Ontology Learning and Population in SmartWeb

Paul Buitelaar*, Nicolas Weber*[1], Philipp Cimiano+

*DFKI GmbH - Language Technology Lab & Competence Center Semantic Web
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
paulb@dfki.de

+ University of Karlsruhe, Institut für Angewandte Informatik und Formale Beschreibungsverfahren
D-76128 Karlsruhe, Germany
cimiano@aifb.uni-karlsruhe.de

1 Introduction

The work described in this paper is concerned with the population and extension of a soccer ontology in the context of the SmartWeb[2] project. The SmartWeb system 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. The system is thus able to provide intelligent information services that are accessible via mobile broadband devices. The main scenario for SmartWeb is the FIFA World Cup 2006. Here we describe SOBA and ISOLDE, two sub-components of the SmartWeb system.

SOBA (SmartWeb Ontology-based Semantic Annotation), described in section 2, is a system for ontology-based information extraction from web pages for the automatic population of a knowledge base used in 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.

The ISOLDE (Information System for Ontology Learning and Domain Exploration) system we describe in section 3 generates a domain ontology by extracting class candidates from the linguistic context of a given set of ontology instances and by deriving further knowledge on these class candidates from available web resources. The last few years saw a continuing increase in the availability of open source web-based information resources such as Wikipedia and similar initiatives. In this paper we show how Wikipedia, Wiktionary and a German online dictionary (DWDS) can be used in combination with a domain corpus and a general purpose named-entity tagger to derive a domain ontology.

2 Ontology Population with SOBA

SOBA 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. The extracted information is defined with respect to

[1] Nicolas Weber is currently at the Know-Center in Graz, Austria: http://www.know-center.at
[2] http://www.smartweb-projekt.de/

an underlying ontology (SWIntO: SmartWeb Integrated Ontology [Oberle et al. 2006]) 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. [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.

2.1 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[3] and UEFA[4] web sites), automatically downloads relevant documents from them and sends these 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 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 sections we describe the different components of the system in detail.

2.2 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

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

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

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. The extracted data are labeled by IDs that match the filename. They 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.

2.3 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 SProUT with gazetteers, part-of-speech and morphological information.

SProUT has basic grammars for the annotation of persons, locations, numerals and date and time expressions. On top of this, we implemented rules for the 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 names and is supplemented to the general named-entity gazetteer.

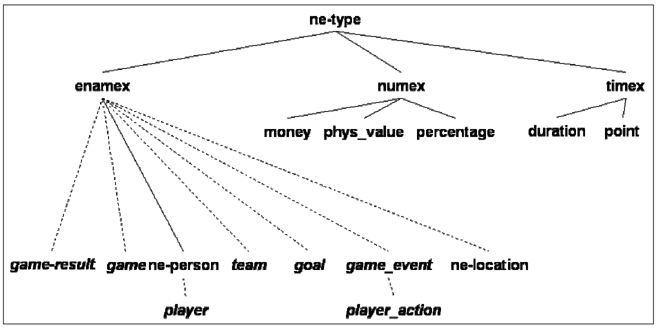


Figure 1: SProUT type hierarchy

2.4 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 sections 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.

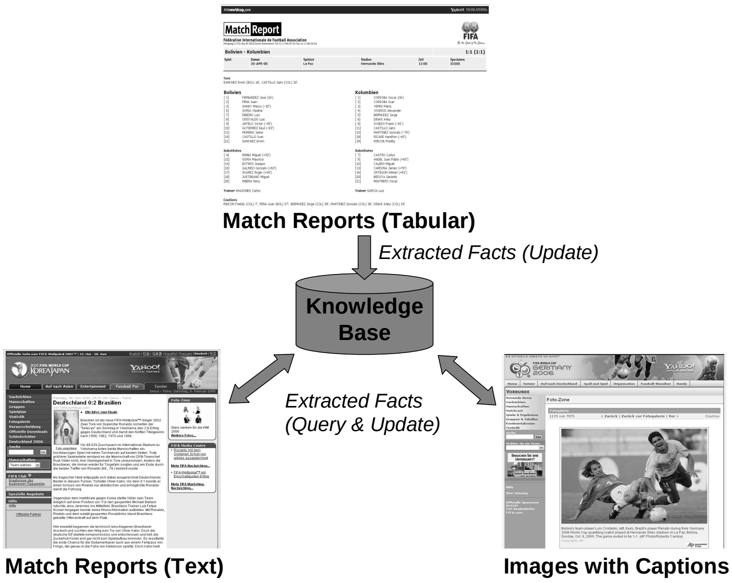


Figure 2: Semantic information integration

2.5 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[5] structures (i.e. class instances) with which the knowledge base is updated. The instances 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.

