illustrates how different situations could influence the relevance of an Event Participant. The newsworthiness of an Event Participant can change at a given time, which can be omitted by NLG systems. NLG system can identify which Event Participants are worth reporting at that moment by keeping track of their usage in media. Unexpected situations could also occur for the content types Team, Location and Event. Such situations increase the value for the news market of these content types and their relevance. It is difficult to keep track of the usage in media for the content types Role, Template and Instance and thus measuring their news market value. But, their newsworthiness could be measured with the help of sports experts or journalists. SE1 stated that some Template and Instance content types appear rarely in news reports because they don't occur that often in the sport. For example red cards in soccer. Therefore, the users of SOCCER framework need to take such situations into account when they implement Metric 2 for the content types Template and Instance.

The sports experts in the interviews are illustrated in B.4.2. Similar with popularity, all the sports experts unanimously considered that the newsworthiness of the Event Participant, Team and Event content types has an influence on the relevance of a SF. PO1 stated „If you want to create a news article, then you want to do so for an audience that recognizes the teams and the players involved. Otherwise, you have to make a very, very, very big story about a completely unknown sportsman or football player." . The statement illustrates B.4.1 can be used by journalists. Additionally, we notice that there are areas in which B.4.2 and B.4.1 overlap in terms of relevance. CE1 shared that the Template and the Instance are important for journalists as they could write a full article around it. SE1 didn't mark the Role, Template and the Instance originally as he thought that it is hard to measure their newsworthiness. Also, he disagreed with the measuring guideline which stated at the time that you need to count how often a content type is mentioned in the news to measure its newsworthiness. Based on the feedback from SE1, we updated the definition and the measuring guideline.

Importance

Sports have different rules and are about reaching a certain goal to win an event. The rules in a sport lead to a difference in the value of certain actions and not only. Kanerva et al. and Zhang et al. leveraged the importance of events to identify relevant content in their NLG systems. The property can be used for different entities and for different purposes. Leppänen et al. used weights to measure the importance of entities types such as locations, persons and others to select which content is relevant to a users. The weights were modifiable at run time which gave the ability to the NLG system to produce tailored and topical news reports for a user.

Importance focuses on the key points of a sport. Not every Event Participant is equally meaningful for an event, for example the players and the referees in a soccer match. While we might see some facts about the referees, the players play a more important part to the event. The situation is similar in other cases. For example, international competitions such as the Champions League have a higher degree of importance than the national leagues. The property is quite important for *Templates*. For example, when a Template involves an important action that happens in the sports such as red cards or goals in soccer. An NLG systems needs to select and filter the Templates that are about important actions of that event. For example, when selecting between a fact about a goal and a pass, the system should select the one about the goal.

Metric 3: Event Participant importance

It should measure the importance of the Event Participant to the sport, the event or the team they are part of.

It can be derived from the monetary payment the event participant receives, with the help of sports experts or other data sources that illustrate the importance of the event participant.

Metric 3 identifies the valuable Event Participants of a sports, event or a team. The metric should be implemented depending on the goal of the NLG system. For example, if the NLG system needs to generate content related to the soccer match between Arsenal and Liverpool, it is worth to identify the important Event Participants from that event and for those teams. We could determine the value of Metric 3 from the salary of an Event Participant. The players, referees and managers have different wages based on their experience and performance. Event Participant who are valued more by their team would have a higher wage. SF about the high performance Event Participants would have a higher relevance. That was opinion of sport expert CE2 who stated „The bigger the player or the team, the more important the fact is because you reach a broader audience or more of the audience is interested in it.”. Overall, all sports experts agreed that Metric 3 has an influence on the relevance of a SF.

Metric 4: Role importance

It should measure the importance of a Role to the sport, the team or the event.

It with the help of sports experts or other data sources that display the importance of a Role.

The Role content type captures the responsibilities and characteristic of the Event

Participants. It might be their position on the field in soccer, their role in the team and the sport or their nationality. Different sports have different roles and those roles have different levels of importance. By focusing on the importance of the role, an NLG system can generate content that it is of value to its users. For example, fact **CF10** has an increase relevancy because it is about the captain of soccer team Tottenham. The captain is often an experienced player but also a play maker for the team. Thus, his importance is higher compared with other roles in the team.

The sports experts expressed their low expectations for Metric 4 in the interviews as illustrated by their choices in Appendix B.4.3. Sports experts CE2 and PO2 disagreed that Metric 4 has an influence on relevance due to different reasons. CE2 remarked that Metric 3 already captures the same area of a SF's relevance as Metric 4, especially when the sports fact it is about an important player. CE2 stated that „If you make a statistic about Lionell Messi or Cristiano Ronaldo it is not about their role, it is irrelevant at that moment". His statement is similar to the one PO1 had for Metric 1 who identified the the popularity of the Role is covered by that of Event Participants when he is extremely popular. As such, the impact on relevance on the metric related to the Role content type is reduced when that of Event Participant metrics is substantial. PO2 stated that Metric 4 has a small influence on the relevance of a SF, thus it is not worth measuring. During the interviews, the sports experts had to answer with yes and no to the questions. PO3 stated that he doesn't find that all metrics have an equal influence on relevance and that is why he wouldn't consider some of the metrics. Thus, he chose to give estimated percentages for some of the metrics. He estimated that Metric 4 would maybe have only 35% influence on the relevance of a SF compared with other importance metrics. CE3 stated that the influence on relevance of Metric 4 depends on the type of statistic that is part of a fact. CE3 expressed that Metric 4 it matters for **CF1** because you don't expect that a defender would have assists. In his example, SF1 touched the unexpectedness property.

Metric 5: Team importance

It should measure the importance of the team to the sport or the event they are part of.

It can be derived from the market value of the team, with the help of sports experts or other data sources that demonstrate the importance of the team.

There are sports where athletes compete against each other such as golf, swimming and others. Other sports involve teams that compete such as soccer, baseball and so on. In soccer, 20 teams compete in Premier League. Out of the 20 teams, not all of team are equally important because some might have better players, a longer history, a larger fan base,a bigger influence on the competition or due to other reasons. All these reasons make important teams relevant to sports stakeholders. Therefore, NLG systems can implement Metric 5 to generate content related to important teams.

During the interviews, all the sports experts agreed that Metric 5 has an influence on relevance. SE1 and PO1 gave the example of how teams are important and popular nationally but not internationally. PO2 considered that only Metric 6 has a higher impact on relevance than Metric 5. PO3 stated that the team could also be an Event Participant, thus Metric 5 would influence the relevance similarly to Metric 3. Initially, CE3 confused the relevance metrics scope. He thought that the metrics would measure if a content type should be part of a SF. Thus, we offered additional explanations regarding the meaning of

the metrics. In the end, CE3 stated that „So, in general, I think that high level teams will be more important and have more relevant than lower-level teams, but it all depends on the user of the facts". The statement brings up the subjectiveness of the metrics. Relevance is dependant on the users of the NLG system to a certain degree.

Metric 6: Event importance

It should measure the importance of the event to the sport compared with other similar events.

The importance of an event could be derived from the monetary value of the event, with the help of sports experts, from the structure of the competition or from other sources that deliver data on the importance of an event.

In a sport, it is common to have different type of events. A tournament is an event where event participants or teams compete against each other in an establish format. Thus, the tournament is comprised of several phases, each with a different degree of importance to the whole tournament. As such, the final of the tournament has a higher degree of importance than a preliminary round. Also, the tournament itself has a certain degree of importance when we compare it with other tournaments. For example, in soccer, there are the tournaments UEFA Champions League and UEFA Europa League. Champions League is a top tier tournament where only the best teams of a country can participate. Europa League is a second tier tournament where the best of the remaining compete. The difference is illustrated also by their monetary value where the total earnings for a team in Europa League could be a maximum of 21.34€millions while for Champions League is 82.45€millions €(*Infographic: The Financial Mismatch in European Cup Football*, z. j.).

The sports experts unanimously agreed that Metric 6 has an influence on relevance as showcased by their choices in Appendix B.4.3. PO2 considered that Metric 6 has the highest influence between the importance based metrics. We found it easy to explain Metric 6 in terms of relevance. CE2 identified importance similarly where he gave the example of Premier League as an important event compared with others.

Metric 7: Template importance

It should measure the importance of the sports action part of the Template to the sport.

It can be measured with the help of sports experts.

The Template content types captures the core of a SF and most templates describe an action that happens in the sport. For example **CF1** refers to the action of assisting in a goal while **CF9** is focused on the number of ball touches. From the perspective of the sport, it would be more important to assist a goal than touching the ball. Thus, facts that contain important action should have a higher relevance.

Originally, we used the name *Action* for the content type Template. But, it confused the sports experts as they understood it to be only an action that happens in the sport and not the whole query behind the SF. Also, the Template content type was confused at the time with the Instance content type. Therefore, at times we had to provide additional examples to the sports experts. PO2 and CE3 marked that Metric 7 has an influence on relevance after seeing only the definition of the property. PO2 and CE2 agreed that Metric 7 has an influence on relevance after a short discussion. With PO2 we also discussed

the impact the Role content type for a Template, in particular goals in soccer scored by players with different roles. We came to the conclusion that important Templates are not influenced by the Instance or other content types as PO2 stated „... a defender that scores is of course more valuable". Metric 7 came up in interview with SE1. SE1 believes that several cases discussed for Metric 2 could be covered by Metric 7. For example, he mentioned how red cards in soccer can be identified using Metric 7 instead of Metric 2.

Metric 8: Instance importance

It should measure the importance of a Instance for sports stakeholders compared with Instances part of the same Template.

It can be measured with the help of the sports stakeholders or other data that proves the importance of an Instance.

A large amount of Instances could be part of a Template. For example, the template *X player has played Y matches for team Z* could be filled for each player of a team, for each game they played for that given team. But, it would be a milestone for the 25th, 50th or 100th match.We used the same example previously for Metric 1. Milestones are not only popular but important too for sports stakeholders. Additionally, Metric 8 can be used to identify results which are important but not popular. CE2 gave a similar example when he motivated his choices during the interview. Overall, all sports experts identified that Metric 8 has an influence on the relevance of a SF.

Significance

During the literature review, we identified that NLG systems use significance in different ways to determine the relevance of content. Leppänen et al. calculated the newsworthiness of facts by identifying out liners in data. ge Yao Jianmin Zhang Xiaojun Wan Jianguo Xiao Institute of Computer Science et al. deemed phrases that happen close to key changes in the game as relevance. Sports generate large amounts of data which will uncover out liners, either the best performing athletes or the worst one. In both, cases, sports fans would find relevant to read content about it.

Significance is a property that focuses on the surrounding context of a piece of content. By context, we mean the given situation at the moment related to the fact for the content type. For example, Liverpool would play multiple matches during the group stage of Champions League. Lets assume that the last match is deciding if the team moves forward in the competition. While all group stage matches are equally important to Liverpool, the last match has a higher significance as it is influencing the future of the team in the tournament. We cannot define a context for all content types. As such, this property can be measure and applied only to content of type Event, Template and Instance.

Metric 9: Event significance

It should measure the influence an event has for the sports event participants or the teams involved in it considering their situation at the moment.

It can be measured by identifying and quantifying the impact of the results from the sports event to its participants or teams using the related data.

Metric 9 can be used in NLG systems to identify Events that have a significance for the

Event Participants or the Teams. It is not only deciding Events that could be significant. Event Participants or the Teams participating in the event could have a long lasting rivalry. As such, the matches between them have an increase significance for them but also for their fans. NLG systems that implement Metric 9 can identify SFs related to rivalries. Additionally, Metric 9 can be used to generate a content reel of the important Events in a tournament for the Teams or Event Participants.

Most of the sports experts had different miss conceptions about the metrics related to significance. During the interviews, the sports experts first saw the description of the significance property: „how crucial is the content type given the context". The sports experts found it difficult to understand the meaning of *context*. But, after further discussions in the interview, all sports experts agreed that Metric 9 has an influence on the relevance of a SF as illustrated by their choices in Appendix B.4.4. We modified the definition of the related significance metrics using the feedback from the interviews.

Metric 10: Template significance

It should measure the influence of the sports action part of the Template ,compared to similar sports actions, to the outcome of an event taking into account the situation at that moment.

It can be measured using sports data and an algorithm that quantifies the sports action impact to the outcome of an event.

During a sports event, the same action could happen multiple times, thus an NLG system can generate multiple SF that describe them. For example, multiple goals could be scored during a soccer match. All of them would be equally important but, the deciding goal would have a higher significance. The same logic can be applied to other actions which are captured by the content type Template. As such, Metric 10 can be used to identify SF that had an impact and are of interest to the stakeholders.

As with Metric 9, the sports experts had different misconceptions about Metric 10. CE2 confused Metric 10 with Metric 6 as he gave the example of how winning a final is significant. PO2 identified how an injury becomes significant depending the athlete that it occurred to. For RE1, Metric 10 would happen when a team is down 2 goals and it would influence the mind state of the players. In the second part of the interviews related to significance, the sports experts got acquainted with additional examples. After this, they all agreed that Metric 10 has an influence on the relevance of a SF.

Metric 11: Instance significance

It should quantify the statistical differences between Instances that are part of the same Template.

It can be measured using regular algorithms from statistics such as average, standard deviation and so on.

The Instance content types describes the data behind a Template but also the results. For example in the Template *scoring X goal during Y matches*, X describes the result and Y the data behind the template. It is of higher significance to score 10 goals in 10 matches than scoring 10 goals in 20 matches or 5 goals in 20 matches. Metric 11 plays an important role in measuring the relevance of a SF as fans are impressed by the numbers. Also, Metric 11 can be used to identify significant performance of players by comparing their results

with the others. Also, the data can identify out liners in the data as done by (Leppänen et al., 2017). Metric 11 is one of the most important SOCCER metrics as it captures how impressive are the results described by a SF. Journalists are interested in results that are impressive or out liners in the data as they can write a news article around them. Sports fans want to know who was the best or who was the worst. At the same time, gamblers need similar information to increase their success rate. Metric 11 measures and quantifies which data has out liners, superior or inferior results.

Metric 11 measures the significance of the numbers that are part of a SF. The sports experts identified such situations themselves based on the description of the significance property. PO1 have the example of breaking records. RE1 gave the example of high values in ball possession in soccer. For CE2, SFs are about significant Instances and Templates, thus he identified from the start that Metric 10 and Metric 11 have an influence on the relevance of a SF.

Unexpectedness

Unexpectedness is a subjective measure of a pattern interestigness in the data mining field. A pattern is unexpected, and thus interesting, when it contradicts what a user knows or believes about the data (Kontonasios et al., 2012). Sports stakeholders expect that teams and athletes that perform well will continue to do so in the next events. But, sometimes the unexpected happens and they lose. Such is the case in SF **CF9** where it is unexpected that the captain of the team touched the ball only 11 times. Unexpected SF deliver new and contradicting information to sports stakeholders. Such SF are interesting for sports fans, helpful and valuable to journalists and gamblers.

Metric 12: Template unexpectedness

It should measure the level of unexpectedness a Template can create for a sports fan.

It can be measured with the help of sports experts or data that can identify the level of unexpectedness a Template can create for a sports fan.

The content type Template represents the data query behind a SF and its core structure. Thus, the Template can be general such as *goals as X role in Y matches* or specific such as *goals as defender in Y matches*. In the second case, we can attribute a level of unexpectedness to the Template as it is unexpected that a defender scores goals. Sometimes, the data query can be made to identify specific unexpected situations such as *top ranked team losing against a low ranked team* so that NLG system would create SF that surprise the users. Sports experts could give measurements for Metric 12 so that the NLG system identifies SF that surprise its users.

Metric 12 was not part of the first version of SOCCER. It was added after analysing the feedback from the sports experts. During the interviews, we debated several unexpected situations that can happen in sports and how could they be identified with the help of the SOCCER framework. CE1 stated that Event Participants and Teams could do something unexpected and thus he would measure unexpectedness for them. CE1 marked the Template and the Instance content types because she found **CF1** unexpected due to the role and the action. Also, CE1 identified that the Event could also be unexpected since a Team can do well in international events such as Premier League but under perform in national events. NLG system can detect such situations using Metric 12 and Metric 10.

CE2 and PO2 both gave the example of when a top team loses to a team that is in the bottom of the rankings, with PO2 stating „But for me, for example if Ajax loses against the number last of the ranking, that is very unexpected". RE1 also stated that it can be unexpected to lose to certain teams. The sports experts uncovered how unexpected situations could happen in sports and those situations could be identified with the help of the Template content type. Thus we created Metric 12 to assist NLG systems to identify SF that describe surprising events.

Metric 13: Instance unexpectedness

It should measure the level of unexpectedness an Instance could create for a sports fan.

It can be measured using data analysis on the patterns in the data or by other methods that derive the unexpectedness of an Instance content type.

Even though the Template content type can capture specific unexpected situations that occur in sports, some of them could also be derived and captured using Instance content type. Previously, we gave the example of a specific template content types *top ranked team losing to bottom ranked team.* The template could also be identified more generically as *X team losing to Y team.* In the generic case, unexpectedness would be derived from the data if one of the team is ranked last and one first. Therefore, an NLG system can measure the unexpectedness of a SF using the data. Metric 13 can identify unexpected results for different athletes and teams. RE1 mentioned that it would be unexpected for a soccer team to receive 10 goals in an official tournament match. Metric 13 is similar to Metric 11 as it deals with out liners in data, but it should measure when those out liners surprise the stakeholders.

Complexity

Sports contain different rules and have objectives which lead to complex situations which are explained using SFs. Such SFs are complicated therefore hard to understand by some of the users of an NLG system. Also, regular sports fans might not know all the rules of a sport. Thus, complex SF would be less relevant to the sports fans. Additionally, the text produced by an NLG system might have to be placed in a limited space Lampouras en Androutsopoulos. News websites and written press have a limited space to place their content. Betting website present their offer and have even less space to present SF to gamblers. But, complex facts require additional words to describe the targeted situation. Complexity metrics can be implemented in NLG systems that target a wider audience of sports fans or to limit the size of the produced text.

Metric 14: Template complexity

It should measure the level of sports knowledge is required to understand the Template or its structural complexity.

It can be measured with the help of sports experts or other methods that would illustrate the complexity of different Templates.

Templates can be easy to understand or complex when they require a person to know the rules of a sport. For example, in soccer, a simple Template would be *scoring X goals.*

A person needs to understand the meaning of goals. But, the sport has other rules, for example offsides. An offside happens when a player has any part of his body, except the hand and arms, in the opposing team side of the pitch than the last opponent except the goalkeeper. A fact that involves a Template related to offsides has a higher complexity that one about goals. Additionally, facts could be complex due to their structure. For example, we can transform the fact about goal scoring into *scoring X goal in Y day of week while Z weather*. As we added a new dimension to the fact, its complexity grew. The example can be extended with other dimensions which would further increase its complexity. Thus, the fact becomes harder to understand by the fan, therefore less relevant.

Metric 14 wasn't part of the initial version of SOCCER. But, PO3 illustrated its necessity when he stated „This ("number of passes over a certain period of time") is pretty complex. For example, I can make it more complex. The number of forward passes given in the opposite half, in the second half between 75 and 89 (minutes). For me this is an extremely complex action.". CE2 and CE3 also came up with a few examples of complex Templates. We presented the example PO2 in the follow up interview with the other sports experts and they all agreed that Metric 14 has an influence on the relevance of a SF.

Metric 15: Instance complexity

It should measure the level of sports knowledge the user of the sports facts needs to understand an Instance.

It can be measured with the help of sports experts or other methods that would illustrate the complexity of different content of type Instance.

The Instance content type covers the data and the result behind a SF. While the result is simple at time, the data behind that result could be complex. The data could exclude and include only certain competitions or it can cover different ranges of time. For example, **CF1** mentions that that the fact is related only to the English league while **CF8** is related to all competition in which the teams play. A fan needs to know the other competitions the teams are part of to understand the result.

CE2 didn't agree that the content type Instance can be complex as he looked only at the result part. CE3 also had the same opinion at first, but after we elaborated on a few complex situations that CE3 agreed that the Instance content type could be complex. At the time of the interviews, we used the term Statistic instead of Instance. Also, the examples in the interview illustrated mostly that Instance content type captures the result. Thus, we had to further accentuate that the Instance content type captures also the data behind a SF. PO3 and SE1 identified from the start that Metric 15 has an influence on the relevance of a SF. CE2, CE3 and SE1 marked that the Event content type might be complex due to having a complex structure or being unknown. After further discussion, the sports experts changed their opinion. SE1 identified that it is the data that is complex and not the event itself. SE1 also marked the Role content type for the complexity property and gave the example of role in rugby. In the end SE1 changed his opinion as rugby find should know the roles involve in the sport.

NLG system can use Metric 14 and Metric 15 to tailor their output to different categories of sports fans based on they involvement and knowledge of the sports. The generated output would be then easy to understand or complex enough to be relevant to the final users.

Sentiment

Sports fans invest energy and feelings towards the sports itself, the teams and the athletes they support. Their attitude could vary depending on the performance of the team, athlete or competitors. Lee et al. developed an NLG system that generated game reports at the end of a match that were tailored to the fans of the two soccer teams. Thus, for the targeted audience, the system would use a cheer full tone if the team won or a more disappointing and frustrated tone in the case it lost. Other sports stakeholders are also sentimentally invested. Journalists would use different attitudes and tones in their articles and thus they would focus on subjective characteristics of an athlete Kanerva et al.. For example, the journalist could emphasize that a player did not perform as expected and state „Player A did not manage to score more than one goal". Other situations could occur in sports where sports stakeholders are sentimentally involved such as rivalries between certain athletes or teams. We devised the *sentiment* property for NLG systems to captures and identify emotionally invested content.

Metric 16: Event Participant, Team, Event, Action sentiment

It should measure the attitude and its intensity sports fans have towards the content type.

It can be measured with the help of sports experts, sentiment analysis of data or other means that would compute the attitude of stakeholders towards a SF.

