Generating and Visualizing a Soccer Knowledge Base

Paul Buitelaar, Thomas Eigner, Greg Gulrajani, Alexander Schutz, Melanie Siegel, Nicolas Weber
Language Technology Lab, DFKI GmbH
Saarbrücken, Germany
{paulb,siegel}@dfki.de

Philipp Cimiano, Günter Ladwig, Matthias Mantel, Honggang Zhu
Institute AIFB, University of Karlsruhe
Karlsruhe, Germany
cimiano@aifb.uni-karlsruhe.de

Abstract

This demo abstract describes the SmartWeb Ontology-based Annotation system (SOBA). A key feature of SOBA is that all information is extracted and stored with respect to the SmartWeb Integrated Ontology (SWIntO). In this way, other components of the systems, which use the same ontology, can access this information in a straightforward way. We will show how information extracted by SOBA is visualized within its original context, thus enhancing the browsing experience of the end user.

1 Introduction

SmartWeb[1] is a multi-modal dialog system, which derives answers from unstructured resources such as the Web, from automatically acquired knowledge bases and from web services.

In this paper we describe the current status of the SmartWeb Ontology-Based Annotation (SOBA) system. SOBA automatically populates a knowledge base by information extraction from soccer match reports as available on the web. The extracted information is defined with respect to SWIntO, the underlying SmartWeb Integrated Ontology (Oberle et al., in preparation) in order to be smoothly integrated into the system.

The ability to extract information and describe it ontologically is a basic requirement for more complex processing tasks such as reasoning and discourse analysis (for related work on ontology-based information extraction see e.g. Maedche et al., 2002; Lopez and Motta, 2004; Müller et al., 2004; Nirenburg and Raskin, 2004).

2 System Overview

The SOBA system consists of a web crawler, linguistic annotation components and a component for the transformation of linguistic annotations into an ontology-based representation.

The web crawler acts as a monitor on relevant web domains (i.e. the FIFA[2] and UEFA[3] web sites), automatically downloads relevant documents from them and sends them to a linguistic annotation web service.

Linguistic annotation and information extraction is based on the Heart-of-Gold (HoG) architecture (Callmeier et al. 2004), which provides a uniform and flexible infrastructure for building multilingual applications that use semantics- and XML-based natural language processing components.

The linguistically annotated documents are further processed by the transformation component, which generates a knowledge base of soccer-related entities (players, teams, etc.) and events (matches, goals, etc.) by mapping annotated entities or events to ontology classes and their properties.

Finally, an automatic hyperlinking component is used for the visualization of extracted entities and events. This component is based on the VieWs system, which was developed independently of SmartWeb (Buitelaar et al., 2005). In what follows we describe the different components of the system in detail.

2.1 Web Crawler

The crawler enables the automatic creation of a football corpus, which is kept up-to-date on a daily basis. The crawler data is compiled from texts, semi-structured data and copies of original

[1] http://www.smartweb-projekt.de/start_en.html

[2] http://fifaworldcup.yahoo.com/

[3] http://www.uefa.com/

HTML documents. For each football match, the data source contains a sheet of semi-structured data with tables of players, goals, referees, etc. Textual data comprise of match reports as well as news articles.

The crawler is able to extract data from two different sources: FIFA and UEFA. Semi-structured data, news articles and match reports covering the WorldCup2006 are identified and collected from the FIFA website. Match reports and news articles are extracted from the UEFA website. The extracted data are labeled by IDs that match the filename. The IDs are derived from the corresponding URL and are thus unique.

The crawler is invoked continuously each day with the same configuration, extracting only data which is not yet contained in the corpus. In order to distinguish between available new data and data already present in the corpus, the URLs of all available data from the website are matched against the IDs of the already extracted data.

2.2 Linguistic Annotation and Information Extraction

As mentioned before, linguistic annotation in the system is based on the HoG architecture, which provides a uniform and flexible infrastructure for building multilingual applications that use semantics- and XML-based natural language processing components.

For the annotation of soccer game reports, we extended the rule set of the SProUT (Drozdzynski et al. 2004) named-entity recognition component in HoG with gazetteers, part-of-speech and morphological information. SProUT combines finite-state techniques and unification-based algorithms. Structures to be extracted are ordered in a type hierarchy, which we extended with soccer-specific rules and output types.