2.6 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 instances created from the tabular reports. The linking is achieved by querying the KB for players mentioned in the text, thus

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

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 depicted in Figure 3.

```
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 3: Instances derived from a text match report on the Uruguay-Bolivia match

2.7 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.

3 Ontology Learning with ISOLDE

The SOBA system populates the SWIntO ontology as defined in the SmartWeb project. Obviously, this definition is however only a snapshot of the soccer domain. As with all topical domains, the soccer domain will change over time in terms of relevant entities and events that it covers. Tracking and modeling this change can be handled by ontology learning systems such as ISOLDE[6].

The ISOLDE (Information System for Ontology Learning and Domain Exploration) system generates a domain ontology by extracting class candidates from the linguistic context of a given set of ontology instances and by deriving further knowledge on these class candidates from available web resources. The ISOLDE approach is based on techniques for unsupervised named-entity recognition as developed among others by [Yangarber et al. 2000, Cimiano and Staab 2004, Cimiano et al. 2005, Etzioni et al. 2005] but the results are used in a different way. Whereas for these approaches the assignment of named-entities with extracted classes is the final goal, ISOLDE goes one step further by trying to find additional information on the extracted classes in order to organize them into a taxonomy or full ontology.

[6] For a recent overview of recent research in ontology learning systems, see [Buitelaar et al., 2005]

3.1 ISOLDE System Overview

The ISOLDE system can be defined by three analysis steps that can be defined as follows (see also Figure 4):

1. Named-entity recognition (NER) for the extraction of instances for classes in the base ontology
2. Linguistic pattern analysis for the extraction of class candidates from the context of the instances extracted in step 1
3. Collecting web-based knowledge for the class candidates extracted in step 2 and integrating this into a new or extended taxonomy/ontology

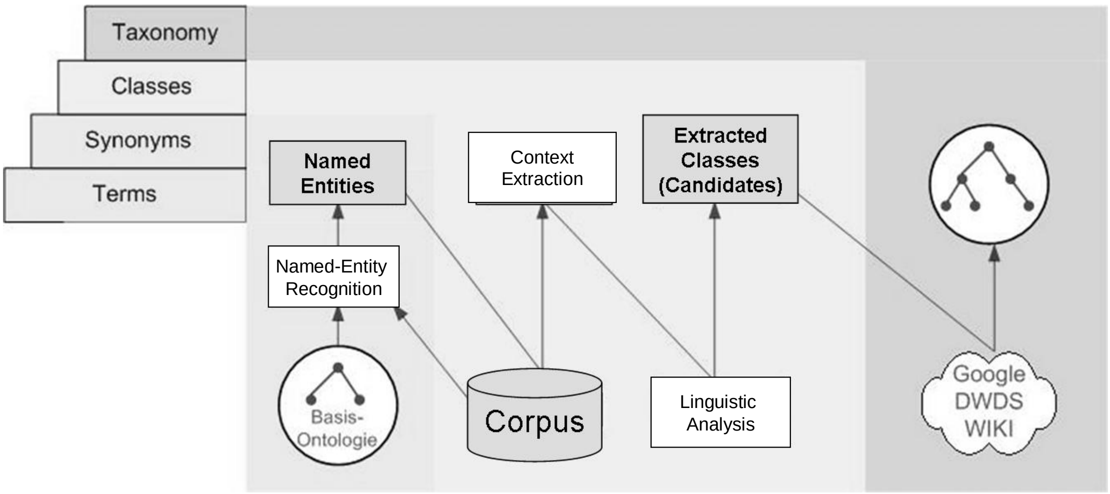


Figure 4: ISOLDE system overview