Sports stakeholder could have a certain attitude with a given intensity towards content. Sports fans can have favourite athletes or teams, enjoy reading about certain type of actions, results or events. Sports fans would often play the sport themselves and have a favourite role they enjoy, thus making content related to that role relevant to them. NLG can implement Metric 16 to identify and tailor the content for sports fans based on their passions. For example, the NLG could focus the generated content on events which involve a long time rival of a team. In soccer, such an event is called a derby and it can increase the demand of content (Tyler, Morehead, Cobbs & DeSchriver, 2017). The derby can be measured and identified using geographical markers, specifically by research or sports experts, by attendance number or other means.

The sports experts expressed different attitudes towards Metric 16 depending on the content type. All sports experts identified that the Metric 16 should be measured for the Event Participant and the Team as sport fans would be highly interested in content related to their favourite teams and athletes. Metric 16 was overlooked by the sports experts for the content type Event. PO3 only marked the content type Event for Metric 16 because some fans might enjoy Champions League or the World Cup more due to nationalism sentiments. In the second part we exemplified the situations of derby matches and how they can be identified by measuring Metric 16 for the content type Event and all sports experts agreed that the metric has an influence on relevance in this case. Except PO3, the other sports experts identified just from the description of the sentiment property that Metric 16 should be measured for the content type Template. Also, PO3 also agreed that Metric 16 should be measured for the content type Template after debating a few examples in the second part of the debate. CE3 stated that sports fans would have an attitude towards the content type Instance if it is positive or negative about their favourite team. RE1 had a similar idea and stated „if your team scores a goal it makes you happy,

if the opposing team scores a goal it makes you sad". The sports experts found it difficult to grasp the utility of Metric 16 for the content type Location. RE1 and CE3 gave the example of how the home stadium matters for fans. But, in the follow up discussions we discussed sports with multiple stages such as auto racing or cycling. In sports as cycling, fans enjoy certain regions of the race such as the mountains ascents in Tour De France. After this, except PO3, the sports experts agreed that Metric 16 should be measured for the content type Location.

Initially, Metric 16 included also the content type Role but we removed due to the feedback we received in the interviews. The sports experts stated that sports fans would not have a certain attitude towards certain Roles, maybe for referees in soccer. Also, the sports fans would mostly care about the Team and the Event Participant as PO3 stated „If you're talking sentiment in the true sense of the word, I would say it's really about the event participant and team that's you know, nobody is sentimental about the striker or the defender".

In some cases, it can be a challenge to measure Metric 16 in an NLG system. But, the results can increase the capability of the system to produce content which is targeted to the beliefs and attitudes of the users. Thus, the content would have a higher relevance to them and increase their engagement. Metric 16 should be adapted according to the requirements but also the limits of the NLG system.

Timeliness

Sports events repeat over the course of time. For example, the Olympics event takes place every 4 years while national tournaments in soccer repeat yearly. Time plays a crucial role for the news domain as the story needs to follow a chronological order (Kanerva et al., 2019). As time passes, events fade out and their related content becomes less relevant. Also, during the course of an event certain actions happen over time, some having a larger impact on the outcome than others. ge Yao Jianmin Zhang Xiaojun Wan Jianguo Xiao Institute of Computer Science et al. valued content from commentary texts which was produced around key changes in the match.

Timeliness is a property that captures the relevance of different entities from a time perspective. Event participants will change teams, retire, thus become irrelevant for sports stakeholders. Events also get less relevant as time goes on unless the NLG system tries to generate facts for an old event. For the Instance content type, *timeliness* has a similar meaning as for other pieces of content. Certain Instances might capture data about past events which are not relevant to the current period. The *timeliness* property has the goal to identify the situation where a content type becomes more or less relevant due to the flow of time. NLG system might want to generate content related to current events or past ones. NLG system would generate content for a certain time range which we will refer to as *target time period*.

Metric 17: Event participant timeliness

It should measure how active was the Event Participant in the targeted time period by the NLG system.

It can be measured using data that shows when a player was active in the sport.

In most sports, athletes have a short career due to the physical requirements. Other

athletes retired due to injuries or other reasons. SF that involve retired athletes become less relevant to the sports fans. Metric 17 would assist the NLG system to identify SF that involve relevant athletes.

The sports experts found it difficult to understand the scope of the property during the interviews, in particular with the meaning of the *target time period*. RE1 understood the meaning of the property easier which we attribute to his background as a researcher. To facilitate the debate, we gave additional examples to the sports experts in the first phase. CE1 stated that Metric 17 can be used to identify when a player retired and thus it would fit the definition. We presented similar examples to the other sports experts, after which they agreed that Metric 17 has an influence on relevance.

Metric 18: Event timeliness

It should measure the time span between the moment the event takes places and the target time period of the NLG system.

It can be measured with data that identifies the time when the event happened or will happen.

Certain events, such as Champions League finals happen every year. **CF1** to **CF10** in Sect. 4.2 are related to the Champions League final, season 2018/2019. The SFs were relevant to the sports stakeholders at that time. But, as time passes on they become less relevant. Metric 18 captures situations as this that would influence the relevance of a SF.

As sports experts found it difficult to understand the meaning of the *timeliness* property, we used an examples related to the content type Event to illustrate its meaning. Thus, most sports experts then identified that Metric 18 would have an influence on relevance. After discussing Metric 18, PO2 stated that he has a finer understanding on what is relevance and suggested „you could almost add time as a content type in general". PO2 brought up an option that we didn't consider before, that of using time as a content type. We investigated the option and found that a few properties can be measured for it as a content type such as popularity or sentiment. But, it wouldn't be useful to measure for other properties such as predictability or novelty. Therefore, we opted to not follow PO2's suggestion but keep it as a venue for further research.

Metric 19: Instance timeliness

It would measure the relevance of the time period covered by the data and the results to the time period targeted by an NLG system.

It can be measured by identifying the time difference between the time period covered by the data and the result to the time period targeted by the NLG system.

SF are derived from data that covers different lengths of time. For the next examples, lets take into account that an NLG system needs to generate SF related to the present. Then, a SF about goals scored in the last year is more relevant than a fact about the goals scored two or more years ago. The sports experts found it hard to understand the meaning of Metric 19 as they disregarded the data that is part of the Instance content type. But, all agreed that Metric 19 has an influence on relevance to a certain degree after further discussions. PO1 was sceptic about Metric 19 as he considered that the time dimension of the data is covered by the Event content type. That was the case of a few examples we debated, but the data behind a SF might extend behind the time range described by

the Event. The data behind a SF can span different lengths of time which can influence its relevance for sports fans, gamblers or journalists. While sports fans or journalists are interested in historical facts, gamblers would find those facts irrelevant. NLG system can use Metric 19 to produce facts which cover data from a time period that is relevant to its users.

Timeliness was one of the properties that the sports experts found most difficult to understand. But, they understood the property and offered valuable feedback back after receiving additional examples. PO2 suggested to make *timeliness* into a content type. PO3 stated that we should measure timeless for the content type Team as it contains the same characteristics as the content type Event Participant. Teams are similar with athletes for certain properties. But, they are almost static from a time perspective compared with the short career of athletes. Metric 17 captures the flow of time for athletes who retire which reduces their relevance. Metric 18 measures the relevance of sports events as they repeat over time. Metric 19 measures the relevance of the data based on the period it captures. NLG system can implement Metric 17, Metric 18 and Metric 19 to deliver facts that deliver actual content to its users.

Novelty

Novelty is a subjective measure of the interestingness of association rules used in data mining. Novelty is a characteristic of information which the user has not seen before or could not have inferred it from another source in data mining (Kontonasios et al., 2012). As with the users of rules data mining systems, sports fans would find interesting SF which they don't hear that often or are just new for them.

Metric 20: Template novelty

It should measure the possibility of a Template being novel for sports fans compared with the other Templates.

It can be measured with the help of sports experts or with data that identifies when a Template was created and how often was delivered to sports fans.

Sports have rigid rules that rarely change. Also the same events, data and output repeats over time. That is one of the reasons NLG system are employed to automate match reports. Such reports become dull to the users as the same type of information repeats itself. But, new types of statistics are developed in sports, for example sabermetrics in baseball. New statistics lead to the creation of new Template content types that describe the SF related to them. Metric 20 can be used to identify new Template content types to create novel SF that would create interest to a sports fan. As a Template is used by an NLG system it would become less novel to its users and thus, the implementation of Metric 20 needs to account for this.

Metric 20 was created after the interview with the sports experts based on their feedback. PO3 opinion was that novel facts are those which contain creative stats as he stated „... everybody knows like all these boring goal stats. You couldn't care less. You care about creative stats. And, creative stats come with novel actions, I would say". SE1 and CE1 stated that it is the combination of Template and Instance which creates the novelty in a SF. Although SE1 first marked the content type Template for novelty measurement, in the second part he doubted his choice based on the examples part of the interview. SE1

agreed in the debate that Metric 20 can be used to identify fun facts.

Metric 21: Instance novelty

It should measure the possibility of an Instance being novel for a sports fan compared to Instances of the same Template.

It can be measured using statistical algorithms on data that depicts the usage rate of Instances for a Template.

A diverse range of Instances can fill a Template. Some of those Instances are common for a given Template. For example, it can happen often that a player scores a goal or none but rarely that he scores 5 goals in a match. Also, an Instance can occur for the first time. For example, the first defender to score 5 goals in a match. Similar situations take place for most Templates and Instances content types. Instances that appear often are well known by the sports fans and journalists, thus they are duller. Therefore, they are less relevant due to the low level of interestingness. NLG systems can track which Instance content types they have used for different Templates and give a higher relevance score to those which are used rarely.

The sports experts were mostly positive about Metric 21 as illustrated in their choices in Appendix B.4.9. Although CE1 didn't choose to mark Metric 21 in the first part of the debate regarding the novelty property, he did mention that it is a combination of action and statistic that creates novelty. In the second part of the interview, we discussed further instances when something happens for the first time or rarely after which he agreed that Metric 21 has an influence on relevance. PO1 identified only Metric 21 for the novelty property because „something novel has to be the first ever". SE1 found the property difficult to understand and asked for further clarifications during the interview. After which, he tried to inquire if it is novel when someone achieves something for the first time.

NLG system can implement Metric 20 and Metric 21 to generate content about events that are interesting to its users due to its novelty. Standard, machine like reports become dull and thus irrelevant to the audience. Additionally, the novelty metrics can assist NLG systems to add a certain level of objectiveness to their content by identifying the rare situations or are new to its users.

Utility

Sports stakeholders have certain goals that they try to reach through sports content. Sports fans want to be entertained, read about their favourite team or player and to keep up to date. Journalists search for content that can help them write news articles. Gamblers are looking for information that can help them place a successful bet. Utility is defined as a property that can help a user reach a certain goal in the data mining field (Shaharanee, 2012). NLG systems can have a wide scope but it is more often that they are fine tuned for certain tasks. For example, the NLG could have the goal to deliver content for news domain or for betting domain. In these cases, it can broadly identify the goal of its users. As such, it can use relevance metrics based on *utility* metric to produce content that will help its users reach that goal.

Metric 22: Content type utility

It should quantify the value the content type gives in assisting the sports stakeholder to reach his goal.

It can be measured by identifying the goal of the user of an NLG system or the system itself and quantifying the utility of the content type for achieving that goal.

Metric 22 can be used in NLG system which have a defined goal or can identify the users of its content. For example, an NLG system can specifically produce SF for the betting domain to assist gamblers in placing their bets. If the gamble happens on a website, then the system could receive data on which team, athlete or event the gambler wants to place a bet. Thus, the NLG system can use Metric 22 to specifically generate content related to that team, athlete or event or betting types. In the news market, an NLG system could have the goal to generate news articles on certain events. Then, the end user are sports fans who are interested in that event and its teams. Therefore, the NLG system can use Metric 22 to specifically generate content related to that event including its participants, teams, location and roles. NLG systems can use Metric 22 to create specific content that is suited to its goals or the needs of its users.

Initially, we included only the content type Instance in Metric 22 as we assumed that it can measure the entertainment and informational value a SF holds for a sports fan or a gambler. But, the sports experts brought compelling arguments to include other content types. All the sports experts identified that the Template content type can also have utility for gamblers when they place a bet. For example, when you bet on who will win there will be a group of Templates that can give you useful information and a group that will not. CE3 stated that „all of these content types are at some point relevant (of utility) to the different stakeholders". CE2 also identified that utility should be measured for the content types Event Participant and Team as fans and gamblers want to read about certain Teams or Event Participants. PO2 found it hard to imagine the goal of a newsreader but he brought the argument on how gamblers want to see Templates and Events that can help him place a successful bet. PO2 also identified that Metric 22 should be measured for the content types Event Participant and Team as newsreaders have an interest in particular teams and athletes. But, he wasn't sure on how to measure or identify the interest of newsreaders. We decided to include all content types for the *utility* property based on the sports expert choices which are illustrated in Appendix B.4.10.

Peculiarity

Data mining system can discover a large amount of associations rules and thus different methods have been devised to reduce the irrelevant ones through pre-processing and post-processing the data (Miani & Junior, 2018). The rules discovered by a data mining system can be similar when you define a distance function for them. It is time consuming and uninteresting to read the same information twice. Peculiarity is a property in data mining that is used to identify patterns which are different from others according to a specified distance measurement (Kontonasios et al., 2012). Similarly, NLG systems which are template or end-to-end based could generate facts that are similar in a given situation. The similarity of facts is illustrated mostly through the content types Template and Instance. For example, it can become dull to a soccer fan to read a news article filled with SF about

pass accuracy. The *peculiarity* property can be applied to collections of SF to identify the ones that are unique or different from others.

Metric 23: Template peculiarity

It should measure the similarity of a given Template in a collection compared with the other Template types.

It can be measured using a defined distance function or another method that can compute the similarity of a given Template to other Templates from a collection.

An NLG system can generate SFs using multiple variations of a Template. For example, the NLG system could generate various facts regarding goal scoring which would be described by the Templates *scoring goals on X day*, *scoring goals from X location* and so on. All Templates regarding goals have a degree of similarity between them. As in the data mining field, the content can become dull for users when it is repeating itself. NLG systems can implement Metric 23 to diversify their output.

The initial version of SOCCER didn't include Metric 23. Initially, we defined generic Template types that covered a wide range of SF such as *score goals* which included all SF about scoring goals. Therefore, it was the Instance content type that could become common in a collection of SF. But, the sports experts identified the issue and gave compelling arguments to include Metric 23. All the sports experts chose that Metric 23 has an influence on relevance as illustrated in Appendix B.4.11. The Template is the content type that creates diversity in the content and in „what you measure" as stated by CE3. Also, the Template content type creates the creativity as noticed by CE1 who stated „with the action (Template), I think you can be as creative as possible and this makes the content different from other statistics". NLG systems can use Metric 23 to deliver a diversified content to its users.

Metric 24: Instance peculiarity

It should measure the frequency of a given Instance type in a collection of SF for a given Template.

It can be measured by counting how many times an Instance content type appears in a given collection of sports facts.

Metric 24 has the same goal as Metric 23, to increase the diversity of the content generated by NLG systems. Metric 24 can be used especially in the case of generic Templates that can cover a range of SF but which can have similar results. For example, for the soccer Template *X role scoring a goal* can repeat several times for the role attacker or midfielder but rarely for defender or goalkeeper. Similar situations could happen for other Templates and Instances. NLG systems can use Metric 24 to measure the diversity in their output based on the pecularity of the data and the results involved in the SFs.

CE1, CE3 and PO3 stated that it is a combination of Template and Instance that can create diversity for a collection SF. CE3 opinion was that Metric 23 has a higher influence on relevance than Metric 24 as „(Template) is more important than statistic(Instance) in this case(peculiarity)". PO1 didn't mark Metric 24 in the first part of the interview but he agreed in the second part after we provided additional examples and arguments. PO1 motivated his choice with the example used in the interview, where the role created the diversity for him in the case of **CF1**. PO1 also thought of peculiarity as identifying outliers

and thus confusing Metric 24 with Metric 11. Additionally, we can argue that is the data and the result that delivers the novelty in the case of **CF1** as the data covers assists for defenders and the result doesn't appear often for that given Template.

Metric 23 and Metric 24 can be implemented in NLG systems to increase the diversity of the content they deliver. They can be measured using standard frequency algorithms, distance based metrics or other methods that can asses the similarities between Templates or Instances.

Predictability

The sports analytics field employs different measurements to evaluate the performance of athletes but also to get an insight on their future performance. Betting on sports is one of the most popular forms of gambling (van Wijk, 2012), therefore researchers have employed different methods to predict the outcome of matches. van Wijk used machine learning techniques and predicted the outcome of SOCCER matches with an accuracy of 53%. Bhattacharjee en Talukdar applied similar methods to predict the outcome of cricket matches in the Twenty20 format. Sports fans, journalists and other sports stakeholders also have an interested in predictive content. Sports statistician developed several predictive statistics such as expected goals in soccer. The *predictability* property can be used by NLG systems in the news and betting markets to increase the relevance of the content that is delivered to gamblers and sports fans.

Metric 25: Template predictability

It should measure the ability of a Template to predict the future performance in the sport for an Event Participant, Team or a given target.

It can be measured using data analysis on the historical data related to that Template type, sports experts or other data sources that can measure the predictability of a Template.

Goals are the actions that contribute the most to the outcome of an match in soccer. The team that will score the most goals wins the match. But, during the course of a soccer match other actions will take place such as passing, assists, red and yellow cards, missed shots and so on. The sports actions lead to the creation of different Templates such as *X pass accuracy*, *X ball possession*. The templates about goal scoring would have the highest predictive value. But, the outcome of a match can be linked to other Template types such as *X ball possession*. A similar approach can be devised for other sports. Thus, NLG systems can use Metric 25 to identify the Templates that can be correlated with the outcome of an event. It can happen that Templates that have predictive value are also important. Thus, Metric 25 and Metric 7 can cover overlapping sections on the relevance of a SF. But, Metric 25 should track and identify the ability of a Template to predict its outcome ignoring its importance. SE1 only marked Metric 25 for predictability as he found that the Template *assist as a defender* can predict future performance.

Initially, Metric 25 wasn't part of the SOCCER framework as sport statisticians use data and results to do their predictions. Most of the sports experts had a similar opinion as it is illustrated by their choices in Appendix B.4.12. But, the sports experts recognized that the Template content type can also be linked to the *predictability* property because it describes *what you measure* as stated by CE3. CE2 has a similar opinion to CE3 as he

stated that the Template and the Instance content type are used by betting companies to predict when a team is performing well. SE1 stated Template itself can predict the future performance as he stated *assists as a defender ever is a predictable thing (Template) from experience*. Metric 25 can be used by NLG systems in situations where the Template itself can be linked to future performance. Additionally, Metric 25 can be used as a control factor for Metric 26 in relevance measurement heuristics as the two metrics are linked to each other.

Metric 26: Instance predictability

It should measure the ability of an Instance type to predict the future performance in the sport for an Event Participant, Team or a given subject compared with Instances of the same Template.

It can be measured using data analysis on the historical data related to Instances of a Template.

Sports fans, journalists and gamblers find interesting and valuable sports predictions due to their own reasons. Sports fans want to learn that their team will win. Journalists can generate content around such facts. Gamblers want to learn which team will win an event so that they can place a bet. Sports predictions are enticing to all stakeholders. The Instance content type describe the data and the result behind SF. The data can be analysed to identify which results correlate with success in sports. The NLG system can implement statistical analysis algorithms that determine which results correlate with the success or failure of a team and deliver that data to the user. The sports analytics field already employs algorithms to compute predictive statistics such as expected attempts in basketball (Franks, Miller, Bornn & Goldsberry, 2015), expected goals in hockey (Macdonald, 2012) and soccer(Spearman, 2018). Predictive statistics have gained popularity as they are visible on a regular basis in sports websites. Although the predictive statistics such as *expected goals* fall under Metric 25, their accuracy would fall under Metric 26. Also, Metric 26 can be computed for regular statistics. For example, the NLG system can compute win rate for teams that score 1, 2 or 3 goals in a soccer match and selected the SF that correlate with a higher win rate. The identified SF are then relevant to sports stakeholders due to their predictive value.

During the literature review we identified the interest in sports analytics for predictive statistics, therefore we included the *predictability* property in SOCCER. We discussed possible relevance metrics with the sports experts related the *predictability* property. CE2 stated that you can expect certain teams to perform well due to their history. He then marked Metric 26 for *predictability* as he found that the predictability come from the results of that team. CE3 marked Metric 26 as he found that the result, if it is high, low or average offers *predictability* for a SF. PO2 had a similar opinion to CE2 and CE3 as he found that the result is the one that gives you the feeling if a team will win again. Initially, SE1 only marked Metric 25 for *predictability*. After further discussion, SE1 agreed that Metric 26 also creates predictability for a SF but in linkage with the Metric 25.

4.4 SOCCER in practice

The SOCCER framework divides a SF into seven general types of content and links relevance to twelve properties that can be assigned to one or more of the content types. Each

Correlation Coefficient	Interpretation
± 1	Perfect
± 0.7 - 0.9	Strong
± 0.4 - 0.6	Moderate
± 0.1 - 0.3	Weak
0	None

Table 4.2 – Spearman rank correlation coefficient interpretation (Akoglu, 2018)

property impacts the relevance on the content type differently depending on the content type. We defined twenty six measuring guidelines which describe what and how you can quantify the impact on relevance of a given property and its respective content type. Table 4.1 contains an overview of the content types, properties and the relevance metrics. We developed the framework following the research method described in Chapter 2. We evaluated the ability of the SOCCER framework to guide the relevance measurement process for SF by executing a case study at Gracenote Sports company.

Gracenote Sports company developed the Omega project, a template based NLG system. The Omega project will generate a set of facts for a soccer match based on current and previous results of the teams involved in the match. Currently, the project measures the relevance of a SF by quantifying the importance of the template and the significance of the results. Therefore, the current relevance measurement is done according to Metric 7 and Metric 11. Currently, several employees of Gracenote Sports company select and filter the facts generated by the Omega project. Gracenote Sports company is looking to improve the relevance measurements done by the Omega project. Our goal is to evaluate the applicability of the SOCCER framework in practice. Thus, we used metrics from the SOCCER framework to measure the relevance of the facts generated by the Omega project. We have to mention here that the initial research project was started at the proposal of Gracenote Sports company.