SProUT has basic grammars for the annotation of persons, locations, numerals and date and time expressions. On top of this, we implemented rules for soccer-specific entities, such as actors in soccer (trainer, player, referee …), teams, games and tournaments. Using these, we further implemented rules for soccer-specific events, such as player activities (shots, headers …), game events (goal, card …) and game results. A soccer-specific gazetteer contains soccer-specific entities and names and is supplemented to the general named-entity gazetteer.

As an example, consider the linguistic annotation for the following German sentence from one of the soccer game reports:

Guido Buchwald wurde 1990 in Italien Weltmeister (Guido Buchwald became world champion in 1990 in Italy)

```
<FS type="player_action">
  <F name="GAME_EVENT">
     <FS type="world champion"/>
  <F name="ACTION_TIME">
     <FS type="1990"/>
  <F name="ACTION_LOCATION">
     <FS type="Italy"/>
  <F name="AGENT">
     <FS type="player">
         <F name="SURNAME">
            <FS type="Buchwald"/>
         <F name="GIVEN_NAME">
            <FS type="Guido"/>
```

2.3 Knowledge Base Generation

The SmartWeb SportEventOntology (a subset of SWIntO) contains about 400 direct classes onto which named-entities and other, more complex structures are mapped. The mapping is represented in a declarative fashion specifying how the feature-based structures produced by SProUT are mapped into structures which are compatible with the underlying ontology. Further, the newly extracted information is also interpreted in the context of additional information about the match in question.

This additional information is obtained by wrapping the semi-structured data on relevant soccer matches, which is also mapped to the ontology. The information obtained in this way about the match in question can then be used as contextual background with respect to which the newly extracted information is interpreted.

The feature structure for *player* as displayed above will be translated into the following F-Logic (Kifer et al. 1995) statements, which are then automatically translated to RDF and fed to the visualization component:

```
soba#player124:sportevent#FootballPlayer
[sportevent#impersonatedBy ->
  soba#Guido_BUCHWALD].


soba#Guido_BUCHWALD:dolce#"natural-person"
[dolce#"HAS-DENOMINATION" ->
 soba#Guido_BUCHWALD_Denomination].

soba#Guido_BUCHWALD_Denomination":dolce#"
natural-person-denomination"
[dolce#LASTNAME -> "Buchwald";
 dolce#FIRSTNAME -> "Guido"].
```

2.4 Knowledge Base Visualization

The generated knowledge base is visualized by way of automatically inserted hyperlink menus for soccer-related named-entities such as players and teams. The visualization component is based on the VIeWs[4] system. VIeWs allows the user to simply browse a web site as usual, but is additionally supported by the automatic hyperlinking system that adds additional information from a (generated) knowledge base.

For some examples of this see the included figures below, which show extracted information for the Panama team (i.e. all of the football players in this team in Figure 1) and for the player Roberto Brown (i.e. his team and events in which he participated in Figure 2).

3 Implementation

All components are implemented in Java 1.5 and are installed as web applications on a Tomcat web server. SOAP web services are used for communication between components so that the system can be installed in a centralized as well as decentralized manner. Data communication is handled by XML-based exchange formats. Due to a high degree of flexibility of components, only a simple configuration over environment variables is needed.

4 Conclusions and Future Work

We presented an ontology-based approach to information extraction in the soccer domain that aims at the automatic generation of a knowledge base from match reports and the subsequent visualization of the extracted information through automatic hyperlinking. We argue that such an approach is innovative and enhances the user experience.

Future work includes the extraction of more complex events, for which deep linguistic analysis and/or semantic inference over the ontology and knowledge base is required. For this purpose we will use an HPSG-based parser that is available within the HoG architecture (Callmeier, 2000) and combine this with a semantic inference approach based on discourse analysis (Cimiano et al., 2005).

Acknowledgements

This research has been supported by grants for the projects SmartWeb (by the German Ministry of Education and Research: 01 IMD01 A) and VIeWs (by the Saarland Ministry of Economic Affairs).

References

Paul Buitelaar, Thomas Eigner, Stefania Racioppa *Semantic Navigation with VIeWs* In: Proc. of the Workshop on User Aspects of the Semantic Web at the European Semantic Web Conference, Heraklion, Greece, May 2005.

Callmeier, Ulrich (2000). *PET – A platform for experimentation with efficient HPSG processing techniques.* In: Natural Language Engineering, 6 (1) UK: Cambridge University Press pp. 99–108.

