SOBA: SmartWeb Ontology-based Annotation

Paul Buitelaar*, Philipp Cimiano+, Anette Frank*, Stefania Racioppa*

* DFKI GmbH, Saarbrücken, Germany {paulb, frank, sracioppa}@dfki.de
+ AIFB, University of Karlsuhe, Germany cimiano@aifb.uni-karlsruhe.de

Abstract: We describe SOBA, a sub-component of the SmartWeb multi-modal dialog system. SOBA is a component for ontology-based information extraction from soccer web pages for automatic population of a knowledge base that can be used for domain-specific question answering. SOBA realizes a tight connection between the ontology, knowledge base and the information extraction component. The originality of SOBA is in the fact that it extracts information from heterogeneous sources such as tabular structures, text and image captions in a semantically integrated way. In particular, it stores extracted information in a knowledge base, and in turn uses the knowledge base to interpret and link newly extracted information with respect to already existing entities.

Introduction

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

In this paper we describe the current status of the SmartWeb Ontology-Based Annotation (SOBA) component, which automatically populates a knowledge base by information extraction from soccer match reports as found on the web. The SOBA system consists of a web crawler, linguistic annotation components and a component for the transformation of linguistic annotations into a knowledge base, i.e. an ontology-based representation.

Web Crawler

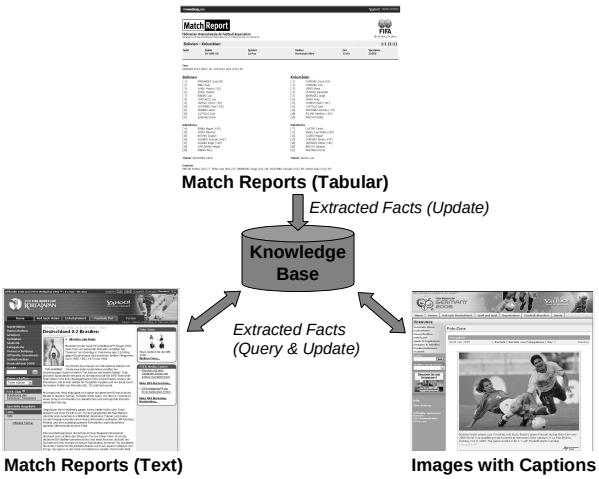
The crawler extracts data from the web: semi-structured data, match reports and images covering the World Cup 2002 and 2006 are identified and collected from the FIFA website. The extracted data are labeled by IDs that match the filename. IDs are derived from the corresponding URL and are thus unique. 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.

Linguistic Annotation

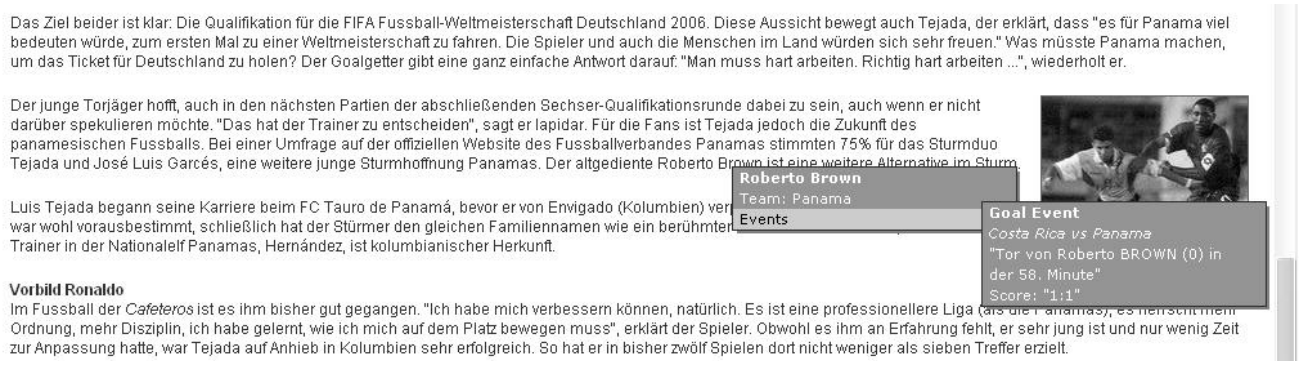
Linguistic annotation in SOBA is based on components that are available in the "Heart of Gold" (HoG) architecture for integrated shallow and deep linguistic processing developed at DFKI [Schäfer 2006], in particular the information extraction system SProUT [Drozdzynski et al. 2004]. SProUT combines finite-state 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 as defined by the soccer ontology developed in SmartWeb [Oberle et al. 2006]. SProUT has basic grammars for the annotation of persons, locations, numerals and date and time expressions. On top of this, we implemented rules for extraction of soccer-specific entities, such as actors in soccer (trainer, player, referee …), teams and tournaments. Using these, we further implemented rules for the extraction of soccer-specific events, such as player activities (shots, headers …), match events (goal, card …) and match results. A soccer-specific gazetteer contains soccer-specific entities and names and is supplemented to the general named-entity gazetteer.