The relevance metrics cover different areas and situations of relevance SFs. In some cases, the relevance metrics could cover the same areas of relevance. For example, a popular content type could also be newsworthy. We used a Likert scale based evaluation that placed the relevance ratings and several relevance metrics into ordered categories according to their measurement. The Spearman rank correlation coefficient can be used to measured the strength and the direction of an association between two variables that are ordered (Schober, Boer & Schwarte, 2018). Therefore, we used the coefficient to draw tentative conclusions about the relation between different metrics and relevance itself. We will interpret the correlation coefficient as illustrated in Table. 4.2.

4.4.1 Metrics implementation

First, we evaluated the applicability of different metrics to derive the relevance of the facts generated by the Omega project. We took into consideration available data, the goal of the case study, the available time for the case study and the technological limits. Therefore, we couldn't implement all metrics from the SOCCER framework. We choose to implement the metrics whom would pose the least difficulties in implementation and design. Therefore, we implemented Metric 1, Metric 2, Metric 3, Metric 7, Metric 11, Metric 14 and Metric 22 in the case study. Also, ML models need greater amounts of training data as you increase the number of input features. But, we collected a reduced training data set due

to the availability of the sports experts. Therefore, this also posed a limit to the number of SOCCER metrics we can implemented in the case study.

Metric 1 can help NLG systems to identify SFs about athletes and teams that are well known to them. Therefore, this was one of the first metrics we implemented in the case study. We used Twitter as a source of data because it verifies the identity of celebrities. A well known athlete who is verified will have a blue check mark which illustrates that he is the same person in real life. Although, we found that most soccer players and coaches don't follow this practice. But, we did find fan pages with followers and unverified accounts with their name. We computed the value of Metric 1 for Event Participants and Teams based on the number of followers they had on Twitter. Additionally, we computed the Metric 1 for the Role. Here we gave arbitrary values based on the feedback from the interviews with the sports experts and the literature review. Therefore, we attributed the highest value to forwards, followed by midfielders, defenders, goalkeepers, coaches and assistant coaches. The measurements of Metric 1 have a few weaknesses. First, we used the search query of Twitter to identify the soccer players on the network. In some cases, the query returned a user who wasn't the soccer player. But, the query worked properly for popular athletes who had a Twitter account. Therefore, the measurements would succeed for the cases of relevance we are trying to identify. Secondly, some coaches didn't have an official account on social media. As such, we selected different fan pages as a data source. We assume that these weaknesses impacted the contribution to the relevance of a SF of Metric 1.

Teams value their soccer players based on their performance and the value they add to the team. Sports fans have an interest in the players who play a crucial role in their team. Therefore, we implemented Metric 3 in the case study to identify facts about the important players. We measured Metric 3 at team and league levels. We extracted the player wages from the website Spotrac that contains different information related to the economic in sports such as wages, transfer values and so on (*EPL Contracts*, z. j.). We compared the salary of the player with his teammates to identify his importance for the team. We compared a player salary with others from Premier League to measure his importance for the league.

The template content type describe the core of a SF. Based on the template, a fact can be about goal scoring, pass accuracy, track record and so on. The Omega project uses over three hundred different templates to generate the SFs. We measured Metric 1, Metric 2, Metric 7, Metric 14 and Metric 22 for the content type Template with the help of sports experts CE1 and CE2. We asked CE1 and CE2 to evaluate the value of the enumerated metrics using a Likert based scale. For Metric 22 we asked the sports experts to focus on the utility of the Template for the betting domain. As such, Metric 22 had a high impact on relevance for the ML models that we developed for the betting domain. Figure 4.1 display the correlation coefficient values between the values given by CE1 and CE2 for the Template based metrics. There is a strong relationship between the measurements done by the sports experts between newsworthiness, popularity and importance. The three metrics cover similar areas of the relevance as an important template will also be popular and newsworthy most of the time and vice versa. But, they cover different edge cases. Therefore, the users of the SOCCER framework should evaluate if they need all three metrics in their use case. Also, we can see that there is a moderate correlation between the measurements done by the sports experts for the given metrics. Each expert used his tacit knowledge and experience to measure a given metric, thus the differences. Ideally, we would've used several sports experts to perform the measurements and select

the ones where most of them agreed on. But, we had to implement the case study with two sports experts. Therefore, we randomly selected the measurements for the metrics from the available data to reduce the risk of overfitting the measurements to the opinion of a sport expert.

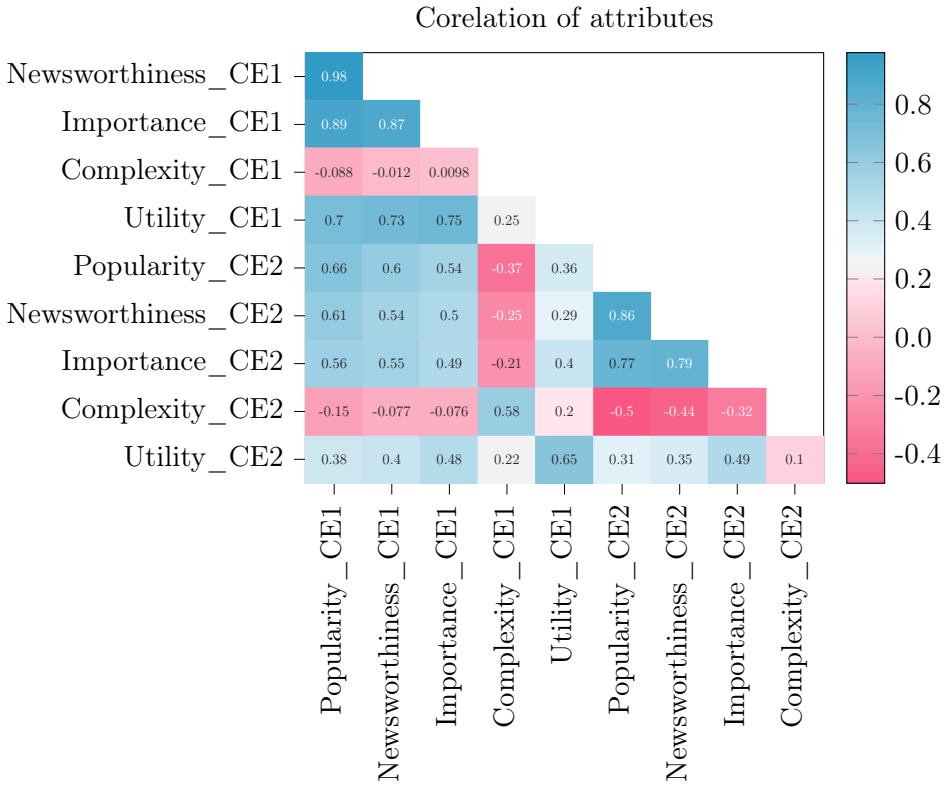

Figure 4.1 – Template related relevance measurements correlation coefficient values

Sports fans are interested in facts about impressive results. Similarly, gamblers can use such facts to guide their bets. Journalists can also use the most impressive stats to build their reports. As such, one of the most important metrics of SOCCER is Metric 11. The Omega project generated over 60000 SF for the English Premier League. The facts can describe impressive feats such as Liverpool winning streak of 40 matches, but also dull stats such winning only 2 of the last 5 matches. Therefore, it was crucial to differentiate between different sports facts based on the values present in them. We picked out the results from the SFs by parsing the text and extracting the numerical values. Then, we encoded the results in what we called focus variables. For a fact such as *A team won X matches of their last Y* we would encode X into focus variable 1 and Y into focus variable 2. We could then compare the same results of a fact with the results of other facts of the same type. Researchers and practitioners could also implement a parser to extract relevant content types from existing text to measure their relevance measurements. The approach would be helpful for NLG systems that use human made text as a source of data (Kanerva et al., 2019; ge Yao Jianmin Zhang Xiaojun Wan Jianguo Xiao Institute of Computer Science et al., 2017; Renjun et al., 2016).

The relevance metrics measurements would output values of different scales. For example, the measurements of Metric 11 would output a low value for the number of matches won by a team and 50 or 100 for a fact about the number of matches played by a certain

Relevance rating	CE1		CE2	
	News	Betting	News	Betting
Not at all relevant	288	109	327	150
Slightly relevant	176	34	184	161
Moderately relevant	164	311	135	270
Very relevant	82	224	66	104
Extremely relevant	38	70	36	63

Table 4.3 – Sports experts relevance ratings count

player. Also, the scale would change across metrics. The different measurement scales could create two issues in our case. First, it would be difficult to train the ML models as the metrics with higher values could have a larger impact for the model. Secondly, it would affect the adaptability of the models to predict the relevance of facts from other sports that would have different scales. We tackled this problem by applying a Min-Max normalization step in the relevance measurements to get a value between 0 and 1 (Patro & Sahu, 2015). The same procedure could be applied to measurements done for other sports. Therefore, the models that we trained with soccer data could be used to predict the relevance of facts about other sports. We didn't evaluate this assumption, but the models predicted relatively accurately the relevance of other SFs outside the dataset we used for training and testing.

4.4.2 Data collection

ML models need labelled data from which they can learn a task. In our case, we needed SFs with relevance measurement labels. However, relevance is a highly subjective matter. Additionally, there are no data sets available or definitive answers on what makes a SF relevant or not. Gracenote Sports company has a department where the employees select and use the SF generated by the Omega project to produce content or deliver it to their clients that deal with the news or the betting domains. Therefore, the employees of the Gracenote Sports company contain the tacit knowledge on what makes a SF relevant for the enumerated domains. As such, we got the help of sports experts CE1 and CE2 to label the facts generated by Omega project.

We designed the data labelling process by taking into consideration a few issues related to the case study. The main issue was that the sports experts had a limited time to label the data as they also had to do their regular job. Therefore, we had to adapt the process around this limitation. The Omega project can generate over 300 SF per match. It would require a great deal of time from the sport experts to label only the SF of a match. Therefore, we did a filtering of the facts to reduce their number to around twenty. First, we did the filtering by selecting facts that would fit in different relevance levels ourselves. Then, we used data analysis to increase the diversity of the labeled facts. Therefore, we got labeled facts related to all the templates used by the Omega project and at different levels of result significance. There can be a total of 760 matches played in Premier League as each of the 20 teams would play 38 matches. We had to do the data collection in a limited span of time. Therefore, we decided to label facts related only to the teams of Liverpool, Arsenal and Aston Villa from match week 16 on wards. The three teams were positioned on the first, ninth and seventeenth position in the rankings at the time. We picked the teams as their rankings would increase the diversity in the data due to their

given performance at the time. Using this procedure, we limited the amount of data the sports experts had to label.

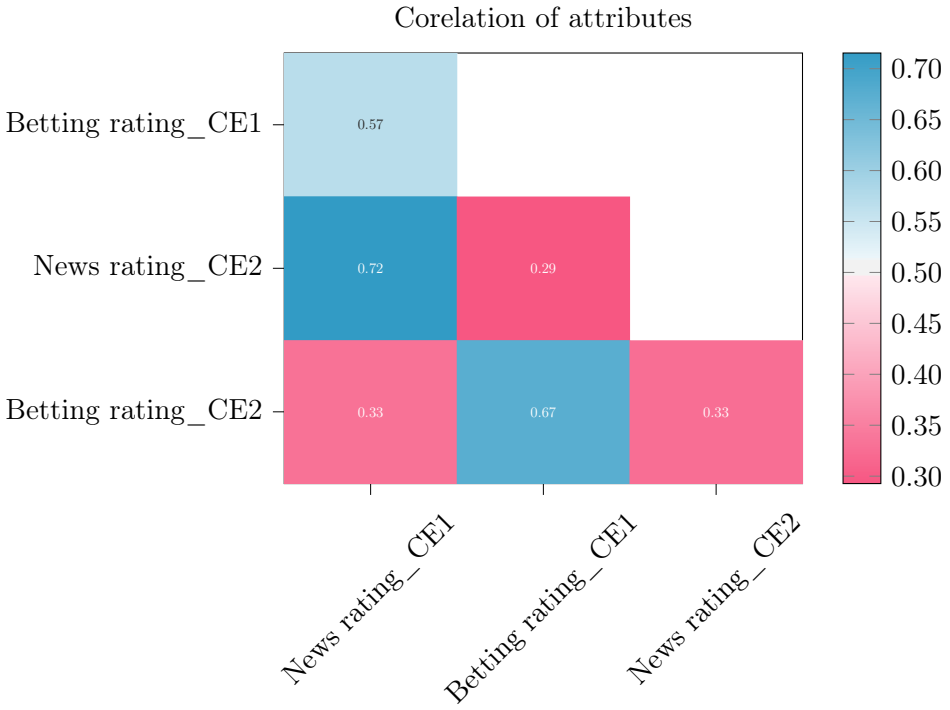

Figure 4.2 – Spearman rank correlation coefficients for relevance ratings

We collected labelled data on 748 sports facts. The sports experts evaluated the relevance of a sports facts on a Likert scale for the news domain as described in Sect. 2.4. The sports experts received instructions on the meaning of each relevance rating to guide their choices. Table 4.3 illustrates the distribution of the relevance ratings. CE1 rated a higher number of facts as being moderately relevant. CE2 found most sports facts as not relevant, in particular for the news domain. The Likert scale orders the facts in different relevance categories. There is a strong correlation between the ratings that CE1 and CE2 gave for the news domain as illustrated by the vale of $\rho = 0.72$. But, the value is not close to 1 which signifies a perfect correlation. The strong but not perfect correlation demonstrates a certain level of subjectivity involved in the choices made by the two sports experts. The differences are accentuated for the betting domain where there is a moderate correlation between the relevance ratings given by CE1 and CE2 as the value of $\rho = 0.67$. But, the values of the coefficients are close to each other, therefore consistent. Thus, we took into consideration that the two sports experts measure the relevance of a SF with certain differences. Ideally, we would use multiple sports experts to evaluate relevance and pick the most common choice. But, we were limited in terms of availability expertise for this case. For both experts, there is a moderate positive correlation between the relevance ratings for the news and the betting domains. The correlation displays the link between the relevance of SF for the two domains. Relevance is a subjective measurement as we can see from the values of the correlation coefficients. Also, we can identify a certain overlap in terms of relevance between the news and the market domains, especially for the ratings given by CE1.

4.4.3 Relevance measurements

The SOCCER framework provides a list of relevance metrics that can assist researchers and practitioners in developing or improving relevance measurements. In Sect. 4.3 we described how we implemented several metrics from the framework. The metrics can be used in a given heuristic to compute the relevance of SF. Li et al. used a neural network to develop an end-to-end NLG system that generated text based on the performance of the players but also their popularity and importance. Zhang et al. developed an NLG system that selected sentences from live event commentaries to generate the final output. The NLG system used a sentence selection ML model that leveraged a match event importance and the athlete popularity among other metrics to pick the sentences part of the output. For the case study, we followed a similar approach as we used ML models to perform the relevance measurements. However, the metrics could be integrated also in a simple weight based algorithm when the NLG system has a clear scope and definition of relevance.

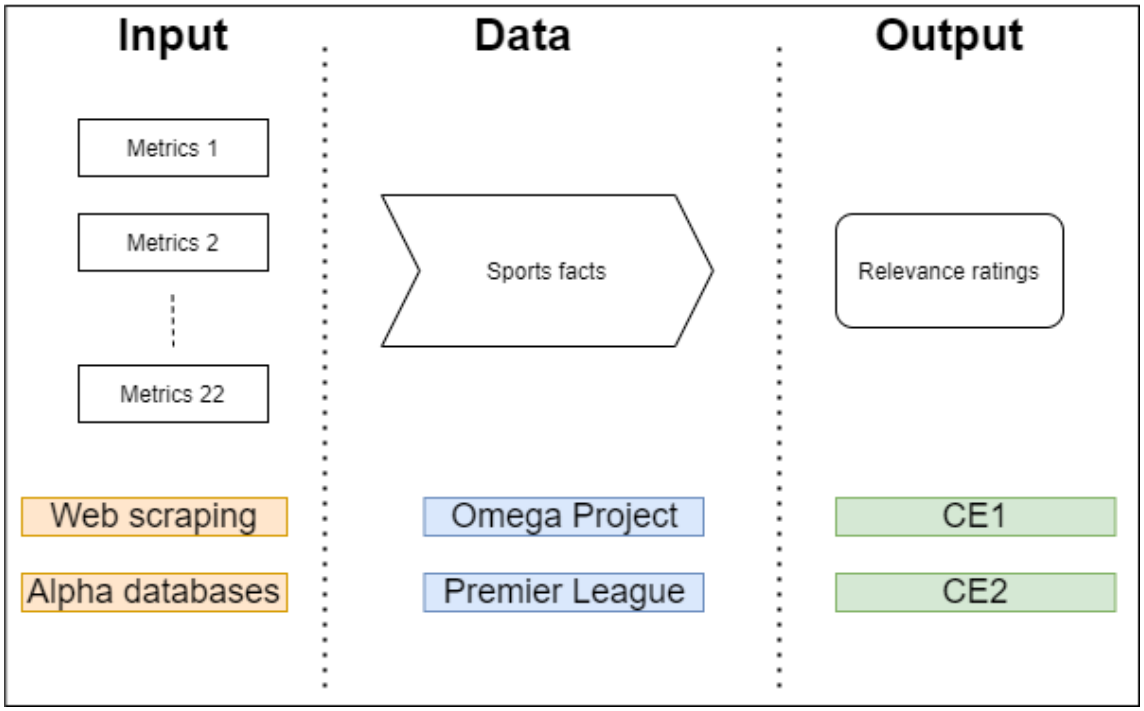

Figure 4.3 – Machine Learning architecture for the case study

The Gracenote Sports company has a department which filters and selects the relevant SFs for their clients in the news and the betting domains from those generated by the Omega project. CE1 and CE2 have 16 years of experience together in their current roles. They select and filter SFs on a daily basis. Therefore we labelled a set of SFs with their help to train several ML models. We trained the ML models on the labelled data by CE1 and CE2 to identify the links between the SOCCER metrics that we implemented in the case study project. The architecture of the ML models data flow is illustrated in 4.3. Therefore we scraped data and used the databases of Gracenote Sports company to compute the metrics of SOCCER framework as described in Sect. 4.3. At the same time, we collected a labelled dataset following a structured process as defined in Sect. 4.4.2. In this section we will present the ML algorithms that we used to train and the process we used to improve their predictions.

Domain	Model	Train Accuracy	Validation Score	Test Accuracy	Average		
					Precission	Recall	Fscore
News	Omega	17.4	-	14.67	0.19	0.32	0.15
	Random Forest	77.63	63.48	71.11	0.72	0.63	0.66
	Bagged Decision Tree	73.04	63.66	68.89	0.71	0.62	0.65
	Voting Classifier	74.00	63.48	68.00	0.67	0.62	0.64
Betting	Omega	36.71	-	37.78	0.29	0.25	14.29
	Bagged Decision Tree	93.12	71.89	75.56	0.69	0.66	0.66
	Random Forest	94.84	71.32	75.11	0.68	0.66	0.66
	Radial SVC	78.78	68.27	74.67	0.62	0.62	0.62

Table 4.4 – The accuracy of the relevance measurements for the Omega project and top three ML models

We used the Kaggle platform as a source of inspiration for the ML models, in particular the notebooks related to the Titanic dataset. The Titanic dataset widely used by inexperienced users to learn how to develop ML models, but also by advanced users who push the limits of such models. Therefore, there is a rich suite of notebooks that can serve as documentation. Additionally, the users tackle a binary classification problem as they have to predict if a passenger survives or not the disaster. We classified the SFs in five categories based on the Likert scale. Therefore, the the problem falls in the area of multinomial classification. We selected several ML algorithms that we trained for the labelled datasets for the news and the betting domains.

We followed an iterative process to develop and improve the accuracy of the predictions done by the ML models. First, we analysed and cleaned the data. The Omega project generates additional meta data for the OneLiners that we didn't need. Additionally, we encoded the different template types as numbers which we used as an input feature so that the ML models can link certain metrics to a template type. We randomly selected measurements and relevance ratings from the datasets labelled by CE1 and CE2 to reduce overfitting the models to the opinion of a sport expert. We were left with sixteen input features for the ML models after preparing the data. But, we had a small training set, thus we had to reduce the number of input features. The scikit-learn library provides a helper module to select the best predictor features for a target variable. We used the helper module to reduce the number of features to twelve. The module removed Metric 1 for the content types Team and Event Participant and several values related to Metric 11. As describe in Sect. 4.4.1, we generated several focus variables for Metric 11, some of which were applicable only for a subset of the facts which reduced their prediction value. There was a small set of facts related to the Event Participants in the training set. As such, the measurements of Metric 1 for Event Participant had a reduced impact on predicting the relevance of a SF. But, Metric 1 was still part of the model as the measurements for the content type Template were it was deemed to be a predictor of relevance by the helper module. Similarly, several focus variables related to the measurements of Metric 11 passed the test and were part of the final models. Initially, the models did well on the training set but did not generalise effectively to unseen data. Therefore, we used a module from skit-learn library to identify the best parameters for the models that would increase their performance on unseen data.

Omega project also performs relevance measurements by using Metric 7 and Metric 11. The application uses the two metrics and attributes relevance points between 0 and 100 to a fact. We use the Omega project relevance measurements as a baseline. The baseline

allows us to evaluate the performance of the ML models to those of the existing method. To do so, we converted the relevance points to our Likert scale based evaluation by dividing the 100 relevance points into five equal intervals. Next, we used the same performance measurements that we applied for the ML models to measure the accuracy of the Omega application relevance measurements.