Callmeier, Ulrich, Eisele, Andreas, Schäfer, Ulrich and Melanie Siegel. 2004. *The DeepThought Core Architecture Framework* In Proceedings of LREC 04, Lisbon, Portugal, pages 1205-1208.

Cimiano, Philipp, Saric, Jasmin and Uwe Reyle. 2005. *Ontology-driven discourse analysis for information extraction,* Data Knowledge Engineering 55(1).

Drozdzynski, Witold, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. 2004. *Shallow processing with unification and typed feature structures – foundations and applications.* Künstliche Intelligenz, 1:17-23.

Kifer, M., Lausen, G. and J.Wu. 1995. *Logical Foundations of Object-Oriented and Frame-Based Languages.* Journal of the ACM 42, pp. 741-843.

Lopez, V. and E. Motta. 2004. *Ontology-driven Question Answering in AquaLog* In Proceedings of 9th International Conference on applications of natural language to information systems.

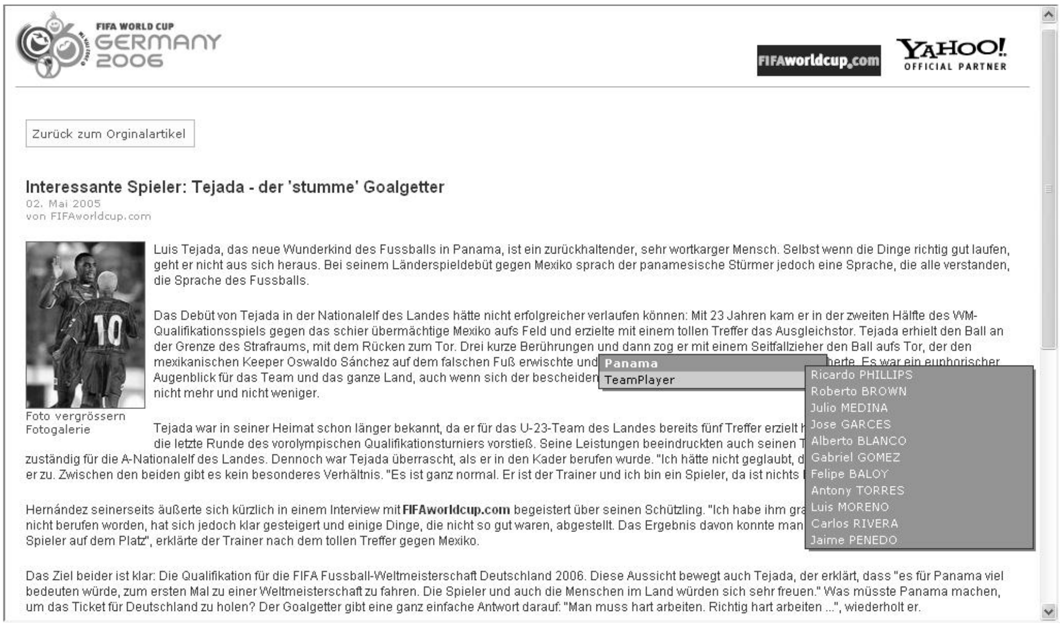
Maedche, Alexander, Günter Neumann and Steffen Staab. 2002. *Bootstrapping an Ontology-Based Information Extraction System.* In: Studies in Fuzziness and Soft Computing, editor J. Kacprzyk. Intelligent Exploration of the Web, Springer.

Müller HM, Kenny EE and PW Sternberg. 2004. *Textpresso: An ontology-based information retrieval and extraction system for biological literature.* PLoS Biol 2: e309.

Nirenburg, Sergei and Viktor Raskin. 2004. *Ontological Semantics.* MIT Press.

Oberle et al. The SmartWeb Integrated Ontology SWIntO, in preparation.

[4] http://views.dfki.de



Zurück zum Orginalartikel

Interessante Spieler: Tejada - der 'stumme' Goalgetter
02. Mai 2005
von FIFAworldcup.com

Luis Tejada, das neue Wunderkind des Fussballs in Panama, ist ein zurückhaltender, sehr wortkarger Mensch. Selbst wenn die Dinge richtig gut laufen, geht er nicht aus sich heraus. Bei seinem Länderspieldebüt gegen Mexiko sprach der panamesische Stürmer jedoch eine Sprache, die alle verstanden, die Sprache des Fussballs.

