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 Information Extraction System (SOBIE). A key feature of SOBIE is that all information is extracted and stored with respect to the SmartWeb ontology. 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 SOBIE is visualized within its original context, thus enhancing the browsing experience of the end user.

1 Introduction

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

The SmartWeb component that we describe here automatically populates a knowledge base by information extraction from soccer game reports as found on the web. The extracted information is defined with respect to the underlying SmartWeb ontology in order to be smoothly integrated into the system - 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). The ability to extract information and describe it ontologically is a basic requirement for more complex processing tasks such as reasoning and discourse analysis.

In this paper we describe the current status of the SmartWeb ontology-based information extraction (SOBIE) system.

2 System Overview

The SOBIE 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/events with ontology classes/properties.

Finally, an automatic hyperlinking component is used for the visualization of extracted entities/events. This component is based on the VieWs system, which was developed independently of SmartWeb (Buitelaar et al., 2005).

2.1 Web Crawler

The crawler enables the automatic creation of a football corpus, which is kept up-to-date on a

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

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

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

daily basis. The crawler data is compiled from texts, semi-structured data and copies of original 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 is 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 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:

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

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

 sobie#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 he is now 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 centralized as well as decentralized. Data communication is handled by XML-based exchange formats. Because of 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 visualizing the extracted information through automatic hyperlinking. We argue that such an approach is innovative and enhances the user experience.

Future work includes expansion into 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 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).

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

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, 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, Manchester.

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.

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

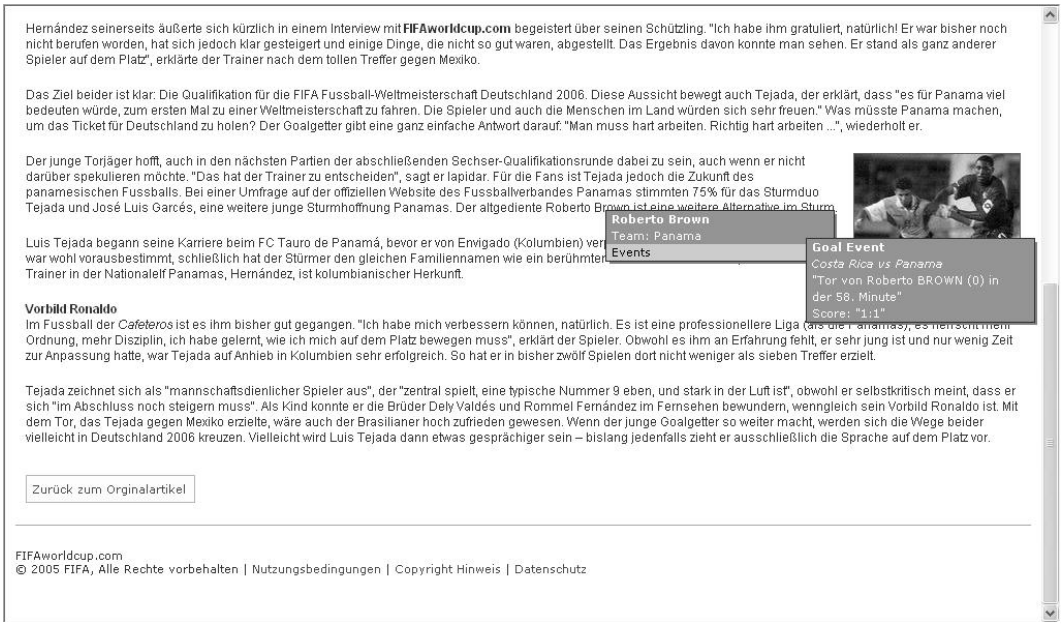

Figure 2: Generated hyperlink on „Roberto Brown“ with extracted information on his team and events in which he participated