The models successfully predicted the relevance of SFs using metrics from the SOCCER framework as illustrated by the accuracy metrics in 4.4. In the news domain, the best models had an accuracy of 71.11% for the test set with a Fscore of 0.66. The ML models outperformed the Omega application relevance measurement which have an accuracy of 17.4% for the train set and 14.67$ for the test set in the news domain. The models were influenced by the imbalanced dataset which contained a dominant amount of irrelevant SF as illustrated in Table 4.3. The Random Forest model has an FScore of 0.91 for the prediction class related to irrelevant SFs in the news domain. Similar, the Bagged Decision Tree had an FScore of 0.95 for the irrelevant class in the betting domain. The Omega application performed better for the betting domain where it scored 36.71% accuracy for the train set and 37.78% for the test set. The ML models could be put in practice in the Gracenote Sports company as it can reduce the work of the sports experts by filtering out irrelevant SFs compared to the current relevance measurements. The ML models performed even better for the betting domain where the models correctly identified the relevance of a SF of the test set in around 75% of the cases. The training set contained predominantly moderately and very relevant SFs. Therefore, the ML models for the betting domain had a higher performance for the after mentioned facts with an Fscore of 0.8 and 0.67 respectively by the Bagged Decision Tree model. In both domains, the best models had an Fscore of 0.66 for predicting the extremely relevant SFs which pose the highest interest for the Gracenote Sports company. The performance of the models is too low to automate the task of identifying the extremely relevant SFs. But, they can assist the sports experts in their job as the Omega project can generate around 300 SFs for a soccer match. The ML models would classify as extremely relevant 30 facts out of which 16 are truly extremely relevant.

4.4.4 Evaluation

In the case study we trained several ML models that predict the relevance of a SF using input features based on the relevance metric from the SOCCER framework. We measured the performance of the models by splitting the data set in two parts, one for training and one for testing. We discussed these results in Sect. 4.4.3. We further evaluated the performance of the ML models by measuring their accuracy in a simulated real life case. Additionally, we used a survey to grasp the perception of sports experts CE1 and CE2 on the relevance prediction of the best models for the betting and the news domains.

The Omega project generates a set of SFs for a given soccer match from which the sports experts CE1 and CE2 select the most relevant ones. We selected three soccer matches outside the range of the labelled dataset to measure the prediction accuracy of the ML models. The ML models relevance predictions would be used by the sports experts. Therefore, we evaluated their perceptions on the relevance predictions on two other sets of SFs related to known teams to them. The sports experts could then use their knowledge about the two teams to evaluate which facts would then be relevant and how the ones predicted by the ML models related to their selection. As such, we selected SFs from the following soccer matches to evaluate the accuracy and perceived performance of the ML

Model	Accuracy	Omega acc.	Sports expert	Domain	Match
Radial SVC	78.20	21.80	CE1	News	Match 1
Logistic Regression	71.43	13.53	CE2		
Radial SVC	81.95	38.35	CE1	Betting	
Logistic Regression	63.91	32.33	CE2		
Random Forest	78.38	14.86	CE1	News	Match 2
Naive Bayes	65.54	12.16	CE2		
Random Forest	81.76	35.81	CE1	Betting	
Logistic Regression	74.32	45.27	CE2		
Random Forest	63.68	17.04	CE1	News	Match 3
Decision Tree	58.74	11.66	CE2		
Radial SVC	74.44	34.08	CE1	Betting	
Decision Tree	56.95	37.67	CE2		

Table 4.5 – The ML models with the best score in the evaluation step and the accuracy score of the Omega application

models by the sports experts:

Match 1 Accuracy. The match between West Ham United and Sheffield United in round 10.

Match 2 Accuracy. The match between Newcastle United and Everton in round 20.

Match 3 Accuracy. The match between Leicester City vs Manchester United in round 38.

Match 4 Perception. The match between Manchester City vs Arsenal United in round 30.

Match 5 Perception. The match between Liverpool and Chelsea United in round 37.

We trained the ML models using SF related to three teams and their matches from round 16 until round 29. The models can overfit the labelled dataset and fail to generalise to predicting the relevance of SF for SFs related to other teams or from other rounds. Therefore, we further measured the accuracy of the ML models using labelled data by sports experts CE1 and CE2 for soccer matches of other teams from different rounds of the Premier League. The use case simulates a real life case in which the Gracenote Sports company and the sports experts would use the relevance predictions. The ML models and the sports experts gave relevance ratings to all the SFs that the Omega project generated for each of the tree matches. We extracted the best results of the ML models in Table 4.5. As illustrated in Table 4.5, the ML models surpassed the current relevance measurements of the Omega application in all cases. The ML models had a higher accuracy towards the relevance ratings given by CE1 compared to the ones by CE2. Also, the Logistic Regression model had the highest accuracy score in multiple cases for the news and the betting domain. It demonstrates that we can implement relevance measurement heuristics with simple algorithms. Also, the algorithms delivered similar results as with the training and the test sets. As a result, we can state that the ML models generalise and are able to predict the relevance of SF related to other Premier League soccer matches. But, we didn't evaluate the model with SF from other leagues, such as Bundesliga, or other sports. Hence, further research is needed to determine the generability of the ML models across different leagues and sports.

The ML models recommend a relevance category for a SF based on the data set that they were trained with. Similarly, recommender systems are trained and designed using

Perception	Questions	Match 4		Match 5	
		CE1	CE2	CE1	CE2
Quality	The ratings for the news market match my interests.	Agree	Agree	Agree	Agree
	The ratings for the betting market match my interests.	Agree	Agree	Agree	Neither agree nor disagree
	The ratings identify most relevant sports facts accurately.	Agree	Agree	Agree	Agree
	The ratings identify the least relevant sports facts accurately.	Agree	Agree	Strongly agree	Agree
	The ratings identify sports facts that are useful for gamblers.	Agree	Agree	Agree	Agree
	The ratings identify sports facts that are useful for journalists.	Agree	Agree	Agree	Agree
	The news ratings of sports facts take into consideration the requirements of the journalists.	Agree	Agree	Agree	Agree
	The betting ratings of sports facts take into consideration the requirements of the gamblers.	Strongly agree	Agree	Strongly agree	Agree
Ease of use	I became familiar with the relevance ratings very quickly.	Strongly agree	Neither agree nor disagree	Strongly agree	Agree
	I easily found the relevant sports facts.	Strongly agree	Disagree	Strongly agree	Agree
	I would quickly become productive with the relevance ratings.	Agree	Neither agree nor disagree	Agree	Neither agree nor disagree
	Finding relevant soccer facts, even with the help of the ratings, consumes too much time. *	Disagree	Neither agree nor disagree	Disagree	Neither agree nor disagree
Usefulness	The sports facts ratings could effectively help me find the ideal facts. *	Agree	Agree	Agree	Neither agree nor disagree
	The ratings would influence my selection of sports facts. *	Neither agree nor disagree	Agree	Neither agree nor disagree	Neither agree nor disagree
	I feel supported in selecting the sports facts that I need with the help of the relevance ratings.	Agree	Agree	Agree	Agree
Attitudes	Overall, I am satisfied with the relevance ratings of the sports facts.	Agree	Agree	Agree	Agree
	The relevance ratings made me more confident about my selection/decision.	Agree	Agree	Agree	Neither agree nor disagree
	The sports facts relevance ratings can be trusted.	Agree	Agree	Agree	Agree
Intentions	If a system that rates sports facts on relevance as this exists, I will use it to find relevant sports facts.	Agree	Agree	Agree	Agree
	I would use these relevance ratings frequently.	Agree	Agree	Agree	Agree
	I prefer to use this type of relevance ratings in the future.	Agree	Agree	Agree	Agree

Table 4.6 – Sports experts answers for the Resque survey

user data to predict a user preference. In the case study, the users group is formed by the sports experts of the Gracenote Sports company who select and filter SFs. As such, we evaluate the sports experts perception towards the relevance ratings through CE1 and CE2. We predicted the relevance categories of the facts generated by Omega for **Match 4** and **Match 5** using the most performing ML models on the test set. We predicted the relevance ratings for the news domain using the Random Forest ML model and Bagged Decision Tree for ratings related to the betting domain. The sports experts then evaluated their perception on the quality, ease of use, usefulness, their attitudes and intentions towards the relevance ratings and the system that predicted them. Table 4.6 displays the questions and the choices made by the sports experts for the given relevance ratings. The sports experts found the quality of the predicted relevance ratings of sufficient quality as they agreed with the positive statements. CE1 has the opinion that the relevance prediction for the betting market are in accordance with the needs of gamblers as it strongly agreed with the related statement on both sets of SFs. The goal of the system is to assist the sports experts to identify relevant SFs easier and faster. But, CE2 was neutral towards several statements that evaluated the ease of use of the relevance predictions, and even disagreed in one case. CE2 stated in the feedback that „Most ratings are accurate in my opinion, good job. Some are a bit high or a bit low, but that might be my preference". His opinion can also be seen in the accuracy metrics where the system predictions had a higher precision towards the ratings of CE1.Overall, the sports experts found the relevance predictions useful as they illustrated their intention to use a similar system in the future if it would exist.

Table 4.4 pictures the ML models and Omega project accuracy for the training and the test set. The results display the integration ability of the relevance metrics from the SOCCER framework into relevance measurement heuristics. Table 4.5 contains accuracy measurements of the ML models and of Omega project for data outside the range of the training set. The results are similar to those on the training set. The ML models accomplished a higher accuracy than the current relevance measurement performed by the Omega project in all data sets. The relevance measurements accuracy results demonstrate the versatility of the ML models to measure the relevance of related SFs compared to the current Omega project relevance measurements. Therefore, the relevance measurements could fulfill the automated journalism domain adaptability requirement for NLG systems. Finally, table 4.6 lays out the opinion of sports experts CE1 and CE2 towards the quality, ease of use and usefulness of the relevance measurements. The relative positive results demonstrate that the relevance measurements could be integrated in an expert based system to evaluate the relevance of the SFs.

In this chapter we have presented the main artifact of this research study, the SOCCER framework. Sect. 4.2 discusses the seven content types that could form a SF. Sect. 4.3 present the twelve relevance properties that we discovered during the literature review and the relevance metrics that we composed and refined with the help of the literature review and several sports experts. Finally, we evaluate the usability of the SOCCER framework in a case study which we present in Sect. 4.4. But, the research has certain limitations which we will present in Sect. 5.3.

Chapter 5
Discussions

In this chapter we will discuss the results of the research project in regards to its goal and sub-research questions in Sect. 5.1. The results of this research project contribute to the main body of knowledge where we identified a need of relevance measurements. We will present the main contributions of the research project in Sect. 5.2. Next, we will discuss the validity threats and the limitations of the research project in Sect. 5.3. The SOCCER framework set up the ground work for relevance measurements. Therefore, there are several paths for future research which we will discuss in Sect. 5.4.

5.1 Research questions

We uncovered a need for a set of relevance measurement guidelines from practitioners and researchers during a preliminary feasibility evaluation of the research subject. We aimed to improve the relevance measurements of SF by creating a relevance measurement framework. To do so, we extracted the requirements and limitations of NLG systems. In Sect. 4.1 we describe the requirements for NLG systems that we discovered during the literature review. We defined the main research question with the research goal in mind. We further expended the research in several steps which we translated into sub-research questions. We present the answers to the sub-research questions in this section which lead to a definitive answer to the main research question in Chapter 6. To this goal, we will enumerate each of the questions and their answers in this section.

SQ1: Which content properties affect the relevance of a sports fact and how can they be quantified?

We further divided SQ1 in three parts to guide our research process. **SQ1.1** captures the first step where needed to investigate and identify which content properties can be linked to relevance. Researchers and practitioners need guidelines how should they quantify the impact towards relevance of a given content type and relevance property. Therefore, we created **SQ1.2** to capture this step. Finally, we focus on the issue of relevance for the news and the betting domains. Therefore, we defined **SQ1.3** to identify the differences in terms of relevance measurements between the two domains.

First, we had to identify and define the content types and their properties which influence the relevance of a SF. We extracted seven content types from the IPTC sports data interchange schema (*SportsML - IPTC*, z. j.) which we describe in Sect. 4.2. NLG systems will produce facts which contain at least one of the content types Event , Event Participant, Role, Team, Location, Template or Instance. Each content type can be part of a SF and it can influence the relevance of a SF when it contains a relevance property. We performed a literature review in which we identified twelve properties that are used to measure the relevance of content in their respective field. We distinguished researchers who used popularity (Li et al., 2019), newsworthiness (ge Yao Jianmin Zhang Xiaojun Wan Jianguo Xiao Institute of Computer Science et al., 2017), importance (Kanerva et al., 2019; Zhang et al., August, 2016), significance (Leppänen et al., 2017; ge Yao Jianmin

Zhang Xiaojun Wan Jianguo Xiao Institute of Computer Science et al., 2017) to measure the relevance of the sports content part of an NLG system. We extracted the properties of unexpectedness, novelty, utility, peculiarity from the data mining field where they are used to identify relevant and interesting association rules (Kontonasios et al., 2012). During the literature review we recognised that stakeholders have an interest on content that can predict the outcome of certain events (van Wijk, 2012; Bhattacharjee & Talukdar, 2019). Therefore, we extracted and defined the property *predictability* which defines the ability of a content type to predict future outcomes. The sports analytics field is constantly evolving and new measures and types of stats are created constantly. The statistics and their corresponding content become increasingly complex as this evolution proceeds. Therefore, we defined the property *complexity* to measure the level of sports knowledge a stakeholder needs to understand a certain piece of information. Sports events take place linearly across time which influences the interest of stakeholders who are mostly keen about the recent events. We defined the *timeliness* property to attribute the reduce of relevance of a content type with the flow of time. We attributed the relevance properties to the content types related to SF, thus becoming sports content properties.

We established a relevance metric for a given content type and relevance property using the knowledge we gathered during the literature review to quantify the impact on relevance. Each relevance metric contains two sections. First, the metric contains a definition of what an NLG system should measure to quantify the impact on relevance of the content type that contains the relevance property. For example, Templates describe certain actions that happen in a sport, each with its own degree of importance in the sport depending on the its rules. In soccer, a goal has a higher degree of importance than a pass. As such, Metric 7 should measure the importance of the sports action that is part of the Template to the sport to determine its influence towards relevance. Secondly, the metric contains a description of how could a researcher, practitioner perform the measurement for that metric. Metric 7 can be measured with the help of sports experts which know the rules of the sport and thus, they can quantify the importance of a sports action. Several sports experts evaluated the validity of the existing relevance metrics and provided insights to create additional ones. We ended with twenty six relevance metrics which describe what and how we can quantify the impact of a content type and relevance property towards the relevance of a SF. The twenty six relevance metrics from Sect. 4.3 provide the answer to **SQ1.2**.

We determined the betting and the news domain as the targets of NLG systems for sports. We developed SOCCER as a unified relevance measurement framework. But, we assumed that there would be differences between the two domains, therefore we established **SQ1.3**. We debated domain specific situations that can influence the relevance of a SF with the sports experts. We got additional information during the design process where we created the *predictability* property which is linked to a prediction algorithm with a target on the betting domain. We enumerated the situations and suitability of a metric for a given situation or domain in the section of of the metrics that required this information. For example, we extended the *utility* property to all content types at the advice of the sports experts. Metric 1 and Metric 2 are suitable for the news domain where the readers are interested in well known players.Metric 11, Metric 26, Metric 25 are suitable for the betting domain as they provide gamblers with valuable information which can assist them in their scope.

SQ2: How can we use the SOCCER framework to build tailored heuristics for relevance measurement of SFs in NLG systems?

SQ1 is focused on the investigation phase of the research project which we answered by designing the SOCCER framework. The goal of the SOCCER framework is to assist researchers and practitioners to develop new relevance heuristics for sports content in NLG projects. It can happen that researchers and practitioners need to tailor the relevance measurements to their use case. For example, Lee et al. developed an NLG system that tailored the output to the fans of the two teams playing a soccer match by picking up the related positive situations and results of a team. We captured this work into SQ2 and its sub-tasks. Through **SQ2.1** answer we are looking to provide researchers with knowledge on how to use the SOCCER framework to measure the relevance of SFs using the SOCCER framework. **SQ2.2** focuses on the relevance tailoring options that could be supported through SOCCER framework. We will elaborate on the answer to SQ2 and its sub-tasks next.

The SOCCER framework metrics can be used to build heuristics that determine the relevance of a SFs. We performed relevance measurements for SF generated by an NLG system from a sports data company using metrics from the SOCCER framework as input features for ML models as described in Sect. 4.4. We scraped social media data and used the Gracenote Sports company databases to implement measurements for Metric 1, Metric 2, Metric 3, Metric 7, Metric 11, Metric 14 and Metric 22. Sports experts and practitioners can implement the relevance measurements by using public, open-source, scraped or commercial data. Gracenote Sports company provides sports data under commercial contracts, but researchers can use the data sources mentioned in Sect. 2.2.1 to implement measurements for the relevance metrics related to the Instance content type. Metric 1 and Metric 2 can be implemented by scraping data on social media or news outlets. The metrics related to the content type Template can be measured with the help of sports experts as they require specific knowledge to the sport. But, researchers and practitioners that use the SOCCER framework decide the data source and method of measurement. The metrics only define what that metric measures and how it can be measured. Therefore, researchers and practitioners can use the data that they have available to build their measurements.

We used the relevance metrics as input features for several ML models that had the goal to predict the relevance of SFs. We tailored the relevance measurements for the news and the betting markets by gathering a labelled dataset of SF with relevance measurements for the two markets. We used each market specific dataset as the predictor variable of the ML models. As such, the ML models could „learn" which input features correspond to a relevance rating. The ML models assign certain weights to the input features to predict the output variable. Therefore, the SOCCER metrics could also be used in a simple, weight based heuristic. Researchers and practitioners can further tailor the heuristics for relevance measurements by implementing a subset of the SOCCER metrics in their NLG systems.

The labelled data set is a crucial step in performing relevance measurements using ML techniques. Therefore, we accentuate that researchers and practitioners need to establish and define a rigorous process that follows established guidelines to gather the labelled data. The process need to take into account the degree of subjectivity involved in the perception of relevance by sports experts or the subjects performing the labelling of the data.

The SOCCER framework provides a set of metrics that can be used to measure the relevance of sports content in NLG systems for a broad number of cases and situations. The metrics act as a set of parameters that can derive the relevance of a SF. They can

be implemented by using sports data, with the help of sports experts or by analysing the interests of the users of an NLG system. Researchers and practitioners can tailor the relevance measurements for the news or the betting markets by selecting a sub-set of the metrics or the methods they use to combine the metrics into a relevance heuristic. We tailored the measurements by gathering market specific relevance measurements with the help of sports experts and then training two sets of ML models, one for the betting market and the other one for the news market.

SQ3: How do experts perceive the quality of the results delivered by the relevance measurement heuristics?

SQ3 acts as the validation step of this research project. We implemented the SOCCER metrics in a case study and demonstrated that the SOCCER framework can be used to develop relevance measurement heuristics. We assessed the performance of the relevance measurements to illustrate the utility of the relevance heuristics built with the help of the SOCCER framework. We measure quality using a two fold method. First, we measure the perceived usefulness of the measurements which is defined in **SQ3.1**. Secondly, we measure the accuracy of the relevance measurements compared to sports experts which we capture in **SQ3.2**. In the following we will answer SQ3 by elaborating on the accuracy and perceived qualities of the relevance measurement heuristics by the sports experts.

We developed several ML models to perform the relevance measurements of SFs generated by the Omega project. As such, we measured accuracy using performance measurement techniques from the ML field. The models would predict one of the five relevance categories for a SF based on a Likert scale, with 1 being not relevant and 5 extremely relevant. For each model, we measured the mean accuracy, precision, recall and F score. We performed the accuracy measurements on several data sets related to the news and the betting markets. We used the current relevance measurements computed by the Omega project as a baseline. The training data set illustrated the ability of the models to learn from the data to predict the outcome variable. The test data set was used to measure the performance of the models on unseen data. The industrial use case data set performance measurements demonstrated the ability of the ML models to be used in industrial systems. Table 4.4 contains the results of the best three ML models on the training and the test set for the news and the betting market. For the news domain, the Random Forest model had a mean accuracy of 77.63% on the training set and 71.11 on the test set. For the betting domain, the Bagged Decision Tree had an accuracy of 93.12% for the training set and 75.56% for the test set. The results are promising considering that we trained the ML models using a small data set. Furthermore, the ML models have a higher accuracy than the current relevance measurement computed by the Omega project. But, the performance of the models are not high enough to be used in an industrial system. Table 4.5 contains the accuracy results of the best performing ML models for the industrial use case. The ML models had a higher accuracy towards the ratings of sports expert CE1. The highest accuracy of 81.95 was reached in the case of the ratings of Match 1 by CE1 for the betting market. In several cases, the ML models had a higher accuracy for the industrial data set than the testing data set. But, the models are biased towards the relevance measurements done by CE1 as the accuracy scores are significantly higher in the case of the measurements done by CE1 compared to the ones of CE2. Overall, the models have performed in an acceptable accuracy range. Although, they aren't ready to automate the jobs of sports experts, they can still offer assistance to the sports experts in filtering the facts in terms

of relevance. The ML models can replace the current measurements done by the Omega project as they have a considerable higher accuracy.

The accuracy measurement gives us an objective measure on the performance of relevance measurements. We used a survey with Likert scale responses to measure the perceived quality, ease of use, usefulness, the attitudes and intentions of CE1 and CE2 towards the relevance measurements predicted by the ML models with the highest accuracy on the test set. The results are displayed in Table 4.6. Overall, the answers of the sports experts were positive. CE2 was neutral towards several questions as he found that some ratings were either a bit to high or too low from his own. The answers correlate with the results in the accuracy evaluation where the ML models had a lower accuracy towards the relevance measurements done by CE2. CE1 and CE2 filter and select SFs from the once generated by the Omega project on a daily basis. Therefore, they evaluated whenever the relevance measurements done by the ML models could assist them in their task. The ML models are not able to automate their task considering their accuracy evaluation results. But, the sports experts were positive in most of their responses. Also, the ML models can identify irrelevant SFs with a high accuracy an acceptable one for the other relevance categories. As such, the relevance heuristics that we developed using the SOCCER framework can ease the fact selection process for CE1 and CE2.

5.2 Main contributions

We make a contribution to the research field through the main artifact of this research project, the SOCCER framework which established a set of content types, relevance properties and metrics that can be used to identify relevant content in the sports domain by NLG systems. The case study results demonstrated the ability of the framework to lead to the development of new relevance measurement heuristics for sports content in NLG systems. The ML models implemented in the case study identified the relevant SFs with an accuracy of up to 78.38% for the news domain evaluation dataset and 81.95% for the betting domain evaluation dataset as illustrated in Table 4.5. The usefulness of the framework and the relevance measurements is further enhanced when compared with the relevance measurements computed by the Omega project. The current measurements had the highest accuracy of 45.27% in the betting domain, while the corresponding model had an accuracy of 74.32% on the same data set. Furthermore, the sports experts perceived the relevance measurements useful as demonstrated by their responses in Table 4.6.