Das Debüt von Tejada in der Nationalelf des Landes hätte nicht erfolgreicher verlaufen können: Mit 23 Jahren kam er in der zweiten Hälfte des WM-Qualifikationsspiels gegen das schier übermächtige Mexiko aufs Feld und erzielte mit einem tollen Treffer das Ausgleichstor. Tejada erhielt den Ball an der Grenze des Strafraums, mit dem Rücken zum Tor. Drei kurze Berührungen und dann zog er mit einem Seitfallzieher den Ball aufs Tor, der den mexikanischen Keeper Oswaldo Sánchez auf dem falschen Fuß erwischte und herte. Es war ein euphorischer Augenblick für das Team und das ganze Land, auch wenn sich der bescheiden nicht mehr und nicht weniger.

Panama
TeamPlayer

Ricardo PHILLIPS
Roberto BROWN
Julio MEDINA
Jose GARCES
Alberto BLANCO
Gabriel GOMEZ
Felipe BALOY
Antony TORRES
Luis MORENO
Carlos RIVERA
Jaime PENEDO

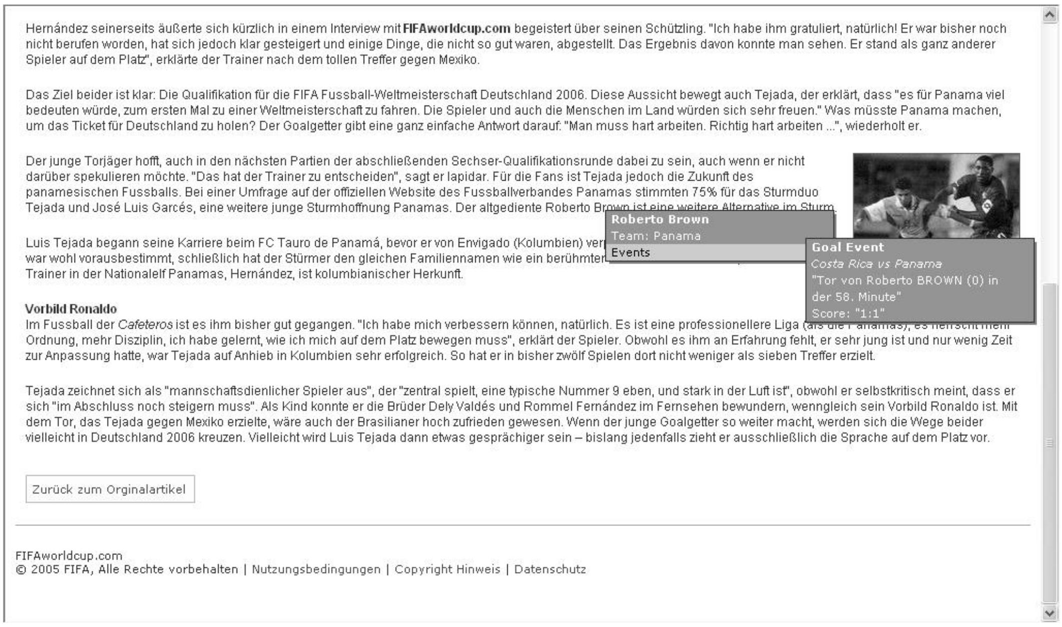
Foto vergrössern
Fotogalerie

Tejada war in seiner Heimat schon länger bekannt, da er für das U-23-Team des Landes bereits fünf Treffer erzielt h die letzte Runde des vorolympischen Qualifikationsturniers vorstieß. Seine Leistungen beeindruckten auch seinen T zuständig für die A-Nationalelf des Landes. Dennoch war Tejada überrascht, als er in den Kader berufen wurde. "Ich hätte nicht geglaubt, d er zu. Zwischen den beiden gibt es kein besonderes Verhältnis. "Es ist ganz normal. Er ist der Trainer und ich bin ein Spieler, da ist nichts

Hernández seinerseits äußerte sich kürzlich in einem Interview mit **FIFAworldcup.com** begeistert über seinen Schützling. "Ich habe ihm gra nicht berufen worden, hat sich jedoch klar gesteigert und einige Dinge, die nicht so gut waren, abgestellt. Das Ergebnis davon konnte man Spieler auf dem Platz", erklärte der Trainer nach dem tollen Treffer gegen Mexiko.