Knowledge Base Generation

At the core of SOBA is the ontology-based transformation component, which semantically integrates the information extracted from tabular and textual match reports, and from associated images, or rather from the image captions. SProUT annotations are mapped to soccer-specific semantic structures as defined by the ontology. The mapping is represented in a declarative fashion specifying how the feature-based structures produced by SProUT are mapped into semantic structures which are compatible with the underlying ontology.



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



Further, the newly extracted information is interpreted in the context of already available information about the match in question, which has been obtained by mapping the extracted semi-structured data on soccer matches to the underlying ontology. The information obtained in this way about the match in question can then be used as background knowledge with respect to which newly extracted information can be correctly interpreted and integrated.

Extraction from Tabular Match Reports

Tabular match reports (semi-structured data) are processed using wrapper-like techniques to transform HTML tables into XML files which are then translated into F-Logic and RDF structures (i.e. class instances) with which the knowledge base (KB) is updated. The KB structures generated for the tabular report include knowledge about the date and time of the match, the stadium it took place in, the number of attendees, the referee, the teams and their players, but also goals, and yellow and red cards in the match.

Extraction from Text Match Reports

In addition to processing tabular reports about each match, SOBA also processes text linked to the match in order to extract additional information, specifically additional events that are represented in the semi-structured data. The semantic transformation component maps extracted events to the ontology and links these class instances to the KB structures created from the tabular reports. The linking is achieved by querying the KB for players mentioned in the text, thus linking the newly extracted information to the ID of the player which is already in the knowledge base. All events that can be extracted from the text are linked to a match instance that was created in processing the tabular match reports. The mapping from SProUT feature structures to KB structures in F-Logic/RDF is specified in a declarative form (XML) and is thus extendable in a flexible manner

Extraction from Image Captions

SOBA also processes image captions for images on the FIFA web pages. Here we use entities and events that can be extracted from the image captions to annotate the corresponding image in the KB to allow for its retrieval given an appropriate question about an entity (i.e. a player) and/or event displayed in the image. To process the captions, SOBA uses the same techniques as for processing free text, but additionally creates a KB structure (i.e. class instance) for the image that corresponds to the extracted information

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[2] system that was developed independently at DFKI. 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 - see the figure above.

Conclusions

We described SOBA, a system for ontology-based extraction, integration and display of information. SOBA as presented here is a domain-specific application. Porting SOBA to another domain can be based on the general purpose NLP components in HoG, but also involves the integration of a domain-specific ontology, extensions and/or modifications of the SProUT gazetteers and rule set and of the KB-related F-Logic rules.

Acknowledgements

This research has been supported by grant 01 IMD01 of the German Ministry of Education and Research (BMB+F) for the SmartWeb project. We would like to thank Thomas Eigner, Greg Gulrajani, Günter Ladwig, Matthias Mantel, Alexander Schutz, Nicolas Weber and Honggang Zhu for implementing parts of the system.

References

U. Callmeier, A. Eisele, U. Schäfer and M. Siegel *The DeepThought Core Architecture Framework*, In Proc. of LREC 2004.

W. Drozdzynski, H-U. Krieger, J. Piskorski, U. Schäfer, F. Xu *Shallow processing with unification and typed feature structures – foundations and applications*, Künstliche Intelligenz, 1:17-23, 2004.

D. Oberle et al. *DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO (SmartWeb Integrated Ontology).* Technical Report[3], 2006.

U. Schäfer *Middleware for Creating and Combining Multi-dimensional NLP Markup.* In Proc. of the Workshop on Multi-dimensional Markup in NLP. Trento, Italy, 2006.

2 http://views.dfki.de

3 http://www.aifb.uni-karlsruhe.de/Publikationen/showPublikation?publ_id=1246