First, we recognise that researchers don't use a defined set of content types in terms of relevance measurements. Therefore, one of the first contribution to the research field are the seven content types part of SOCCER. The seven content types are the main components that can constitute a SF, a piece of natural language. Researchers can develop the architecture and the content generation phase using the seven content types as a standard to form new templates or organise the output. The content types are defined based on the IPTC schema which acts as a media data transfer protocol (*SportsML - IPTC*, z. j.). Researchers can implement the content types a standard in their NLG system to share templates across NLG systems. Additionally, the content types can be integrated into a standard OWL sports ontology which can serve as a source of semantic representation for concept based NLG systems (Q.-M. Nguyen, Cao & Nguyen, 2015).

Secondly and most importantly, the SOCCER framework defines a set of general, transparent and adaptable relevance properties and measuring guidelines that can assist re-

searchers into improving or creating new relevance measurements in NLG systems that target the sports domain. The relevance metrics contain general guidelines on what to measure to identify the relevancy of a SF. Therefore, the metrics are adaptable to all sports based content. During the literature review we uncovered only NLG systems that target a single sports, such as basketball (Li et al., 2019) or soccer (Lee et al., 2017) due to limitations in architecture, data availability or the content determination phase. The SOCCER framework will allow NLG system to employ advanced relevance measurements that can fit multiple sports through its generic content types and metrics.

The SOCCER framework sets up the groundwork for relevance related topics in the field of sports based NLG. Additionally, the development of the framework can act as a guideline for researchers to extract relevance metrics for other areas of interest for NLG systems such as financial, business reporting, medicine or other domains. This research study adds to the growing body of research related to NLG systems by filling a gap in literature related to sports content relevance. We will further expand on the future research areas related to the SOCCER framework in Sect. 5.4.

Although the framework was evaluated in a single study in the sports of soccer, the results reveal the ability of SOCCER to handle relevance measurements in unfavourable conditions. Furthermore, the study was executed on top of an existing NLG system which demonstrates the adaptability of SOCCER and the ease of integration with existing systems. We implemented a prototype that performed relevance measurements for SFs generated by an industrial NLG system used by a company. The results display that SOCCER framework can be used by practitioners to aide existing relevance measurements or even replace them.

Concluding, the SOCCER framework can be of assistance for researchers and practitioners that develop NLG system for the sports domain. SOCCER framework fills in the gap we identified in literature related to the measurement of relevance of sports content in the NLG field. But, this research project has its limitations which we will present in the following section.

5.3 Limitations

In this section we will elaborate on several shortcomings of this research project that we identified before, during and after performing the research. We will consider four levels of validity threats that limit the conclusions of the research project as described by Wohlin et al.. Internal validity concerns the validity of the actual study. External validity are limitations that pose an issue in generalising the research towards industrial practice. The conclusion validity implicates the limitations that reduce our ability to draw conclusions. The construct validity concerns the constraints towards generalising the results towards theory beyond the study. In this section, we will present the threats that can influence the validity of the research project and the strategies we used to mitigate them.

Construct validity

We identify two threats that can influence the construct validity of this study based on the research method we employed. The first threat is derived from the methodology we used to perform the literature review. Sect. 2.2.1 elaborates on the process we used to perform the literature review. The threat is derived from the fact that we didn't perform a forward or backward snowballing process on the identified relevant research papers. We used a

data extraction form to extricate relevant information to build the SOCCER framework. Therefore, there is a risk that we overlooked research papers which could provide us with valuable information. But, we used loosed inclusion and exclusion criteria as illustrated in Appendix A which allowed a sufficient amount of papers to pass to the data extraction stage. We extracted data from forty papers related to the sports analytics field and forty four from computational linguistics area. Therefore, we mitigated the construct validity threat by extracting sufficient theoretical knowledge to design the SOCCER framework.

The second threat to construct validity is represented by miss interpreted information. Besides data from the interview, we used sports experts interviews to validate and extend the SOCCER framework. The interviewees might misinterpret the questions and offer the wrong feedback. We mitigated this risk by piloting the questions and the process with two other employees of Gracenote Sports company. Initially, we planed to do in person interviews, but we had to change the process due to COVID-19. Therefore, we modified the process to use online tools. We further supplied the interviewees with additional instructions regarding the interviewee process. Also, during the interview we asked the interviewees to share their screen so that we can notice when they don't follow the process or have issues. We mitigated the risk of misunderstandings by executing a structured, validated and clear interview process.

Internal validity

The threats that influence the internal validity of this research project are related to single group tests. The main internal validity threat is due to the fact that we executed the case study while the COVID-19 situation was taking place. For the case study, we collected data for training and evaluating the ML models. Due to the COVID-19 crisis, the participants were under additional stress as their well being was at risk. Therefore, there is the chance that the data we collected would be different if it would be gathered at a different time. We mitigated the risk by using a flexible data gathering schedule without any set deadlines. Secondly, the labelling data process can be tedious and lead to boredom. As such, the participants might react differently over time. We mitigated the threat by limiting the amount of work for the sports experts. We used a person first strategy to reduce the internal validity threats due to COVID-19. We kept in touch regularly with the CE1 and CE2 and explicitly stated that the data labelling process can be done at their own pace.

External validity

We performed a case study where we measured the relevance of SFs produce by a commercial NLG system of a sports data company. We developed a prototype following the format of a TAR study as described by Wieringa. Therefore, we implemented a prototype system in an industrial setting at a sports data company. The ML models are not able to automate the task of sports experts as their accuracy can vary and it can be biased to the opinion of a sports experts. But, the current system can assist the sports experts of Gracenote Sports company to identify relevant facts in a larger collection. Therefore, the case study proved the usefulness and the ability of the SOCCER framework to be integrated in industrial systems that perform relevance measurements for SFs, thus mitigating the risks towards external validity of the research project.

Conclusion validity

There are several threats that influence the validity of the conclusions we draw from the case study results. The case study required the assistance of sports experts, but they have a certain disparity and time availability for assisting research tasks that are time consuming. Therefore, we got the help of only two sports experts to gather the data needed to train and evaluate the ML models. We mitigated the risks by applying a two fold evaluation method. First, we measured the accuracy of the relevance measurements compared to the ones done by the sports experts. Secondly, we evaluate the perceptions of the sports experts towards the relevance measurements. But, the population is formed only by two sports experts therefore we are limited in generalising our conclusions.

The researcher is an employee of Gracenote Sports company which constitutes a risk to the conclusion validity as the sports experts could offer optimistic evaluations towards the relevance measurements performed by the ML models. We mitigated this risk by performing accuracy measurements that can illustrate objectively the performance of the ML models. Secondly, we emphasized that sincere feedback would increase the value and the validity of the research project.

We identified the risks and issues of the different steps we took during the research project. We applied different strategies to mitigate such risks and increase the validity of the research. For example, we limited the number of questions during the sports experts interview to reduce mental fatigue and increase the value of the data gathered. We randomise the labelled data that we used to train the ML models to reduce over fitting towards a certain pattern in the data and increase the generalisation of the relevance measurements. But, there are steps we would have perform differently if we had the necessary resources available. We conducted the research project taking into consideration the risk and we have drawn tentative conclusions considering the validity threats and the limitations of the research.

5.4 Future Work

Relevance is a subject of great interest in the field of information science as it is approached on several related domains such as IR (McCrudden & Schraw, 2007), TS (McCrudden & Schraw, 2007), NLG (Nichols et al., February 14, 2012). This research project sets up the ground work for relevance measurement of sports content in the field of NLG. Therefore, we will enumerate next, several areas of research that can use the SOCCER framework. But, firstly we will describe the areas in which the SOCCER framework itself can be further validated and extended.

We performed a single case study in the sport of soccer to evaluate the applicability of SOCCER framework in practice. Future research can further validate the SOCCER framework by implementing relevance measurements for other sports. Also, we perform the relevance measurements on natural language generated by an NLG system. Researchers could use the SOCCER framework to implemented relevance heuristics in the content determination phase of their NLG systems. Lastly, practitioners can use the SOCCER framework to improve the relevance measurements of commercial NLG systems to further validate the usage in industrial settings.

The field of relevance is vast and influences a multitude of domains. Therefore, the subject receives the attention of researchers and practitioners. We identified seven content types and twelve relevance properties following a structured literature review process under

a limited time frame. Thus, we assume that there are other relevance properties that could influence the relevance of SF. As such, future research can broaden the research focus towards uncharted domains by this research project to further develop the SOCCER framework with new relevance properties. Furthermore, the SOCCER framework can be extended with new content types as the NLG and the sports analytics field advance.

Researchers can use the SOCCER framework to further expand on its goal. We aimed to assist researchers and practitioners to develop new, shareable and adaptable relevance heuristics for NLG systems that generate content for the sports domain. Future studies, can use the SOCCER framework to develop sets of heuristics that can measure the relevance of sports content using metrics from the SOCCER framework. To do so, we encourage researchers to create labelled data sets with SFs that have their content types defined and the relevance measurements labelled either by sports experts or through other data collection methods. By doing so, they can developed relevance heuristics based on ML models. The labelled data set can be used for analysis to determine a suitable heuristic that determines the relevance of a SFs. Furthermore, the SOCCER framework, the relevance heuristics and the labelled data set can be used to develop new automatic evaluation methods for NLG system that target the sports domain as current automatic metrics such as BLEU poorly reflect human judgement (Novikova, Dušek, Curry & Rieser, 2017).

Relevance is a concept which requires constant research and improvement to deliver results to its beneficiaries. We have set up the ground work for the research and development of the relevance concept in the NLG field in relation to sports. Future work can extend and further validate the SOCCER framework. Additionally, researchers can use the SOCCER framework metrics to develop unified and comparable relevance measurements heuristics that they can share and improve collectively.

Chapter 6
Conclusion

We identified a gap in literature with regards to relevance measurement in NLG system that target sports data. As such, we started this research project with the goal to fill that gap. We did so by developing the SOCCER framework with goal of assisting researchers and practitioners to develop new relevance heuristics. We validated the framework in a case study where we performed relevance measurements for SFs generated by an industrial NLG system. We answered each of the sub-questions of the research project in Sect. 5.1. Therefore, in this section we will discuss the answer to our main research question. Furthermore, we will present our final conclusion in regards to the results we obtained and the insights we gained by executing the research.

6.1 Main research question

RQ: How can a content relevance measurement framework for SFs be developed to aid practitioners and researchers in improving and constructing relevance measuring heuristics for NLG systems?

Relevance is a subject that has been well researched in domain of information science. Therefore, we performed the first step by executing a literature study to identify the methods and the properties other domains link to relevance. We executed a structured literature review from which we extracted twelve properties that are used in data mining and the NLG fields to identify the relevant content for their use case. The framework needs to take in account the architecture and the needs of the NLG field. As such, we used the IPTC schema to define seven general content types for which we linked the relevance properties. We expanded our knowledge base about sports by including the sports analytics field in the literature review. By doing so, we gained a deeper understanding about the sports rules and the relevant analytics used to measure performance. We used this knowledge to adapt the relevance properties to the sports field and define the measuring guidelines that can assist researchers and practitioners in developing relevance heuristics.

We used the research field body of knowledge to build the initial version of the SOCCER framework. But, practitioners also deal with sports data to create relevant content for fans, gamblers and other stakeholders. Therefore, another part of the knowledge lies with them. Sports experts use their experience and insights to pick and generate relevant sports content. That is the case at Gracenote Sports, a sports data company, which developed the Omega project, a template based NLG project that creates SFs related to soccer matches. Several employees of Gracenote Sports company select the relevant SFs from the ones generated by the Omega project. We identified several sports experts from the Gracenote Sports company and a researcher who developed an NLG system that targeted the sports domain. We executed several interviews with the sports experts to validate, extend and adapt the SOCCER framework with their knowledge.

The framework needs to assist practitioners and researchers to develop and improve the relevance heuristics in their NLG systems. We evaluated the SOCCER framework ability to create relevance heuristics through a case study at Gracenote Sports company.

We implemented several ML models that measured the relevance of the SFs generated by the Omega project using metrics from the SOCCER framework. Researchers and practitioners can implement relevance measurement heuristics in a similar fashion or using custom algorithms due to their versatility. We measured the perceived quality of the relevance measurements and their accuracy with the assistance of sports experts. CE1 and CE2 found the measurements of assistance for their process as their feedback was mostly positive. But, the ML models aren't able to automate the task of selecting relevant SFs as their predictions on the industrial evaluation data set fluctuated between 56.95% and 81.95% mean accuracy and were biased towards the decisions made by CE2. Although the accuracy measurements fluctuated, the sports experts were positive about the results. The ML models can ease their work as they can identify irrelevant SFs with a high accuracy and the other relevance categories within an acceptable range. With the case study and the quality measurements, we demonstrated the ability of the SOCCER framework to aid practitioners and researchers in developing and improving the relevance heuristics in their NLG systems.

The SOCCER framework defines seven types of content that can form a SF. The seven content types can be integrated into template based or end-to-end NLG systems. We identified and used twelve relevance properties to define a set of measurements guidelines that researchers can use to develop relevance heuristics that identify relevant content for the sports domain. We used the SOCCER framework in an industrial NLG system to measure the relevance of SFs. The models performed within acceptable ranges that illustrate the ability of the SOCCER framework to assist researchers and pracitioners in developing new relevance measurement heuristics. We combined scientific research, sports specialists expertise and hands on experience to set up the foundations of relevance measurements for NLG systems that generate sports content.

6.2 Conclusion

Sports plays an important role in our lives through the sports we practice or watch. Media companies provides its users with increasing amounts of content related to sports. Another aspect of sports content is represented by the betting domain. Sports fans and gamblers use the content either for entertainment or as a source of information for placing bets. But, the sports stakeholders can overflow the users with information as they capture increasing amounts of data (Albert, 2010). As such, sports stakeholders need to select and filter the information they deliver to suit the needs of the fans or gamblers. The sports field is a recurring topic for the research field of NLG due to the availability of the data and the almost standard text reports written for a given event. But, the large amounts of data available and the various statistics increase the complexity of the identifying relevant content for NLG systems, journalists and other stakeholders such as Gracenote Sports company. Furthermore, there hasn't been any research until now, from up to know, about what makes a piece of content relevant for the sports domain that can assist NLG projects in identifying suitable content for sports stakeholders. Therefore, we started this research project to fill this gap in literature.

We designed the SOCCER framework by taking into account the different architectures of NLG systems. Therefore, the metrics can be integrated in template based NLG system due to its usage of content types that would usually be part of a given template used to generate content. The relevance measurement heuristics can be implemented in

the content determination as regular algorithms or even with ML based techniques. In end-to-end NLG systems, the content types and their measurements can be part of the input features of the ML models as shown in the case study. We gathered valuable insights by implementing relevance measurements using metrics from the SOCCER framework in an industrial NLG project. We had to implement the measurements on top of the existing project. As such, we used the templates defined for the Omega project to parse the data and extract the content type Instance. The approach can be used by practitioners and researchers who want to evaluate the SOCCER framework in their projects without modifying their content determination stage. The parsing method can also be used in NLG projects that use Twitter data to summarise or generate content related to sports (Nichols et al., February 14, 2012; Kubo et al., November 17, 2013). The parser can extract a content type of the SOCCER framework from a piece of text and perform the measurement of the metrics. Next, the metrics can be used to perform relevance measurements. The SOCCER framework is flexible in implementation through its generic content types, relevance properties and measuring guidelines therefore it can be adapted to all types of NLG systems.

During the case study, we gathered a labelled data set with relevance measurements for the news and the betting markets that were performed by sports experts. The labelled data set is mandatory for training ML models. But, we see that a labelled dataset can be of significant assistance when developing NLG systems using SOCCER metrics or other methods. Researchers and practitioners can use the labelled data set together with the SOCCER metrics to build and compare their relevance measurement heuristics as the users of Kaggle platform compare their ML models for a given data set. Also, a standard labelled data set can be used as a reference for automatic evaluation metrics for content relevance.

The case study illustrates the ability of the SOCCER framework to assist researchers and practitioners to derive new relevance heuristics. The nature of the case study demonstrates that the SOCCER framework metrics can capture the experience of sports experts in ML models. We've set up the ground work for sports relevance measurement in NLG systems through the development of SOCCER framework that will allow researchers and practitioners to develop innovative relevance measurement heuristics that can fulfill the requirements of stakeholders.

Bibliography

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, *18*(3), 91-93.

Albert, J. (2010). Sabermetrics: The past, the present, and the future. *Mathematics and sports*(43), 15.

Arnold, T. & Godbey, J. M. (2011, /01/06). *Introducing linear regression: An example using basketball statistics* (Rapport). Verkregen van `https://papers.ssrn.com/abstract=1736184` doi: 10.2139/ssrn.1736184

Asif, R., Zaheer, M. T., Haque, S. I. & Hasan, M. A. (2016). Football (soccer) analytics: A case study on the availability and limitations of data for football analytics research. *International Journal of Computer Science and Information Security*, *14*(11), 516-518.

Beneventano, P., Berger, P. D. & Weinberg, B. D. (2012). Predicting run production and run prevention in baseball: the impact of sabermetrics. *Int J Bus Humanit Technol*, *2*(4), 67-75.

Bhattacharjee, D. & Talukdar, P. (2019). Predicting outcome of matches using pressure index: evidence from twenty20 cricket. *Communications in Statistics-Simulation and Computation*, 1-13.

Blandford, A., Furniss, D. & Makri, S. (2016). Qualitative hci research: Going behind the scenes.. doi: 10.2200/S00706ED1V01Y201602HCI034

Borlund, P. (2003). The concept of relevance in ir. *Journal of the American Society for Information Science and Technology*, *54*(10), 913-925. Verkregen van `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.10286` doi: 10.1002/asi.10286

Bornn, L., Cervone, D. & Fernandez, J. (2018). Soccer analytics: Unravelling the complexity of "the beautiful game". *Significance*, *15*(3), 26-29. Verkregen van `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2018.01146.x` doi: 10.1111/j.1740-9713.2018.01146.x

Chang, J. & Zenilman, J. (2013). A study of sabermetrics in major league baseball: The impact of moneyball on free agent salaries. *Unpublished dissertation, Washington University in St.Louis*.

Cicoria, S., Sherlock, J., Muniswamaiah, M. & Clarke, L. (2014). Classification of titanic passenger data and chances of surviving the disaster. In *Proceedings of student-faculty research day, csis* (p. 1-6).

Coleman, B. J. (2012). Identifying the "players" in sports analytics research. *Interfaces*, *42*(2), 109-118.

Cotta, L., de Melo, P., Benevenuto, F. & Loureiro, A. A. (2016). Using fifa soccer video game data for soccer analytics. In *Workshop on large scale sports analytics.*

Dale, R. & Mellish, C. (1998). Towards evaluation in natural language generation. In *In proceedings of first international conference on language resources and evaluation.*

Drazan, J. F., Loya, A. K., Horne, B. D. & Eglash, R. (2017). From sports to science: Using basketball analytics to broaden the appeal of math and science among youth. In *Conference paper presented at rensselear polytechnic institute, ny. retrieved from https://www. researchgate. net/publication/314263728.*

Dörr, K. N. & Hollnbuchner, K. (2017). Ethical challenges of algorithmic journalism. *Digital journalism*, *5*(4), 404-419.

Epl contracts (Dl. 2020) (nr. Aug 11,). (z. j.). Verkregen van `https://www.spotrac.com/epl/contracts/sort-value/limit-2000/`

fact | definition of fact in english by lexico dictionaries (Dl. 2019) (nr. Aug 12,). (z. j.). Verkregen van `https://www.lexico.com/en/definition/fact`

Fifa world cup™ (Dl. 2019) (nr. Aug 8,). (z. j.). Verkregen van `http://www.fifa.com/aboutfifa/worldcup/index.html`

Frankel, M. J. (2012). Secret sabermetrics: trade secret protection in the baseball analytics field. *Alb.Gov't L.Rev.*, *5*, 240.

Franks, A., Miller, A., Bornn, L. & Goldsberry, K. (2015). Counterpoints: Advanced defensive metrics for nba basketball. In *9th annual mit sloan sports analytics conference, boston, ma.*

Gangal, A., Talnikar, A., Dalvi, A., Zope, V. & Kulkarni, A. (2015). Analysis and prediction of football statistics using data mining techniques. *International Journal of Computer Applications*, *975*, 8887.

ge Yao Jianmin Zhang Xiaojun Wan Jianguo Xiao Institute of Computer Science, J., Technology, University, P., Beijing, ., of Computational Linguistics, C. T. M. K. L., University, P. & yaojinge and zhangjianmin2015 and wanxiaojun and xiaojianguo@pku.edu.cn, C. (2017, September,). Content selection for real-time sports news construction from commentary texts. *Proceedings of the 10th International Conference on Natural Language Generation*, 31–40. Verkregen van `https://www.aclweb.org/anthology/W17-3504` doi: 10.18653/v1/W17-3504

Gowda, M., Dhekne, A., Shen, S., Choudhury, R. R., Yang, X., Yang, L., ... Essanian, A. (2017). Bringing iot to sports analytics. In (p. 499–513). Berkeley, CA, USA: USENIX Association. Verkregen van `http://dl.acm.org/citation.cfm?id=3154630.3154672`

Greisdorf, H. (2000). Relevance: An interdisciplinary and information science perspective. *InformingSciJ*, *3*, 67-72. doi: 10.28945/579

Grez, M. (2019, June 1,). *Liverpool beat tottenham hotspur in champions league final.* Verkregen van `https://edition.cnn.com/2019/06/01/football/champions-league-final-result-liverpool-tottenham-spt-intl/index.html`

Gwet, K. L. (2011). On the krippendorff's alpha coefficient. *Manuscript submitted for publication.Retrieved October*, *2*(2011), 2011.