Das Ziel beider ist klar: Die Qualifikation für die FIFA Fussball-Weltmeisterschaft Deutschland 2006. Diese Aussicht bewegt auch Tejada, der erklärt, dass "es für Panama viel bedeuten würde, zum ersten Mal zu einer Weltmeisterschaft zu fahren. Die Spieler und auch die Menschen im Land würden sich sehr freuen." Was müsste Panama machen, um das Ticket für Deutschland zu holen? Der Goalgetter gibt eine ganz einfache Antwort darauf: "Man muss hart arbeiten. Richtig hart arbeiten ...", wiederholt er.

Figure 1: Generated hyperlink on „Panama“ with extracted information on this team



Hernández seinerseits äußerte sich kürzlich in einem Interview mit **FIFAworldcup.com** begeistert über seinen Schützling. "Ich habe ihm gratuliert, natürlich! Er war bisher noch nicht berufen worden, hat sich jedoch klar gesteigert und einige Dinge, die nicht so gut waren, abgestellt. Das Ergebnis davon konnte man sehen. Er stand als ganz anderer Spieler auf dem Platz", erklärte der Trainer nach dem tollen Treffer gegen Mexiko.

Das Ziel beider ist klar: Die Qualifikation für die FIFA Fussball-Weltmeisterschaft Deutschland 2006. Diese Aussicht bewegt auch Tejada, der erklärt, dass "es für Panama viel bedeuten würde, zum ersten Mal zu einer Weltmeisterschaft zu fahren. Die Spieler und auch die Menschen im Land würden sich sehr freuen." Was müsste Panama machen, um das Ticket für Deutschland zu holen? Der Goalgetter gibt eine ganz einfache Antwort darauf: "Man muss hart arbeiten. Richtig hart arbeiten ...", wiederholt er.

Der junge Torjäger hofft, auch in den nächsten Partien der abschließenden Sechser-Qualifikationsrunde dabei zu sein, auch wenn er nicht darüber spekulieren möchte. "Das hat der Trainer zu entscheiden", sagt er lapidar. Für die Fans ist Tejada jedoch die Zukunft des panamesischen Fussballs. Bei einer Umfrage auf der offiziellen Website des Fussballverbandes Panamas stimmten 75% für das Sturmduo Tejada und José Luis Garcés, eine weitere junge Sturmhoffnung Panamas. Der altgediente Roberto Brown ist eine weitere Alternative im Sturm.

Luis Tejada begann seine Karriere beim FC Tauro de Panamá, bevor er von Envigado (Kolumbien) ver war wohl vorausbestimmt, schließlich hat der Stürmer den gleichen Familiennamen wie ein berühmter Trainer in der Nationalelf Panamas, Hernández, ist kolumbianischer Herkunft.

Roberto Brown
Team: Panama
Events

Goal Event
Costa Rica vs Panama
"Tor von Roberto BROWN (0) in der 58. Minute"
Score: "1:1"

Vorbild Ronaldo
Im Fussball der *Cafeteros* ist es ihm bisher gut gegangen. "Ich habe mich verbessern können, natürlich. Es ist eine professionellere Liga Ordnung, mehr Disziplin, ich habe gelernt, wie ich mich auf dem Platz bewegen muss", erklärt der Spieler. Obwohl es ihm an Erfahrung fehlt, er sehr jung ist und nur wenig Zeit zur Anpassung hatte, war Tejada auf Anhieb in Kolumbien sehr erfolgreich. So hat er in bisher zwölf Spielen dort nicht weniger als sieben Treffer erzielt.

Tejada zeichnet sich als "mannschaftsdienlicher Spieler aus", der "zentral spielt, eine typische Nummer 9 eben, und stark in der Luft ist", obwohl er selbstkritisch meint, dass er sich "im Abschluss noch steigern muss". Als Kind konnte er die Brüder Dely Valdés und Rommel Fernández im Fernsehen bewundern, wenngleich sein Vorbild Ronaldo ist. Mit dem Tor, das Tejada gegen Mexiko erzielte, wäre auch der Brasilianer hoch zufrieden gewesen. Wenn der junge Goalgetter so weiter macht, werden sich die Wege beider vielleicht in Deutschland 2006 kreuzen. Vielleicht wird Luis Tejada dann etwas gesprächiger sein – bislang jedenfalls zieht er ausschließlich die Sprache auf dem Platz vor.

Zurück zum Orginalartikel

FIFAworldcup.com
© 2005 FIFA, Alle Rechte vorbehalten | Nutzungsbedingungen | Copyright Hinweis | Datenschutz

Figure 2: Generated hyperlink on „Roberto Brown“ with extracted information on his team and events in which he participated