In step 1 we use a domain-specific corpus, a base ontology and a general purpose NER system (SproUT) to find instances for the classes in the base ontology.

In step 2 we collect the linguistic contexts of the instances derived in step 1 and extract class candidates from this by use of lexico-syntactic patterns [Hearst 1992]:

NE „*ist ein*“ {NP}	*Jürgen Klinsmann ist ein Trainer*	*(…is a trainer)*
NE „ , “ {NP}	*Jürgen Klinsmann, Trainer des...*	*(trainer of …)*
{NP} NE	*Trainer Jürgen Klinsmann*	*(trainer …)*

If one of these patterns matches, the head of the NP is taken as a new class candidate. For instance, we can extract the class candidate TRAINER from the context of *Jürgen Klinsmann* – which was instantiated for the class PERSON in the base ontology – in the following sentence:

Jürgen Klinsmann, Trainer der Nationalmannschaft

(Jürgen Klinsmann, trainer of the national team)

The result of step 2 is a list with extracted class candidates for each named-entity. For every class candidate we determine its statistical relevance by use of X^2, which provides a measure over frequency in the domain specific corpus relative to that in a balanced corpus [Manning and Schütze 1999].

In step 3 we collect information on and between extracted class candidates from online resources: DWDS, Wikipedia and Wiktionary. *Das Digitale Wörterbuch der Deutschen Sprache – DWDS* (the digital dictionary of German) is a public online dictionary. It consists of a corpus that currently comprises 80.000 texts and a corresponding dictionary. Wikipedia is a free multilingual encyclopaedia that anyone can edit. As of today there are 345.000 articles in German in various domains. The articles consist of free (i.e. unstructured) text and semi-structured data. Wiktionary, a multilingual dictionary and thesaurus, is a sister project from Wikipedia. The German Wiktionary is still a small project (15.000 articles) but it is the fastest growing Wiki-project.

Whereas Wikipedia provides general knowledge on various topics, DWDS and Wiktionary additionally provide linguistic knowledge on lexical semantic relations (synonymy, hyponymy), morphology and part of speech. In the context of ISOLDE, we aim at deriving the following information from these resources:

- taxonomic: establishing if two class candidates are in a hierarchical relation (compare the RDF/OWL property *SubClassOf*)
- non-taxonomic: establishing if two class candidates are equivalent (compare the OWL property *SameAs*)