Halvorsen, P., Sægrov, S., Mortensen, A., Kristensen, D. K. C., Eichhorn, A., Stenhaug, M., ... Johansen, D. (2013). Bagadus: An integrated system for arena sports analytics: A soccer case study. In (p. 48–59). New York, NY, USA: ACM. Verkregen van `http://doi.acm.org/10.1145/2483977.2483982` doi: 10.1145/2483977.2483982

Hamdad, L., Benatchba, K., Belkham, F. & Cherairi, N. (2018). Basketball analytics. data mining for acquiring performances. In *Ifip international conference on computational intelligence and its applications* (p. 13-24). Springer.

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, *1*(1), 77-89.

Hoeve, L. T. V. (2017). *Geovisual football analytics: towards the development of an interactive visual interface for football coaches, analysts and players* (Academisch proefschrift).

Hossin, M. & Sulaiman, M. N. (2015). A review on evaluation metrics for data classifica-

tion evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1.

Hrovat, G., Jr, I. F., Yermak, K., Stiglic, G. & Fister, I. (2015). Interestingness measure for mining sequential patterns in sports. *Journal of Intelligent and Fuzzy Systems*, *29*(5), 1981-1994.

Infographic: The financial mismatch in european cup football (Dl. 2020) (nr. Aug 6,). (z. j.). Verkregen van `https://www.statista.com/chart/13892/price-money-europa-league-vs-champions-league/`

Joshi, A., Kale, S., Chandel, S. & Pal, D. K. (2015). Likert scale: Explored and explained. *Current Journal of Applied Science and Technology*, 396-403.

Kanerva, J., Rönnqvist, S., Kekki, R., Salakoski, T. & Ginter, F. (2019, Oct 4,). Template-free data-to-text generation of finnish sports news. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Verkregen van `https://arxiv.org/abs/1910.01863`

Kitchenham, B. & Charters, S. (2009). *Guidelines for performing systematic literature reviews in software engineering* (Rapport). CRD. Verkregen van `https://www.york.ac.uk/crd/guidance/`

Kontonasios, K.-N., Spyropoulou, E. & Bie, T. D. (2012). Knowledge discovery interestingness measures based on unexpectedness. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(5), 386-399. Verkregen van `https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1063` doi: 10.1002/widm.1063

Kubatko, J., Oliver, D., Pelton, K. & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, *3*(3).

Kubo, M., Sasano, R., Takamura, H. & Okumura, M. (November 17, 2013). Generating live sports updates from twitter by finding good reporters. In *Web intelligence* (Dl. 1, p. 527-534). IEEE Computer Society. Verkregen van `http://dl.acm.org/citation.cfm?id=2568488.2568811` doi: 10.1109/WI-IAT.2013.74

Kumar, G. (2013). Machine learning for soccer analytics. *KU Leuven*.

Kutlak, R., van Deemter, K. & Mellish, C. (2016). Production of referring expressions for an unknown audience: A computational model of communal common ground. *Frontiers in Psychology*, *7*, 1275. Verkregen van `https://www.frontiersin.org/article/10.3389/fpsyg.2016.01275`

Lampouras, G. & Androutsopoulos, I. (2018). Producing compact texts with integer linear programming in concept-to-text generation. *arXiv preprint arXiv:1811.00051*. Verkregen van `https://arxiv.org/abs/1811.00051v1`

Lee, C. G., Krahmer, E. & Wubben, S. (2017, Jan 1,). *Pass: A dutch data-to-text system for soccer, targeted towards specific audiences.* Verkregen van `https://www.openaire.eu/search/publication?articleId=narcis_____::5d506d6bfec19d96feea75e679f63cae`

Lenca, P., Vaillant, B., Meyer, P. & Lallich, S. (2007). Association rule interestingness measures: Experimental and theoretical studies. *Quality Measures in Data Mining*, *43*, 51-76. doi: 10.1007/978-3-540-44918-8

Leppänen, L., Munezero, M., Granroth-Wilding, M. & Toivonen, H. (2017). Data-driven news generation for automated journalism. In (p. 188–197). Santiago de Compostela, Spain: Association for Computational Linguistics.

Li, C., Su, Y., Qi, J. & Xiao, M. (2019). Using gan to generate sport news from live game

stats. In *International conference on cognitive computing* (p. 102-116). Springer.

Macdonald, B. (2012). An expected goals model for evaluating nhl teams and players. In *Proceedings of the 2012 mit sloan sports analytics conference, http://www. sloansportsconference. com.*

Manoli, E. A. (2019, May 31,). *Champions league final: how money buys success on the pitch.* Verkregen van http://theconversation.com/champions-league-final-how-money-buys-success-on-the-pitch-118087

March, S. T. & Smith, G. F. (1995, December 1,). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251-266. Verkregen van http://www.sciencedirect.com/science/article/pii/0167923694000412 doi: 10.1016/0167-9236(94)00041-2

McCrudden, M. T. & Schraw, G. (2007, Jan 1,). Relevance and goal-focusing in text processing. *Educational Psychology Review*, *19*(2), 113-139. Verkregen van https://www.jstor.org/stable/23363938 doi: 10.1007/s10648-006-9010-7

McMahon, B. (2019, 30 May,). *2019 champions league final, tottenham hotspur vs. liverpool: What you need to know.* Verkregen van https://www.forbes.com/sites/bobbymcmahon/2019/05/30/2019-champions-league-final-tottenham-hotspur-vs-liverpool-what-you-need-to-know/

Miani, R. G. L. & Junior, E. R. H. (2018, 2018). Eliminating redundant and irrelevant association rules in large knowledge bases. In *Iceis* (p. 17-28). Funchal, Madeira, Portugal.

Mizzaro, S. (1997, Sep 1,). Relevance: The whole history. *Journal of the American Society for Information Science (1986-1998)*, *48*(9), 810-832. Verkregen van https://search.proquest.com/docview/216906153 doi: AID-ASI6>3.0.CO;2-U

Myers, M. D. & Newman, M. (2007). *The qualitative interview in is research: Examining the craft* (Dl. 17) (nr. 1). Verkregen van http://www.sciencedirect.com/science/article/pii/S1471772706000352 (ID: 272182) doi: //doi.org/10.1016/j.infoandorg.2006.11.001

Nguyen, Q., Cao, T., Nguyen, H. & Hagino, T. (August 2012). Towards efficient sport data integration through semantic annotation. In (p. 99-106). doi: 10.1109/KSE.2012.21

Nguyen, Q.-M., Cao, T.-D. & Nguyen, T.-T. (2015). A novel approach for automatic extraction of semantic data about football transfer in sport news. *International Journal of Pervasive Computing and Communications*, *11*, 233-252.

Nichols, J., Mahmud, J. & Drews, C. (February 14, 2012). Summarizing sporting events using twitter. In *Intelligent user interfaces iui logo* (p. 189-198). Lisbon, Portugal: ACM. Verkregen van http://dl.acm.org/citation.cfm?id=2166966.2166999 doi: 10.1145/2166966.2166999

Nie, F., Wang, J., Yao, J.-G., Pan, R. & Lin, C.-Y. (2018). Operations guided neural networks for high fidelity data-to-text generation. In (p. 3879–3889). Brussels, Belgium: Association for Computational Linguistics. doi: 10.18653/v1/D18-1422

Novikova, J., Dušek, O., Curry, A. C. & Rieser, V. (2017). Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Ogus, S. (2014). *Sportiq announces a game changing real-time basketball analytics platform* (Dl. 7). Verkregen van https://www.sporttechie.com/sportiq-announces-a-game-changing-real-time-basketball-analytics-platform/

Passfield, L. & Hopker, J. G. (2017). A mine of information: can sports analytics provide wisdom from your data? *International journal of sports physiology and performance*,

12(7), 851-855.

Patro, S. & Sahu, K. K. (2015). Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Dubourg, V. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825-2830.

Pu, P., Chen, L. & Hu, R. (2011). A user-centric evaluation framework for recommender systems. In (p. 157–164). New York, NY, USA: ACM. Verkregen van `http://doi.acm.org/10.1145/2043932.2043962` doi: 10.1145/2043932.2043962

Raynor, D. (2019, May 31,). *Champions league final stats: Sadio mane can make history in madrid.* Verkregen van `https://www.liverpoolfc.com/news/first-team/352264-champions-league-final-stats-sadio-mane-history-madrid`

Reiter, E. & Dale, R. (2000). *Building natural language generation systems.* New York, NY, USA: Cambridge University Press.

Renjun, T., Ke, Z., Shenruoyang, N., Minghao, Y., Hui, Z., Qingjie, Z., ... Jianhua, T. (2016). Football news generation from chinese live webcast script. *Natural Language Understanding and Intelligent Applications*, 778-789. Verkregen van `https://link-springer-com.proxy.library.uu.nl/chapter/10.1007/978-3-319-50496-4_70`

Ruiz, H., Power, P., Wei, X. & Lucey, P. (2017). "the leicester city fairytale?": Utilizing new soccer analytics tools to compare performance in the 15/16 and 16/17 epl seasons. In (p. 1991–2000). New York, NY, USA: ACM. Verkregen van `http://doi.acm.org/10.1145/3097983.3098121` doi: 10.1145/3097983.3098121

Schober, P., Boer, C. & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763-1768.

Schultze, U. & Avital, M. (2011). *Designing interviews to generate rich data for information systems research* (Dl. 21) (nr. 1). Verkregen van `http://www.sciencedirect.com/science/article/pii/S1471772710000412` (ID: 272182) doi: //doi.org/10.1016/j.infoandorg.2010.11.001

Schumaker, R. P., Solieman, O. K. & Chen, H. (2010). Research in sports statistics. In (p. 29-44). Springer.

Shah, P. & Shah, M. (2014). Form-a new cricket statistics. *American Journal of Sports Science*, *2*(3), 53-55.

Shaharanee, I. N. M. (2012). Quality and interestingness of association rules derived from data mining of relational and semi-structured data. *Quality and interestingness of association rules derived from data mining of relational and semi-structured data*.

Sharifi, B., Hutton, M.-A. & Kalita, J. (2010). Summarizing microblogs automatically. In (p. 685–688). Stroudsburg, PA, USA: Association for Computational Linguistics. Verkregen van `http://dl.acm.org/citation.cfm?id=1857999.1858099`

Slater, M. (2019, June 7,). *Champions league final breaks bt sport audience record.* Verkregen van `https://www.independent.co.uk/sport/football/european/liverpool-tottenham-champions-league-final-bt-sport-viewing-figures-ucl-a8948791.html`

Spearman, W. (2018). Beyond expected goals. In *Proceedings of the 12th mit sloan sports analytics conference* (p. 1-17).

Sperber, D. & Wilson, D. (1986). *Relevance: Communication and cognition.* Cambridge, MA, USA: Harvard University Press.

Sportsml - iptc (Dl. 2020) (nr. Jun 22,). (z. j.). Verkregen van https://iptc.org/standards/sportsml-g2/

Stein, M., Häußler, J., Jäckle, D., Janetzko, H., Schreck, T. & Keim, D. (2015, Oct 20,). Visual soccer analytics: Understanding the characteristics of collective team movement based on feature-driven analysis and abstraction. *ISPRS International Journal of Geo-Information*, *4*(4), 2159-2184. Verkregen van https://www.openaire.eu/search/publication?articleId=dedup_wf_001::f142b27e38e85f869ffafa49a11145f0 doi: 10.3390/ijgi4042159

Stensland, H. K., Gaddam, V. R., Tennøe, M., Helgedagsrud, E., Næss, M., Alstad, H. K., ... Johansen, D. (2014, January). Bagadus: An integrated real-time system for soccer analytics. *ACM Trans. Multimedia Comput. Commun. Appl.*, *10*(1s), 14:1–14:21. Verkregen van http://doi.acm.org/10.1145/2541011 doi: 10.1145/2541011

Sukumar, S. S. (2019). *Moneyball or moneyfall? current state of analytics in soccer and what future holds* (Academisch proefschrift).

Tagawa, Y. & Shimada, K. (2018). *Sports game summarization based on sub-events and game-changing phrases.* Verkregen van https://link-springer-com.proxy.library.uu.nl/chapter/10.1007/978-3-319-70636-8_5 doi: 10.1007/978-3-319-70636-8_5

Tyler, B. D., Morehead, C. A., Cobbs, J. & DeSchriver, T. D. (2017). What is rivalry? old and new approaches to specifying rivalry in demand estimations of spectator sports. *Sport Marketing Quarterly*, *26*(4).

van Deemter, K., Theune, M. & Krahmer, E. (2005, Mar 1). Real versus template-based natural language generation: A false opposition? *Computational Linguistics*, *31*(1), 15-24. Verkregen van http://www.mitpressjournals.org/doi/abs/10.1162/0891201053630291 doi: 10.1162/0891201053630291

van Wijk, N. (2012). Soccer analytics. *Predicting the outcome of soccer matches.Master's thesis, VU University Amsterdam, Amsterdam, The Netherlands*.

Verma, A. & Izadi, M. (2016). Cricket prognostic system: A framework for real-time analysis in odi cricket. In *Kdd workshop on large-scale sports analytics.*

Vinué, G. & Epifanio, I. (2017). Archetypoid analysis for sports analytics. *Data Mining and Knowledge Discovery*, *31*(6), 1643-1677.

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering.* Berlin Heidelberg: Springer-Verlag. Verkregen van https://www.springer.com/gp/book/9783662438381

Wilson, J. (2019, May 31,). *Champions league final battle on flanks will underline importance of full-backs | jonathan wilson.* Verkregen van https://www.theguardian.com/football/blog/2019/may/30/champions-league-final-full-backs-liverpool-tottenham

Wohlin, C. (May 13, 2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In (p. 1-10). ACM. Verkregen van http://dl.acm.org/citation.cfm?id=2601268 doi: 10.1145/2601248.2601268

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B. & Wesslén, A. (2012). *Experimentation in software engineering.* Springer Science and Business Media.

Yang, Y. S. (2015). Predicting regular season results of nba teams based on regression analysis of common basketball statistics. *Undergraduate Thesis, UC Berkeley*.

Zhang, J., ge Yao, J. & Wan, X. (August, 2016). Towards constructing sports news from live text commentary. In (p. 1361–1371). Berlin, Germany: Association for

Computational Linguistics. doi: 10.18653/v1/P16-1129

Ángel García, M. (2019, June 2,). *international football the cleanest champions league final in history.* Verkregen van https://www.marca.com/en/football/international-football/2019/06/02/5cf30da2268e3e116a8b4617.html

Appendix A
Literature review protocol

The literature search is performed during 14 October 2019 - To be filled 2019.

A.1 Sports Analytics

A.1.1 Goal

As sports facts are grounded by the analytics behind the sports, it was considered important to identify such statistics.

A.1.2 Search strategy

Search for statistics related to the soccer, football, baseball, basketball and cricket as they are part of the most popular sports. For the search strategy, Google Scholar was used.

Google Scholar

Configured to search results from 2007 on wards without patents.

1. "sports analytics" OR "sports statistics"
2. "soccer analytics" OR "soccer statistics"
3. "football analytics" OR "football statistics"
4. "baseball analytics" OR "sabermetrics"
5. "basketball statistics" OR " basketball analytics"
6. "cricket statistics" OR "cricket analytics"

A.1.3 Inclusion criteria

- uses the keywords *statistics* or *analytics* in it's title, abstract or article keywords and is related to sports

A.1.4 Exclusion criteria

- not written in English
- is a duplicate of another publication
- older than 2007
- is not available to Utrecht University online network
- is a video

Data Item	Value	Additional notes
Data Extractor		
Date of Data Extraction		
Title		
Research domain		
Research goal		
What are the results of the paper in relation to the research goal and sports statistics?		
Algorithms employed for calculating sports statistics		
Sport		
Statistics calculated in the paper		
Data involved in the statistics		
Source		
Type		
Availability		
Time range		

Table A.1 – Data extraction form for sports analytics

A.1.5 Data Extraction Strategy

To extract data from the papers we used NVivo 12 and a data extraction form illustrated in Table A.1. For NVivo we used the following nodes to code the information:

- **Algorithms** - here we coded all the algorithms that had a relation to sports statistics
- **Research goal** - this node was used to code information related to the research goal and targets of the research
- **Research result** - here we coded information related to the results of the research
- **Statistics** - here we coded all the statistics that were mentioned in a paper

A.2 Sports in computational linguistics

A.2.1 Goal

This part of the literature review aims to:

1. Explore and enumerate methods and metrics used to select, sort or filter sports facts.
2. Identify state of the art methods for natural language generation and content selection.
3. Identify trends in literature in relation with sports facts.
4. Identify requirements and applications for SOCCER.
5. Identify relevance and related concepts measuring in information systems.
6. Ground practitioner knowledge to the research field and identify research gaps based on their knowledge

A.2.2 Search strategy

Perform the search using queries on the defined search engines using publications relevant to the field of computational linguistics (NLP). All results go through the inclusion/exclusion criteria. Also, if there are too many results for the query, only the first 10 are selected. For the Google search engine we used a filter to retrive results from 2007 onwards. We couldn't use the same filter for ACL Anthology.

A.2.3 Search Engines

1. Google Scholar (https://scholar.google.com)
2. ACL Anthology (https://www.aclweb.org/anthology/)

A.2.4 Search Queries

Google Scholar

Configured to search results from 2007 on wards without patents.

1. "sports facts" "natural language processing" OR "computational linguistics" OR "natural language generation" OR "referring expressions" OR "content determination"
2. "sport news" "natural language processing" OR "computational linguistics" OR "natural language generation" OR "referring expressions" OR "content determination"
3. "sport" "relevance" OR "interestingness" OR "newsworthiness" "natural language processing" OR "computational linguistics" OR "natural language generation" OR "referring expressions" OR "content determination"
4. "interestingness" "sport"
5. "interestingness" "information systems"

ACL Anthology

1. sports facts
2. sport news
3. interestingness

A.2.5 Inclusion criteria

The following criteria has been used for inclusion of search results:

- has focus on natural language processing regarding sports and tackle the task of filtering, summarizing, selecting or ranking text about sports.
- mentions at least one of the keywords *sport*, *interestigness*, *facts*, *relevance* in their abstract.
- is related to sports

- uses heuristics that measure *interestingness*
- uses heuristics that measure *relevance*
- uses heuristics that measure *newsworthiness*

A.2.6 Exclusion criteria

The following criteria has been used for exclusion of search results:

- not written in English
- is a duplicate of another publication
- older than 2007

A.2.7 Data Extraction Strategy

For the data extraction NVivo will be used. As the scope of the literature review is to identify properties, measures and algorithms related to relevance the following coding tree will be used:

1. Algorithm - In this node we will code information related to algorithms that are used in NLP, deal with text related to sports
2. Content Property - In this node we will code information that related to the content properties of SFs and not only.
3. Measurement guideline - In this node we will code knowledge related to how interestingness or relevance is measured in literature.
4. Heuristic - In this node we will code information about how measurement guidelines and content properties are used in different algorithms or papers to calculate the relevance.
5. Requirement - In this node we will code information about the requirements that we find in literature. We will gather requirements related to data, algorithm limits, computational and knowledge.
6. Research goal - In this node we will capture data related to the research goal of the paper
7. Research results - In this node we will capture data related to the research results of the paper

In addition, the findings were linked to NLG in sports, *relevance* and *interestigness* in other domains and *sports analytics*. These are considered fields of research that contribute to this thesis but to which the thesis can add.

Data Extraction Form

Besides NVivo, the following form will be used to extract data from the papers.

Appendix B
Interview protocol

B.1 Interviewees

PO1 has been working in the sports industry for more than 10 years as a project manager and a product owner. PO1 is responsible for the development of sports related products related to data processing or data entry. PO1 is directly involved in the OneLiner NLG project that generates sports facts for different commercial purposes.

PO2 is leading the local sports engineering department. PO2 has been working in the sports industry for about 22 years. PO2 has worked as a data entry operator and moved to a project management role. PO2 enjoys soccer and different winter sports such as skying.

PO3 has been working in the sports industry for about 12 years. PO2 started in the data entry team. Later, Po2 moved to be the product owner of the IT team that handled the sports data processing which also created the OneLiner project.

CE1 has been working as a content editor for around 6 years. CE1 has been working with sports data to create sports facts for different news outlets or other types of clients from the gambling industry.

CE2 works as a content editor and has around 10 years of experience in his current role. CE2 uses sports data to identify facts and figures for clients from media or the betting market. CE2 actively uses the OneLiner project to identify relevant sports facts.

CE3 is a manager of the editorial team and has work for around 24 years in the sports industry. The team transforms sports data into short sentences previewing life commentary and reviews sports event for media outlets. In addition, the team creates biographical information of sports athletes and coaches.

SE1 has worked as a software engineer in the sports industry for more than 12 years. In his job responsibilities, SE1 has worked on different sports related software programs designed for data entry, data processing and data delivery. SE1 is the main contributor to the OneLiner project. SE1 created the current relevance measurements for the OneLiner system.

RE1 is a PhD student that researches the topic of automated journalism. RE1 stated that his goal is to help journalists by helping some of their tasks. RE1 stated that he tries to automate the generation of text related to amateur soccer matches. RE1 is in his forth year of his PhD.

B.2 Informed Consent

This interview is conducted with regards to my master thesis research at Utrecht University. This research is part of an internship at Gracenote. My research covers the topic of sports fact relevance for the news and the betting domains. Through this interview, I am aiming to gain theoretical and practical knowledge about how relevance is measured for sports facts.

All information gathered during the interview will be treated with respect and will only be used for scientific purposes. The interview will be recorded, transcribed and analyzed to draw scientific conclusions. All the information regarding people, companies and examples mentioned in the interview will remain confidential and they will be used only for the purpose of this scientific research. Entities mentioned in the interview will be anonymized to ensure confidentiality. The recording of this interview will be private, it will not be shared with other employees inside or outside Gracenote, nor other organizations. The recordings will be permanently deleted after the research is completed and the concluded results will be used in my thesis.

Further, the interview does not aim to harm you, nor your organization. Therefore, you have all the right to stop the recording or the whole interview at any point, if you feel uncomfortable to continue. Participating in this interview is totally voluntary and only for supporting scientific purposes. Thank you for participating in this research.

If you have read and agree with the above statement, please sign below.

