Ontology-based Information Extraction with SOBA

Paul Buitelaar*, Philipp Cimiano[+], Stefania Racioppa*, Melanie Siegel

[*] DFKI GmbH
Stuhlsatzenhausweg 3,
66123 Saarbrücken, Germany
{paulb, sracioppa}@dfki.de

[+] AIFB University of Karlsuhe
Englerstraße 11
Kalrsruhe, Germany
cimiano@aifb.uni-karlsruhe.de

Abstract

In this paper 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.

1. Introduction

SmartWeb 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 extracted information is defined with respect to an underlying ontology (SWIntO: SmartWeb Integrated Ontology [Oberle et al. in preparation]) to enable a smooth integration of derived facts into the general SmartWeb system.

Ontologically described information is a basic requirement for more complex processing tasks such as reasoning and discourse analysis. More in particular, there are three main reasons for formalizing extracted information with respect to an ontology - for related work see e.g. [Reyle and Saric 2001], [Maedche et al 2002], [Alani et al. 2003], [Lopez and Motta 2004], [Müller et al 2004], [Nirenburg and Raskin 2004]:

- *Architecture:* The SmartWeb system is based on the representation of information with respect to an ontology. Results from different components are represented in a uniform way according to the SWIntO ontology, such that it makes no difference for the central SmartWeb dialog system where the information has actually come from, i.e. from open-domain question answering, the knowledge base or from a semantic web service. Complying with the ontology therefore allows for a smooth integration of the information from different processing chains.

- *Information Integration:* Representing information with respect to an ontology and storing it in a knowledge base allows for linking different types of information in a well-founded way, establishing connections between extracted entities and events at the semantic level.

- *Reasoning:* Using a formal ontology allows for applying standard inference engines for reasoning over extracted facts (i.e. entities, events), thus enabling the derivation of further information that is not explicitly contained in the text - in SmartWeb the OntoBroker system is used for inference and reasoning [Decker et al. 1999].

SOBA is original and unique in at least two ways. On the one hand, it implements a novel paradigm in which information extraction, knowledge base updates and reasoning are tightly interleaved. On the other hand, it integrates information from heterogeneous sources (semi-structured data such as tables, unstructured text, images and image captions) on a semantic level in the knowledge base. We are not aware of any system which does this in a similarly principled manner.

2. System Overview

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.

The web crawler acts as a monitor on relevant web domains (i.e. the FIFA[1] and UEFA[2] web sites), automatically downloads relevant documents from them and sends these to a linguistic annotation web service.

[1] http://fifaworldcup.yahoo.com/
[2] http://www.uefa.com/

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 XML-based natural language processing components.

The linguistically annotated documents are further processed by the semantic 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.

In the following section we describe the different components of the system in detail.

2.1 Web Crawler

The crawler enables the automatic creation of a soccer corpus, which is kept up-to-date on a daily basis. The corpus is compiled out of texts, images and semi-structured data on world cup soccer matches that are derived from the original HTML documents. For each soccer match, the data source contains a sheet of semi-structured data with tables of players, goals, referees, etc. Textual data consists of one or more associated match reports. Images are stored with their corresponding captions.

The crawler is able to extract data from two different sources: FIFA and UEFA. Semi-structured data, match reports and images covering the World Cup 2002 and 2006 are identified and collected from the FIFA website. Additional match reports are extracted from the UEFA 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.

2.2 Linguistic Annotation

Linguistic annotation in SOBA is based on components that are available in the HoG architecture, 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 - compare Figure 1. For the annotation of soccer match reports, we extended the rule set of the SProUT with gazetteers, part-of-speech and morphological information.

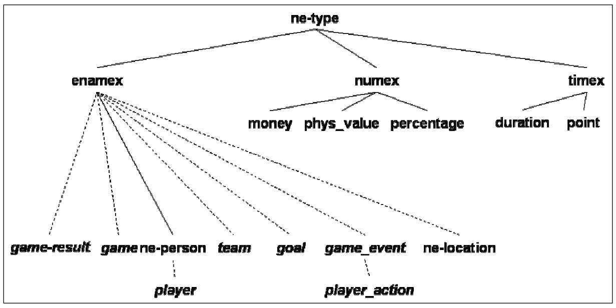

Figure 1: SProUT type hierarchy

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.