A problem occurs if extracted relations are in conflict to each other, as in *SubClassOf(Defender, Goalkeeper)* and *SubClassOf(Goalkeeper,Defender)*. To avoid this, the relations are ranked. If a relation is more frequent than another, it is more likely that this relation is correct. Are two conflicting relations equally frequent, then the two classes could be equivalent as in *SameAs(Goalkeeper, Defender).*

3.2 Experiment

In order to test the ISOLDE system we defined an experiment in generating a domain ontology for soccer by use of a domain-specific corpus and the web resources discussed above. The corpus consists of around 3000 match reports and other news articles on soccer that were downloaded from the web.

In step 1, all named-entities of class PERSON were collected. As presented in Figure 5, there is a correlation between the number of extracted (different) instances of class PERSON and the number of documents these occur in. The higher the thresholds for the occurrence of the PERSON instances the less different PERSON instances are processed. We therefore decided on taking into account only those PERSON instances that occurred at least 40 times.

In step 2 the linguistic context for each of the extracted named-entities is analyzed, e.g. for the named-entity *Michael Ballack:*

Linguistic Context: *Münchens Mittelfeldspieler Michael Ballack ist nach einer Entscheidung …*

Pattern:	*{NP} Ballack*
Extracted Class Candidate:	*Mittelfeldspieler*

Linguistic Context: *Michael Ballack, bester Mann auf dem Platz..*

Pattern:	*Ballack “,” {NP}*
Extracted Class Candidate:	*Mann*

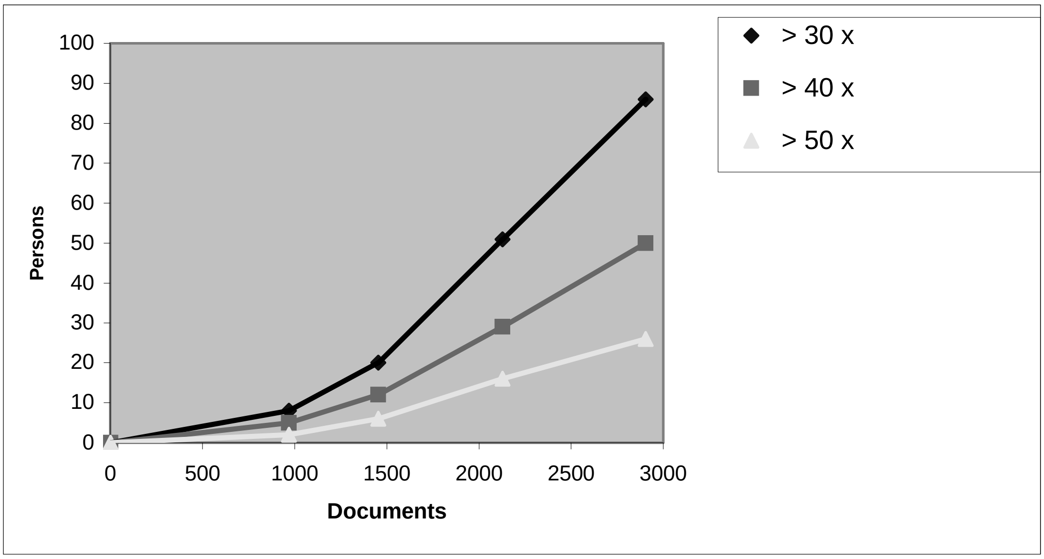


Figure 5: Correlation between the number of different instances of class PERSON and the number of documents they occur in

As discussed above, a frequency and statistical relevance measure is computed for each of the extracted class candidates. For instance, *Stürmer (striker)* occurs 334 times in our corpus (~0.5 mio tokens) and 316 times in the general corpus (~9 mio tokens), which may be represented as follows:

0	**1**	**2**
1	334	316
2	500000	9000000

Using the formula for $\boldsymbol{X}^2$ defined as:

$$\frac{N\,(O_{11}\,O_{22} - O_{12}\,O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