Participant	*Researcher*
Name:	Name
Signature:	Signature:
Date:	Date:
Location:	Location:

B.3 Interview Protocol

**** This page will contain information that can help you answer the questions. Thus, it should be kept in an accessible location.****

SOCCER - Content types that can be part of a sports fact

Event participant - it captures the participants in a sports event participants that are persons such as the athletes, the coaches, referees,etc.
Role - it describes the role of the event participants such as a player, midfielder, etc.
Team - it can be the team that is part of a sports event for which the event participants can belong to.
Location - it captures the place where the event, the action takes place.
Event - It captures events such as a match, a phase, a tournament, a competition, etc.
Action - captures what the fact is about. For example, in soccer it can be passed over a certain period of time, goals, assists, financials such as transfer market activity or payroll, etc.
Statistic - captures the statistical content and the data behind the sports fact. For example, if a fact is about having 70% passing accuracy in the last 10 games, then the statistic is 70% over 10 games and the action is pass accuracy.

Facts:

SF1. „No defender has ever registered more assists in the top division of the English league than Trent Alexander-Arnold's 12 in the season of 2018/2019."
SF2. „Often overlooked is the fact that Spurs was the first British team to win a major European trophy. Back in 1963, Spurs beat Atletico Madrid 5-1 at Feijenoord Stadion to win the now-defunct European Cup Winners' Cup."
SF3. „The Spurs captain, Harry Kane, mustered just 11 touches of the ball in the first half of the Champions League final of 2018/2019 at Wanda Metropolitano, less than every other player in white."

Event Participants: Trent Alexander-Arnold (SF1), Harry Kane (SF3)
Roles: Defender (SF1), Captain (SF3)
Teams: Liverpool (SF1), Tottenham Hotspur (SF2), Atletico Madrid (SF2), Tottenham Hotspur (SF3)
Location: Feijenoord Stadion(SF2), Wanda Metropolitano (SF3)
Event: Premier League season of 2018/2019 (SF1), final of Europeans Cup Winners Cup from 1963 (SF2) , first half of Champions League final from 2018/2019 (SF3)
Action: Assists as a Defender For Season (SF1), First national team to win a major European Trophy (SF2), Ball touches by captain in the first half compared with teammates (SF3)
Statistic: Most (SF1), first to win a trophy (SF2), 5-1 (SF2), 11 (SF3), Least (SF3)

Example

SOCCER - Enjoyability

SOCCER FRAMEWORK		Content Type						
		Event Participant	Role	Team	Location	Event	Action	Statistic
Properties	Enjoyability	☐	☐	☐	☐	☐	☐	☐

Legend: ✓ -> The property can be measured for that content type and the output can influence the relevance of the sports fact.
Definition:

Enjoyability - How enjoyable is a type of content.

E.g. from SF1 a fan would enjoy the player and the action with the statistics as he is a defender.

1. *To calculate the relevance of the enumerated facts based on enjoyability, which of the 7 types of content listed above should be taken into account?*

Ex. I think that SF1 is enjoyable because it discusses an enjoyable player, Trent Alexander, thus I've put the check mark in that column.

2. *In general, to calculate the relevance of sports facts based on enjoyability, which of the 7 types of content listed above should be taken into account?*

*** For each mapping:***

3. *Why do you think the enjoyability of `content type` influences the relevance of a sports fact?*

*** For each missing mapping**

4. *Why do you think the enjoyability of `content type` influences the relevance of a sports fact?*
5. *How would you measure the enjoyability of `content type`?*

Ex. In general, I think that the relevance of a sports fact is influenced by how enjoyable is the event and the action. For example, facts that are about FIFA World Cup are more enjoyable than the ones about a random second League.

SOCCER - Popularity

SOCCER FRAMEWORK		Content Type						
		Event Participant	Role	Team	Location	Event	Action	Statistic
Properties	Popularity	☐	☐	☐	☐	☐	☐	☐

Legend: ✓ -> The property can be measured for that content type and the output can influence the relevance of the sports fact.

Definition:

Popularity - how well known is the content to the sports stakeholders.

1. *To calculate the relevance of the enumerated facts based on popularity, which of the 7 types of content listed above should be taken into account?*
2. *In general, to calculate the relevance of some sports facts based on popularity, which of the 7 types of content listed above should be taken into account?*

SOCCER - Popularity

SOCCER FRAMEWORK		Content Type						
		Event Participant	Role	Team	Location	Event	Action	Statistic
Properties	Popularity	✓	✓	✓	✓	✓	✓	✓

Legend: ✓ -> The property can be measured for that content type and the output can influence the relevance of the sports fact.

Definition:

Popularity - how well known is the content to the sports stakeholders.

Measuring guideline:

Content popularity - It should measure how well known the content is to the stakeholders compared with content of the same type. For example, Cristiano Ronaldo or Lionell Messi are well known players while some are less known. Or, a sports stakeholder might know what an overall ball possession statistic is but the Voronoi statistical model is not that popular.

Certain content is more popular than similar types of content, thus popularity can affect the relevance of a sports fact. As every piece of content can be popular or not, I've marked that the popularity of each content type can influence the relevance of a fact.

**** For each missing mapping: ****

1. *Why do you think the popularity of `content type` does not influence the relevance of a sports fact?*

B.4 Interview Results

B.4.1 Popularity

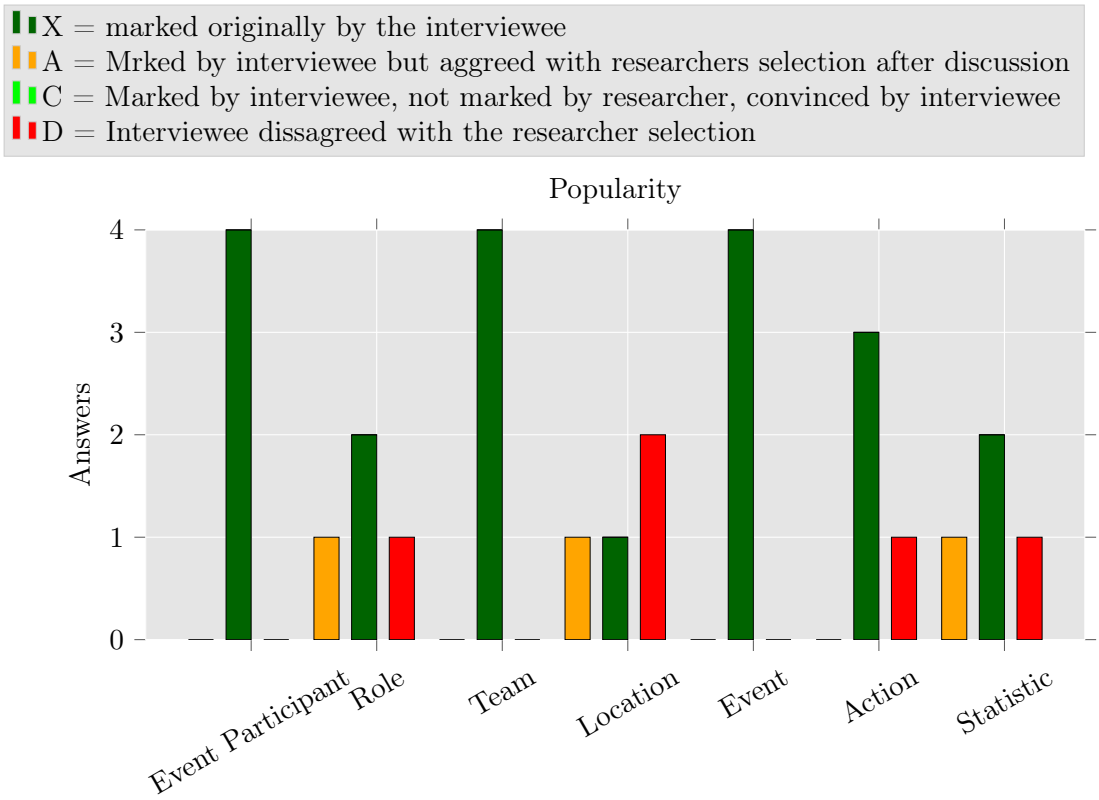

Figure B.1 – Popularity - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Popularity	X	D	X	D	X	D	D	PO1	PO
Popularity	X	X	X	X	X	X	X	CE1	CE
Popularity	X	X	X	D	X	X	A	SE1	SE
Popularity	X	A	X	A	X	X	X	RE1	RE

Table B.1 – Popularity - Interviewee answers

B.4.2 Newsworthiness

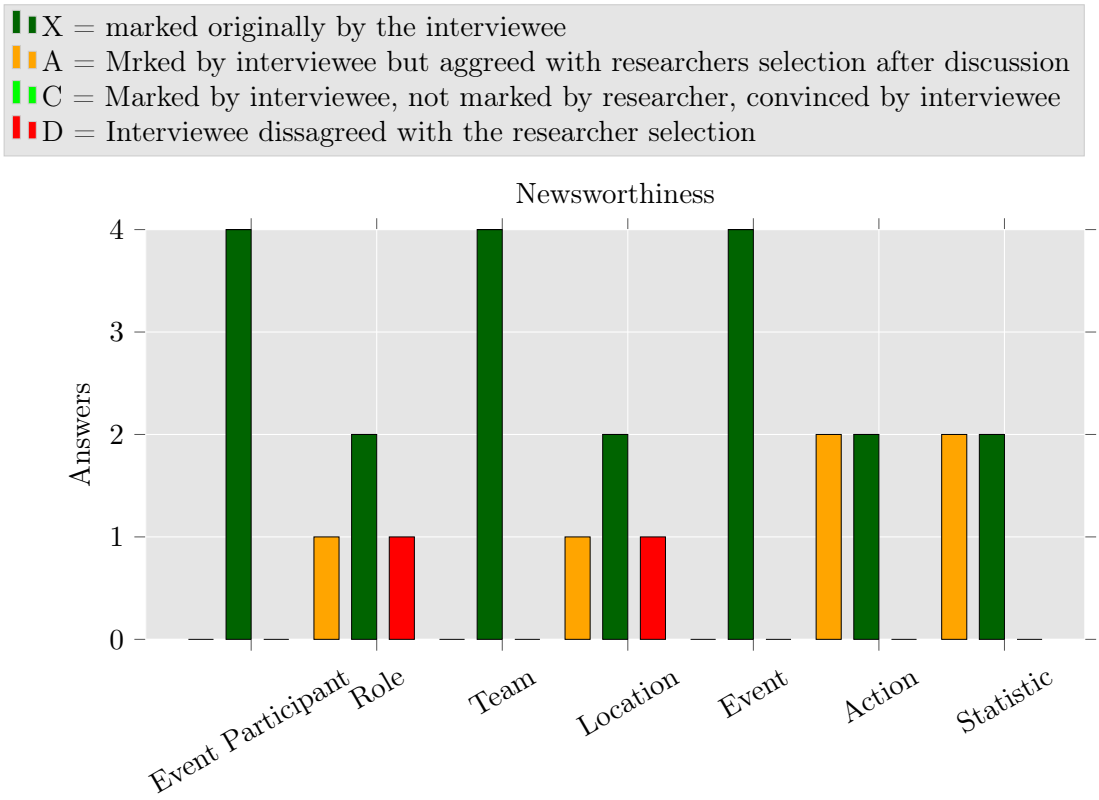

Figure B.2 – Newsworthiness - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Newsworthiness	X	D	X	D	X	A	A	PO1	PO
Newsworthiness	X	X	X	X	X	X	X	CE1	CE
Newsworthiness	X	A	X	X	X	A	A	SE1	SE
Newsworthiness	X	X	X	A	X	X	X	RE1	RE

Table B.2 – Newsworthiness - Interviewee answers

B.4.3 Importance

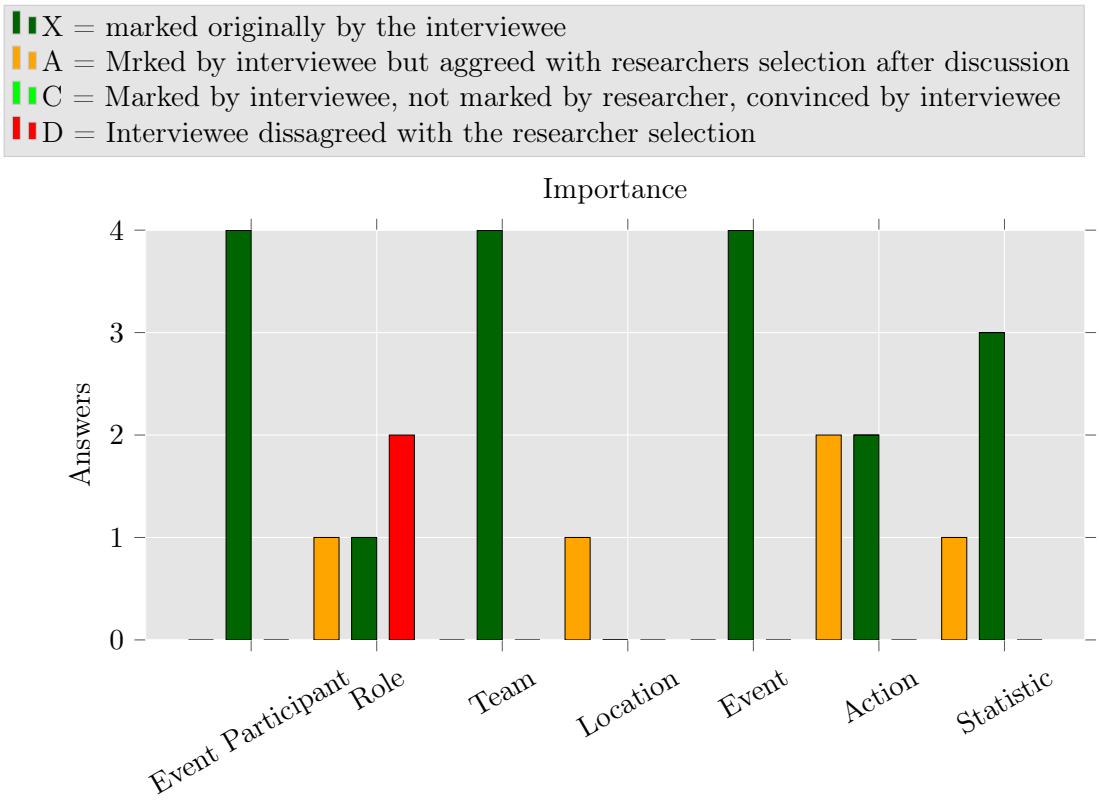

Figure B.3 – Importance - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Importance	X	D	X	NaN	X	A	A	PO2	PO
Importance	X	A	X	NaN	X	X	X	PO3	PO
Importance	X	D	X	NaN	X	A	X	CE2	CE
Importance	X	X	X	A	X	X	X	CE3	CE

Table B.3 – Importance - Interviewee answers

B.4.4 Significance

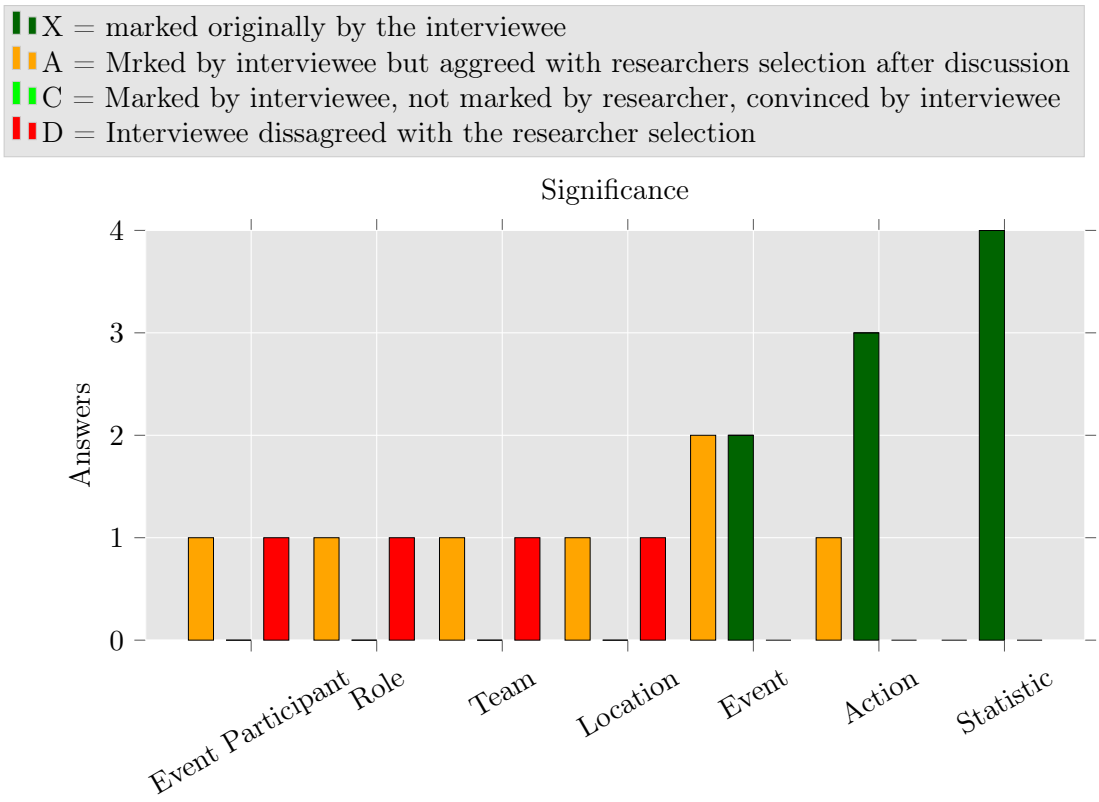

Figure B.4 – Significance - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Significance	NaN	NaN	NaN	NaN	A	A	X	PO1	PO
Significance	D	D	D	D	A	X	X	PO2	PO
Significance	NaN	NaN	NaN	NaN	X	X	X	CE2	CE
Significance	A	A	A	A	X	X	X	RE1	RE

Table B.4 – Significance - Interviewee answers

B.4.5 Unexpectedness

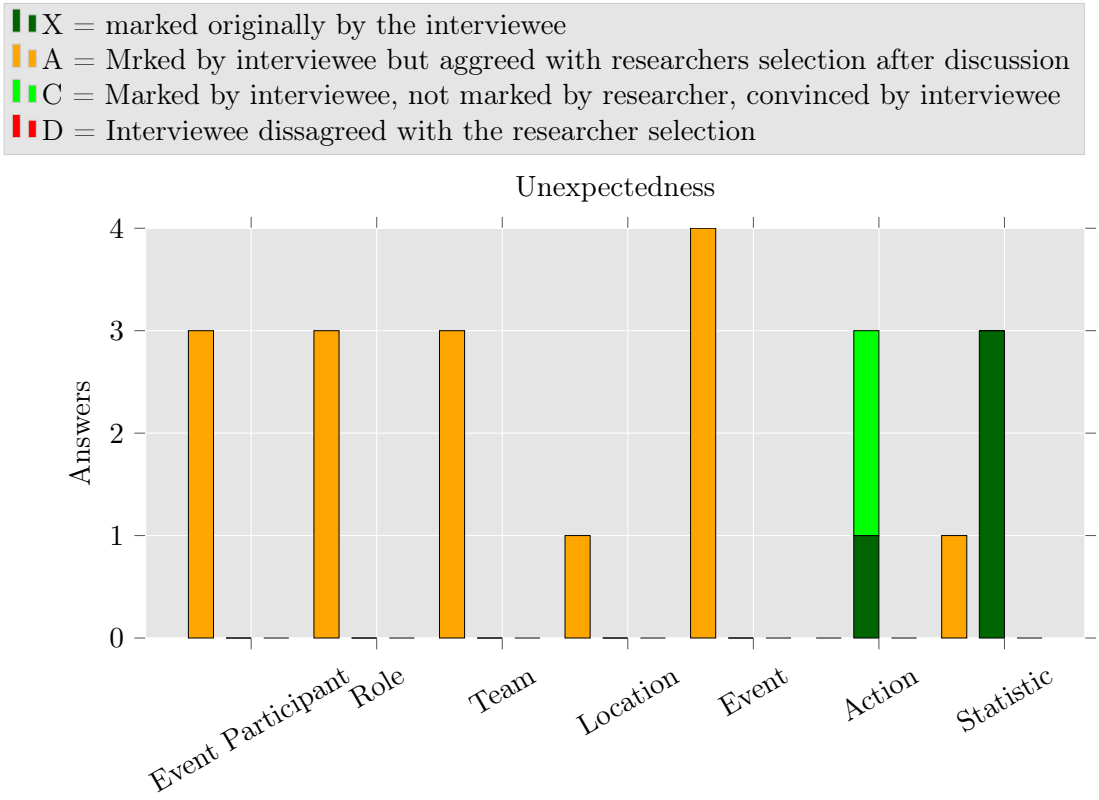

Figure B.5 – Unexpectedness - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Unexpectedness	NaN	A	A	NaN	A	C	X	PO2	PO
Unexpectedness	A	NaN	A	A	A	C	X	CE1	CE
Unexpectedness	A	A	A	NaN	A	NaN	A	CE2	CE
Unexpectedness	A	A	NaN	NaN	A	X	X	RE1	RE