2.3 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.

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.

The Knowledge Base (KB) is at the heart of the transformation component, which not only updates facts into the KB, but also queries it to link newly extracted information from texts and image captions to already existing entities such as matches, players, etc. as illustrated in Figure 2. In the following section we discuss ontology-based information extraction from tabular reports, text and image captions in more detail, focusing on how the information from the different resources is integrated.

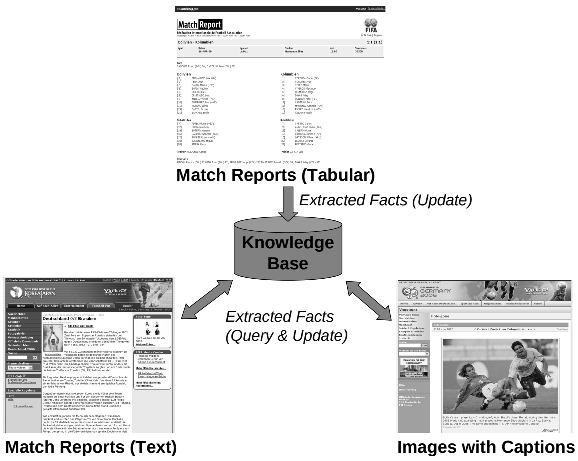

Figure 2: Semantic information integration

```
semistruct#Uruguay_vs_Bolivien_29_Maerz_2000_19:30:sportevent#LeagueFootballMatch
[
                externalRepresentation@(de) ->> "Uruguay vs. Bolivien (29. Maerz 2000 19:30)";
                dolce#"HAPPENS-AT" -> semistruct#"29. Maerz 2000 19:30_interval";
                sportevent#heldIn -> semistruct#"Montevideo_Centenario_29_Maerz_2000_19_30_Stadium";
                sportevent#team1Result -> 1;
                sportevent#team2Result -> 0;
                sportevent#attendance ->49811;
                sportevent#team1 -> semistruct#"Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Uruguay_MatchTeam";
                sportevent#team2 -> semistruct#"Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Bolivien_MatchTeam";
     (…)
]
semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Bolivien_MatchTeam:sportevent#FootballMatchTea
[
      externalRepresentation@(de) ->> "Bolivien";
                sportevent#name -> "Bolivien";
                sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Jose_FERNANDEZ_PFP";
                sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Juan_PENA_PFP";
                sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Marco_SANDY_PFP";
                sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Vladimir_SORIA_PFP";
                sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Luis_RIBEIRO_PFP";
                sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Luis_CRISTALDO_PFP";
      (...)
]
semistruct#"Uruguay_vs_Bolovien_29_Maerz_2000_19 :30_Luis_CRISTALDO_PFP":sportevent#FieldMatchFootballPlayer
[
                externalRepresentation@(de) ->> "Luis CRISTALDO (8)";
                sportevent#number -> 8;
                sportevent#impersonatedBy -> semistruct#"Luis_CRISTALDO"
].
semistruct#"Luis_CRISTALDO":dolce#"natural-person"
[
                externalRepresentation@(de) ->> "Luis CRISTALDO";
                dolce#"HAS-DENOMINATION" -> semistruct#"Luis_CRISTALDO_NaturalPersonDenomination"
].

semistruct#"Luis_CRISTALDO_NaturalPersonDenomination":dolce#"natural-person-denomination"
[
                externalRepresentation@(de) ->> "Luis CRISTALDO";
                dolce#LASTNAME -> "CRISTALDO";
                dolce#FIRSTNAME -> "Luis"
].
```

Figure 3: KB structures (F-Logic) derived from a tabular match report on Uruguay-Bolivia March 29th 2000

3. Ontology-based Information Extraction

3.1 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 [Kifer et al. 1995] and RDF[3] structures (i.e. class instances) with which the knowledge base 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, yellow and red cards in the match. Figure 3 gives an example for the KB structures automatically generated for the match between Uruguay and Bolivia on the 29th of March 2000.