In this way, we are able to determine which class candidates are relevant to the domain. For instance, the class candidates for *Thierry Henry* are *Stürmer (striker), Trainer (trainer), Vater (father)* of which only the first two are over a certain threshold that determines their domain relevance – see the table below. In addition we want to decide which ones of these classes are really of importance relative to the extracted named-entity (NE). For this purpose we keep track of their co-occurrence, which in this case lets us decide to select *Stürmer* and not *Trainer* as a relevant class candidate.

Class Candidate	X^2	Co-occurrence with NE
Stürmer (striker)	*2771.27*	4
Trainer (trainer)	*19.78*	1
Vater (father)	*0.8*	4

In step 3 we extract information from Wikipedia, Wiktionary and DWDS for all of the relevant class candidates. For instance, the following information for *Torwart (goalkeeper)*:

Wikipedia

Der **Torwart** (Torhüter, Tormann, Keeper; Schweiz. Goalie) **ist ein Mitspieler** einer Mannschaftssportart. Er ist der defensivste Spieler seiner Mannschaft und seine Hauptaufgabe besteht darin zu verhindern, dass das Spielgerät (z.B. ein Ball) ins Tor der eigenen Mannschaft gelangt. Daher wird er auch Torhüter genannt….

Wiktionary

Bedeutungen:

Derjenige Fußballspieler, dessen Aufgabe es ist, gegnerische Tore zu vermeiden und der hierfür als einziger Spieler auch seine Hände einsetzen darf.

Herkunft:

aus Tor und Wart

Synonyme:

Tormann, Torhüter, Keeper

Oberbegriffe:

Sport, Fußball, Fußballspieler

Unterbegriffe:

Fliegenfänger (neg.)

DWDS

Torwart, der **1.** Ballspiele *Spieler im Tor, der den Ball fängt, abwehrt*: der T. warf den Ball zum Verteidiger **2.** hist. *Wachmann am Tor*;

Hyperonyme

Spieler

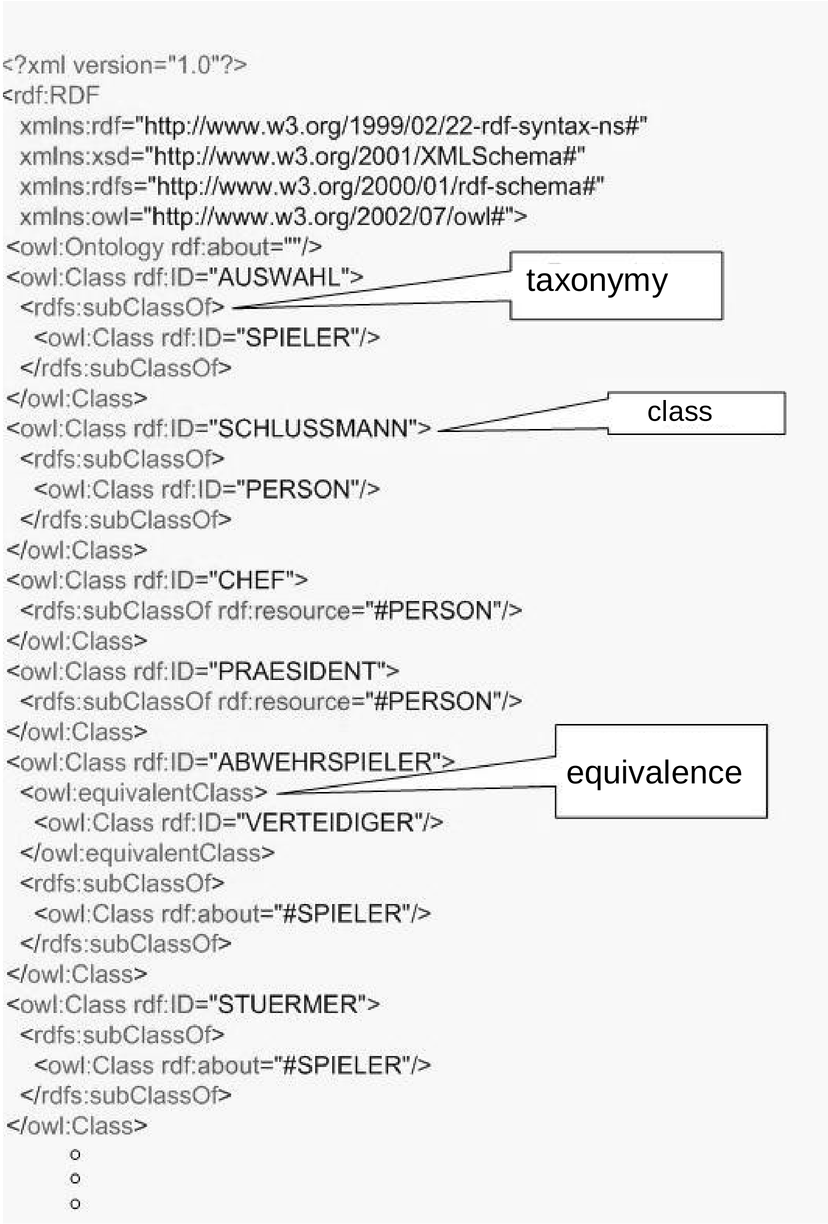
Hyponyme

Abwehr Abwehrspieler

For each of the class candidates, information can be extracted and gathered in this way and merged into an ontology. The target format for the ontology is OWL as we aim to represent knowledge on equivalence with the OWL property SameAs. The soccer ontology we derived in the experiment looks as depicted by Figure 6.

3.3 Evaluation

To evaluate the generated ontology, we compare it to a gold standard or reference ontology. For this purpose we used the manually created SportEvent ontology which was developed in the SmartWeb project as part of the SWIntO ontology mentioned before. For the evaluation however we used only a relevant sub-tree with root *SportiveRole*, which contains 50 classes and 49 direct (taxonomic) relations and 226 indirect relations.