Table B.5 – Unexpectedness - Interviewee answers

B.4.6 Complexity

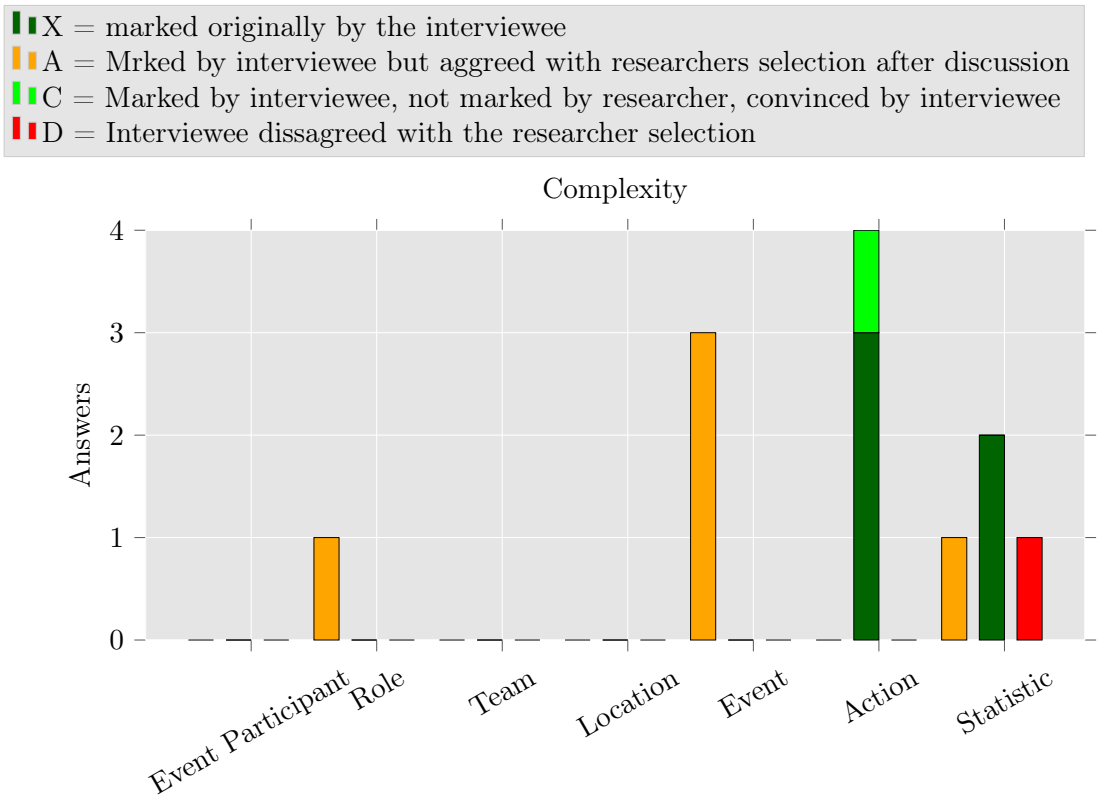

Figure B.6 – Complexity - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Complexity	NaN	NaN	NaN	NaN	NaN	C	X	PO3	PO
Complexity	NaN	NaN	NaN	NaN	A	X	D	CE2	CE
Complexity	NaN	NaN	NaN	NaN	A	X	A	CE3	CE
Complexity	NaN	A	NaN	NaN	A	X	X	SE1	SE

Table B.6 – Complexity - Interviewee answers

B.4.7 Sentiment

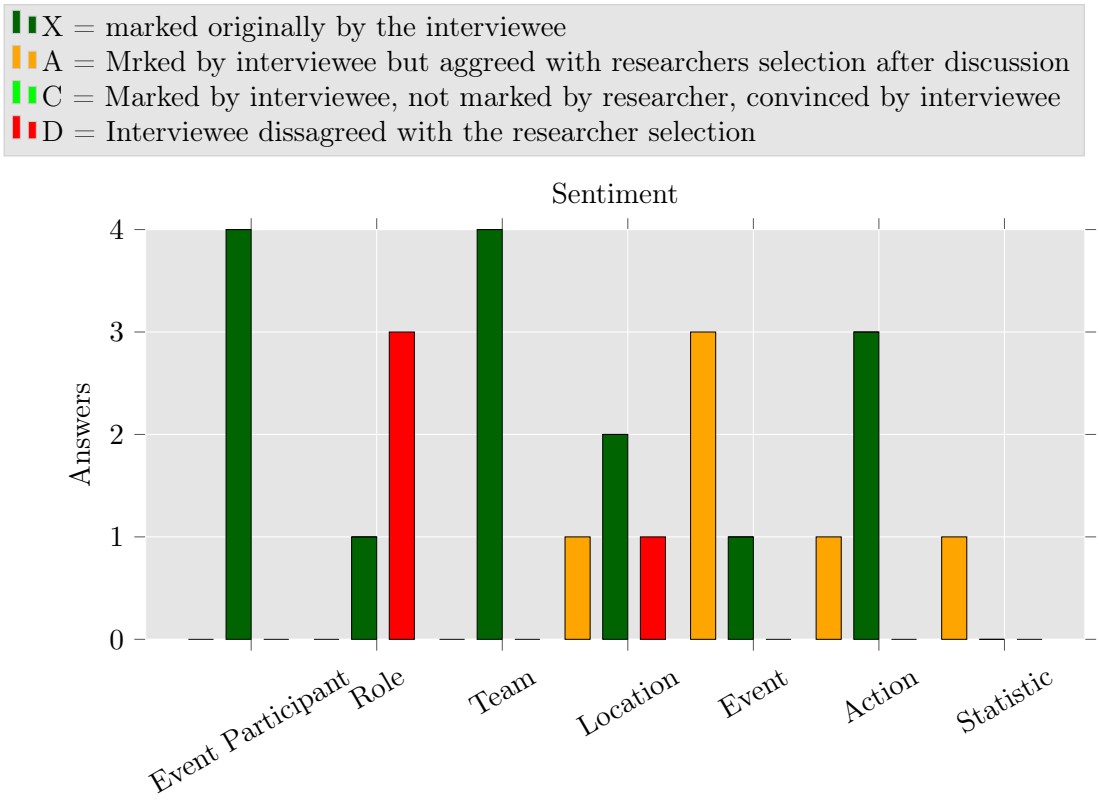

Figure B.7 – Sentiment - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Sentiment	X	D	X	D	X	A	NaN	PO3	PO
Sentiment	X	D	X	A	A	X	NaN	CE3	CE
Sentiment	X	X	X	X	A	X	NaN	SE1	SE
Sentiment	X	D	X	X	A	X	A	RE1	RE

Table B.7 – Sentiment - Interviewee answers

B.4.8 Timeliness

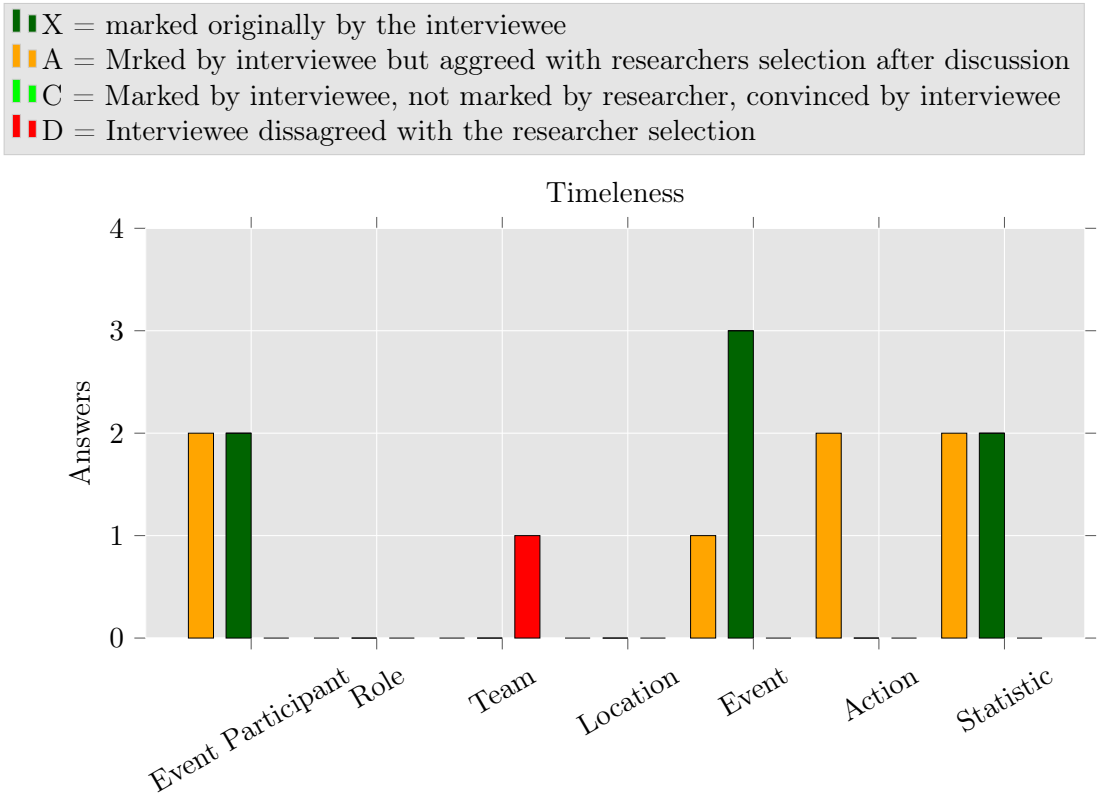

Figure B.8 – Timeliness - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Timeleness	A	NaN	NaN	NaN	X	NaN	A	PO1	PO
Timeleness	A	NaN	NaN	NaN	A	NaN	X	PO2	PO
Timeleness	X	NaN	D	NaN	X	A	X	PO3	PO
Timeleness	X	NaN	NaN	NaN	X	A	A	CE1	CE

Table B.8 – Timeliness - Interviewee answers

B.4.9 Novelty

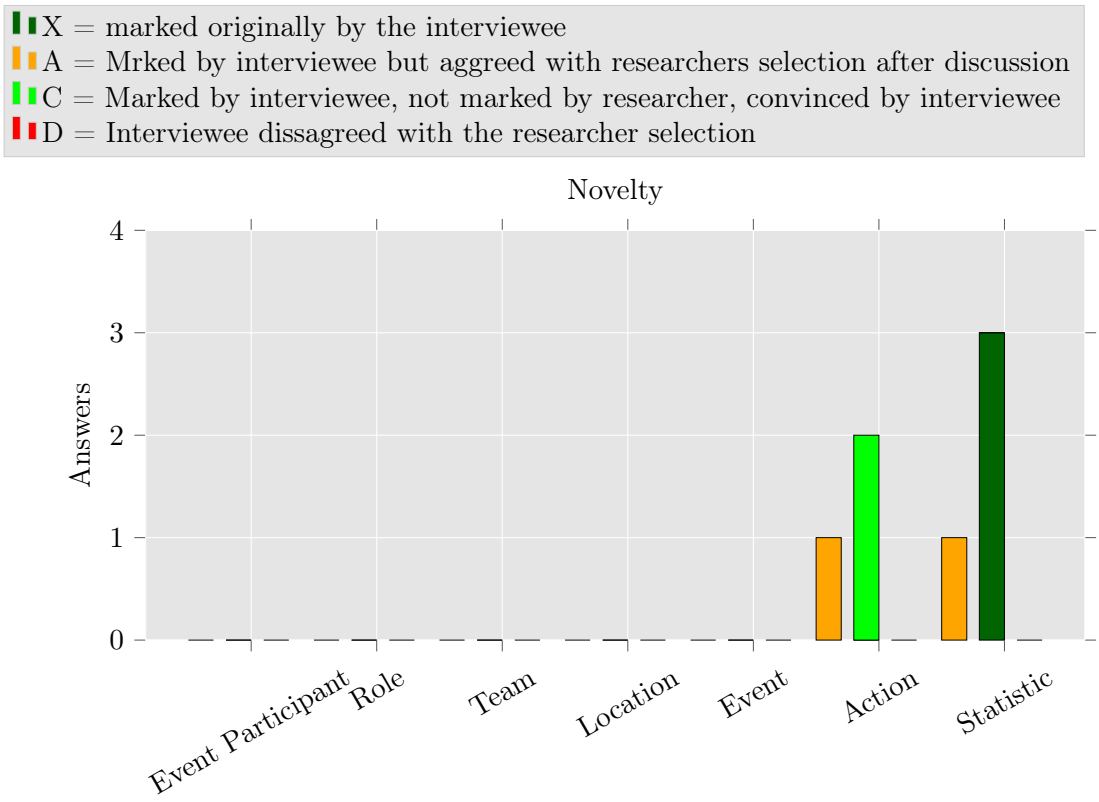

Figure B.9 – Novelty - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Novelty	NaN	NaN	NaN	NaN	NaN	NaN	X	PO1	PO
Novelty	NaN	NaN	NaN	NaN	NaN	C	X	PO3	PO
Novelty	NaN	NaN	NaN	NaN	NaN	C	A	CE1	CE
Novelty	NaN	NaN	NaN	NaN	NaN	A	X	SE1	SE

Table B.9 – Novelty - Interviewee answers

B.4.10 Utility

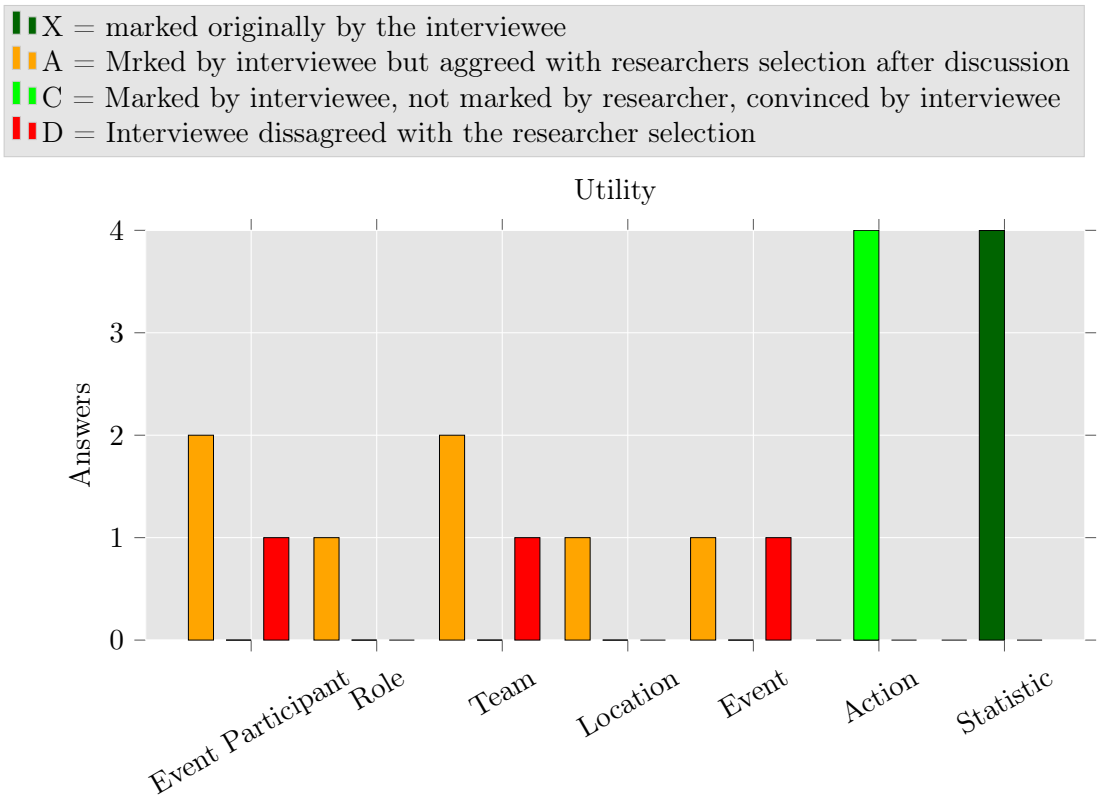

Figure B.10 – Utility - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Utility	A	NaN	A	NaN	D	C	X	PO2	PO
Utility	D	NaN	D	NaN	NaN	C	X	CE2	CE
Utility	A	A	A	A	A	C	X	CE3	CE
Utility	NaN	NaN	NaN	NaN	NaN	C	X	RE1	RE

Table B.10 – Utility - Interviewee answers

B.4.11 Peculiarity

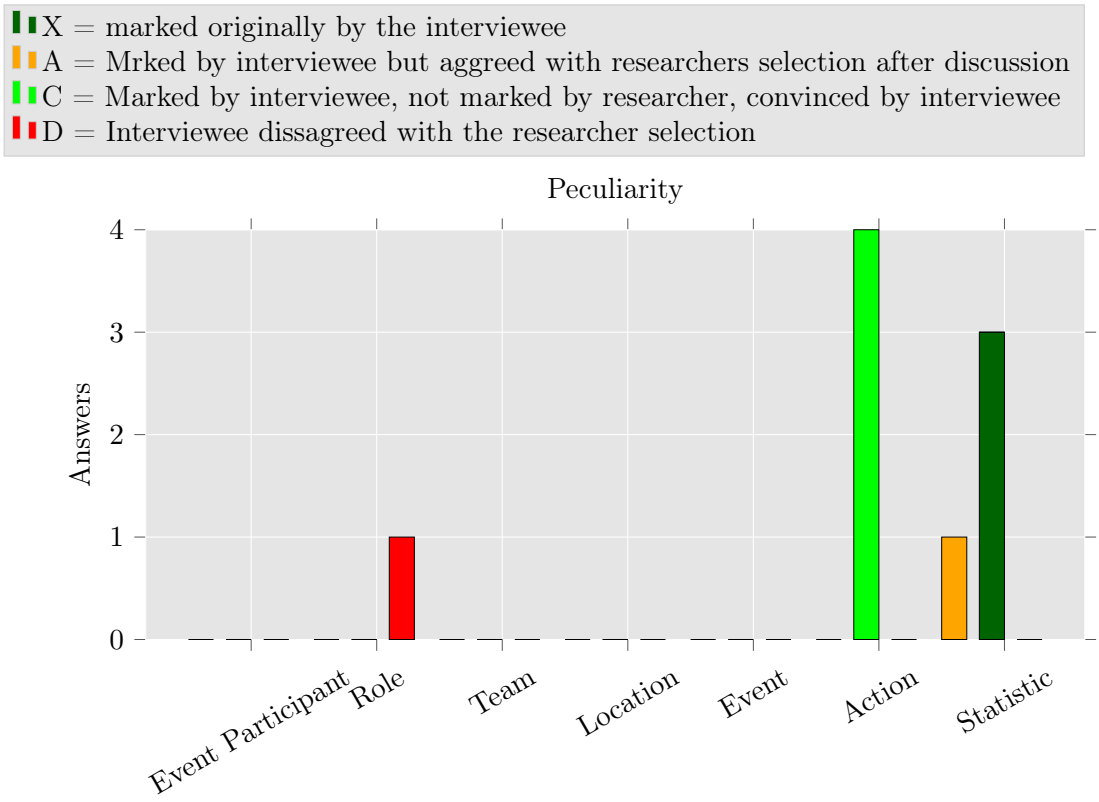

Figure B.11 – Peculiarity - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Peculiarity	NaN	D	NaN	NaN	NaN	C	A	PO1	PO
Peculiarity	NaN	NaN	NaN	NaN	NaN	C	X	PO3	PO
Peculiarity	NaN	NaN	NaN	NaN	NaN	C	X	CE1	CE
Peculiarity	NaN	NaN	NaN	NaN	NaN	C	X	CE3	CE

Table B.11 – Peculiarity - Interviewee answers

B.4.12 Predictability

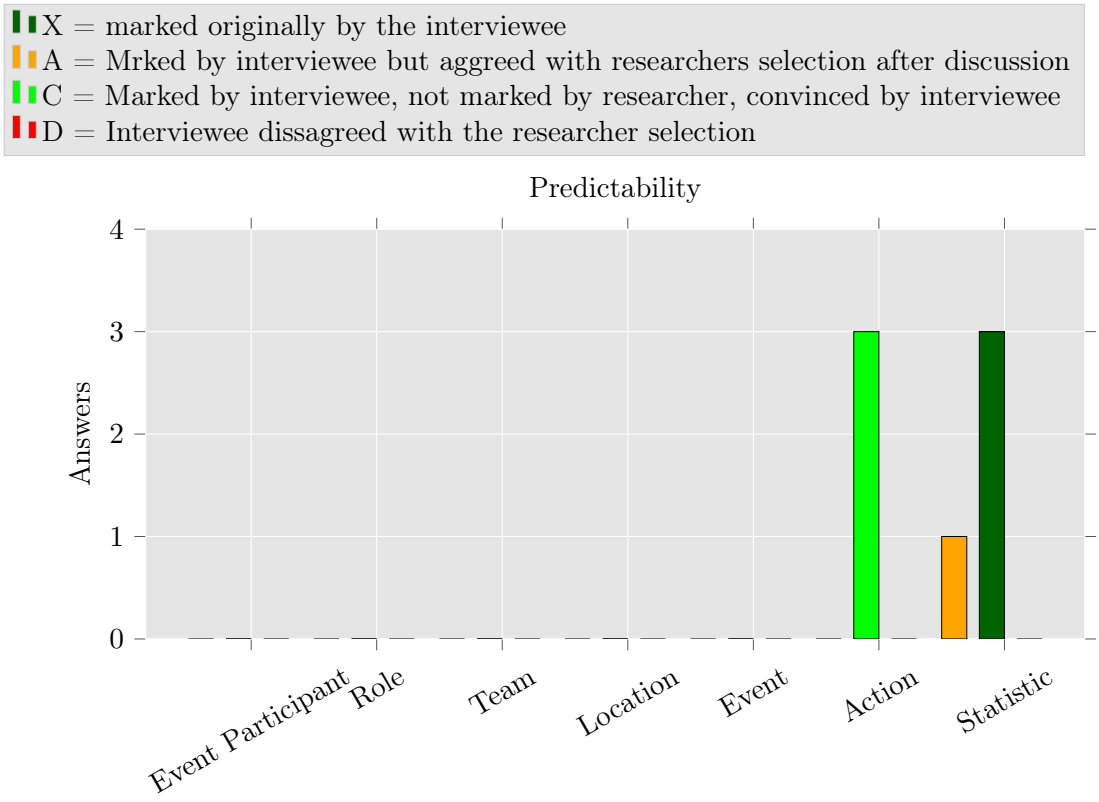

Figure B.12 – Predictability - plot of interviewee answers.

1 Property	Event Participant	Role	Team	Location	Event	Action	Statistic	Interviewee	Group
Predictability	NaN	NaN	NaN	NaN	NaN	NaN	X	PO2	PO
Predictability	NaN	NaN	NaN	NaN	NaN	C	X	CE2	CE
Predictability	NaN	NaN	NaN	NaN	NaN	C	X	CE3	CE
Predictability	NaN	NaN	NaN	NaN	NaN	C	A	SE1	SE

Table B.12 – Predictability - Interviewee answers