3.2 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.

For instance from a text match report on the same match between Uruguay and Bolivia on the 29th of March 2000, we could extract the event that the player *Luis Cristaldo* has been banned. We can then generate an instance for this event and link this to already available information on this match by pointing to the correct ID for *Luis Cristaldo* as shown in Figure 4.

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

```
semistruct#Uruguay_vs_Bolivien_29_Maerz_2000_19:30
[
  sportevent#matchEvents -> soba#ID11
].

soba#ID11:sportevent#Ban
[
 sportevent#commitedOn ->
semistruct#Uruguay_vs_Bolivivien_(…)_Luis_CRISTALDO_PFP
].
```

Figure 4: KB structures (F-Logic) derived from a text match report on the Uruguay-Bolivia match

[3] Resource Description Framework: http://www.w3.org/RDF/

3.3 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 the event described in the image. To process the captions, SOBA uses the same techniques as when processing free text, but additionally creates a KB entity for the image pointing to the extracted information. Let’s assume that SOBA has extracted a foul-event committed by *Luis Cristaldo*, it will create KB structures as depicted in Figure 5.

```
semistruct#Uruguay_vs_Bolivien_29_Maerz_2000_19:30
[
  sportevent#matchEvents -> soba#ID25
].

soba#ID25:sportevent#Foul
[
  sportevent#commitedBy ->
semistruct#Uruguay_vs_Bolivien_(…)_Luis_CRISTALDO_PFP
].

mediainst#ID67:media#Picture
[
  media#URL ->
"http://fifaworldcup.yahoo.com/06/de/photos/124155.jpg";
  media#shows -> ID25
].
```

Figure 5: KB structures (F-Logic) derived from an image caption on the Uruguay-Bolivia match

4 Conclusions and Future Work

We described SOBA, an information-extraction system which relies on an ontology to formalize and semantically integrate (link) extracted information from heterogeneous resources in a knowledge base. We are not aware of other systems that process tables, text and image captions and semantically integrate extracted information according to an ontology in a similar principled way.

In future work, reasoning with respect to the ontology and knowledge-based discourse analysis techniques such as described in [Cimiano et al. 2005] will provide additional functionality to the system. In this respect we intend also to provide deeper linguistic processing, e.g. with HPSG as available within the Heart-of-Gold architecture.

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 our students 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 Proceedings of LREC 04, pp.1205-1208, 2004.

P. Cimiano, J. Saric, and U. Reyle, *Ontology-driven discourse analysis for information extraction*, Data Knowledge Engineering 55(1), 2005.

M. Decker, M. Erdmann, D. Fensel, R. Studer, *Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information,* Database Semantics: Semantic Issues in Multimedia, pp. 351-369, 1999.

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

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

H. Alani, S. Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis, N.R. Shadbolt, *Automatic Ontology-Based Knowledge Extraction from Web Documents*. IEEE Intelligent Systems, 18(1), pp. 14-21, 2003.

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

A. Maedche, G. Neumann and S. Staab, *Bootstrapping an Ontology-Based Information Extraction System*. In: Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web, Springer, 2002.

H.M. Müller, E.E. Kenny, P.W. Sternberg, *Textpresso: An ontology-based information retrieval and extraction system for biological literature*, PLoS Biol 2, 2004.

S. Nirenburg and V. Raskin, *Ontological Semantics,* MIT Press, 2004.

D. Oberle, A. Ankolekar, P. Hitzler, P. Cimiano, C. Schmidt, M. Weiten, B. Loos, R. Porzel,H.-P. Zorn,M. Micelli, M. Sintek,M. Kiesel, B. Mougouie, S. Vembu, S. Baumann, M. Romanelli, P. Buitelaar, R. Engel, D. Sonntag, N. Reithinger, F. Burkhardt, J. Zhou *DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO (SmartWeb Integrated Ontology),* in preparation.

U. Reyle and J. Saric, *Ontology Driven Information Extraction*, Proceedings of the 19th Twente Workshop on Language Technology, 2001.