```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#">
<owl:Ontology rdf:about=""/>
<owl:Class rdf:ID="AUSWAHL">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="SPIELER"/>
  </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="SCHLUSSMANN">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="PERSON"/>
  </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="CHEF">
  <rdfs:subClassOf rdf:resource="#PERSON"/>
</owl:Class>
<owl:Class rdf:ID="PRAESIDENT">
  <rdfs:subClassOf rdf:resource="#PERSON"/>
</owl:Class>
<owl:Class rdf:ID="ABWEHRSPIELER">
  <owl:equivalentClass>
    <owl:Class rdf:ID="VERTEIDIGER"/>
  </owl:equivalentClass>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#SPIELER"/>
  </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="STUERMER">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#SPIELER"/>
  </rdfs:subClassOf>
</owl:Class>
      o
      o
      o
```

Figure 6: Soccer ontology derived from the soccer corpus and additional web resources

A manually generated ontology has always a deeper (hierarchy-) structure than an automatically generated ontology. One reason for this is the fact that for structuring purposes additional classes are defined. Classes like *PositionalMatchFootballPlayer* or *SituationFootballPlayer* are not used in common language but serve only for structuring of the ontology. For this reason the automatically generated ontology is evaluated against the complete SportEvent ontology as well as against a reduced version of this ontology (SportEvent - adjusted) without the classes used for structuring only.

The experiment described in the previous section presented us with 45 classes and 37 relations between these classes which were extracted from the domain corpus and the discussed web resources as follows:

Relations	**all**	**taxonomy**	**equivalence**
	37	*31*	*6*
Wikipedia	2	2	0
Wiktionary	12	8	4
DWDS	19	17	2
GOOGLE	4	4	0

As presented in this table, 85% of relations are obtained from DWDS and Wiktionary and only 15% from Wikipedia and Google. The reason for this is the different type of structure of these documents. DWDS and Wiktionary are semi-structured, i.e. the extraction of the pre-processed relations occurs by the position in text. Wikipedia and Google in contrast only provide unstructured text.

Precision and recall results are shown in the table below. As may be expected, results on the acquisition of ontology classes are much better than on relations, as there are many more possible combinations on the relational level than on the class level, i.e. the system will be able to ‘get it wrong’ more often.

	total	**true positives**	**RECALL**	**PRECISION**
Classes				
SportEvent	50	23	46,0%	31,9%
SportEvent - adjusted	43	23	53.4%	35,3%
Relations				
SportEvent	226	24	10,6%	10,4%
SportEvent - adjusted	107	24	22,4%	21,6%

4 Conclusions

We described SOBA, a system for ontology-based extraction, integration and display of information and ISOLDE, a system for web based ontology learning.

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.

ISOLDE uses web resources such as Wikipedia and Wiktionary in combination with a domain corpus, a general purpose named-entity tagger and a seed or ‘base’ ontology to derive a domain ontology. The experiment shows that the best results may be obtained from semi-structured data resources (e.g. web dictionaries).

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, Anette Frank, Greg Gulrajani, Günter Ladwig, Matthias Mantel, Stefania Racioppa, Alexander Schutz and Honggang Zhu for their contributions to the SOBA system.

References

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.

P. Buitelaar, P. Cimiano, B. Magnini (eds.) *Ontology Learning from Text: Methods, Evaluation and Applications* Frontiers in Artificial Intelligence and Applications Series, Vol. 123, IOS Press, July 2005.

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

P. Cimiano, S. Staab. 2004. *Learning by Googling*. In: SIGKDD Explorations Vol. 6, No. 2.

P. Cimiano, G. Ladwig, S. Staab. 2005. *Gimme´ The Context: Context-driven Automatic Semantic Annotation with C-PANKOW*. In A. Ellis, T. Hagino (eds.) Proceedings of the 14th World Wide Web Conference, Japan. ACM Press.

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, F. Xu. 2004. *Shallow processing with unification and typed feature structures – foundations and applications*. Künstliche Intelligenz, 1:17-23.

O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. 2005. *Unsupervised named-entity extraction from the web: An experimental study*. Artificial Intelligence, 165(1):91–134.

M. Hearst. 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora*. In: Proceedings of the 14th International Conference on Computational linguistics, Nantes.

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.

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.

Ch. D. Manning, H. Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press. Cambridge, MA.

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).* Technical Report, 2006.

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

R. Yangarber, R. Grishman, P. Tapanainen and S. Huttunen. 2000. *Automatic Acquisition of Domain Knowledge for Information Extraction* In Proceedings of COLING 2000, 18th International Conference on Computational Linguistics, Saarbrücken.