SmartWeb Handheld — Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services

Daniel Sonntag, Ralf Engel, Gerd Herzog, Alexander Pfalzgraf,
Norbert Pfleger, Massimo Romanelli, Norbert Reithinger
German Research Center for Artificial Intelligence
66123 Saarbrücken, Germany
firstname.lastname@dfki.de

Abstract

SMARTWEB aims to provide intuitive multimodal access to a rich selection of Web-based information services. We report on the current prototype with a smartphone client interface to the Semantic Web. An advanced ontology-based representation of facts and media structures serves as central description for rich media content. Underlying content is accessed through conventional web service middleware to connect the ontological knowledge base and an intelligent web service composition module for external web services, which is able to translate between ordinary XML-based data structures and explicit semantic representations for user queries and system responses. The presentation module renders the media content and the results generated from the services and provides a detailed description of the content and its layout to the fusion module. The user is then able to employ multiple modalities, like speech and gestures, to interact with the presented multimedia material in a multimodal way.

1 Introduction

The development of a context-aware, multimodal mobile interface to the Semantic Web [Fensel *et al.*, 2003], i.e., ontologies and web services, is a very interesting task since it combines many state-of-the-art technologies such as ontology development, distributed dialog systems, standardized interface descriptions (EMMA[1], SSML[2], RDF[3], OWL-S[4], WSDL[5], SOAP[6], MPEG7[7]), and composition of web services. In this contribution we describe the intermediate steps in the dialog system development process for the project SMARTWEB [Wahlster, 2004], which was started in 2004 by partners from industry and academia.

[1] http://www.w3.org/TR/emma
[2] http://www.w3.org/TR/speech-synthesis
[3] http://www.w3.org/TR/rdf-primer
[4] http://www.w3.org/Submission/OWL-S
[5] http://www.w3.org/TR/wsdl
[6] http://www.w3.org/TR/soap
[7] http://www.chiariglione.org/mpeg

In our main scenario, the user carries a smartphone PDA and poses closed and open domain multimodal questions in the context of football games and a visit to a Football Worldcup stadium. Many challenging task such as interaction design for mobile devices with restricted computing power have to be addressed: the user should be able to use the PDA as a question answering (QA) system, using speech and gestures to ask for information about players or games stored in ontologies, or other up-to-date information like weather forecast information accessible through web services, Semantic Web pages (Web pages wrapped by semantic agents), or the Internet.

The partners of the SMARTWEB project share experience from earlier dialog system projects [Wahlster, 2000; 2003; Reithinger *et al.*, 2005b]. We followed guidelines for multimodal interaction, as explained in [Oviatt, 1999] for example, in the development process of our first demonstrator system [Reithinger *et al.*, 2005a] which contains the following assets: *multimodality*, more modalities allow for more natural communication, *encapsulation*, we encapsulate the multimodal dialog interface proper from the application, *standards*, adopting to standards opens the door to scalability, since we can re-use ours as well as other's resources, and *representation*. A shared representation and a common ontological knowledge base ease the data flow among components and avoids costly transformation processes. In addition, semantic structures are our basis for representing dialog phenomena such as multimodal references and user queries. The same ontological query structures are input to the knowledge retrieval and web service composition process.

In the following we demonstrate the strength of Semantic Web technology for information gathering dialog systems, especially the integation of multiple dialog components, and show how knowledge retrieval from ontologies and web services can be combined with advanced dialogical interaction, i.e., system-initiative callbacks, which present a strong advancement to traditional QA systems. Traditional QA realizes like a traditional NLP dialog system a (recognize) - analyze - react - generate - (synthesize) pipeline [Allen *et al.*, 2000]. Once a query is being started, the information is pipelined until the end, which means that the user-system interaction is reduced to user and result messages. The types of dialogical phenomena we address and support include reference resolution, system-initiated clarification requests and

pointing gesture interpretation among others. Support for underspecified questions and enumeration question types additionally shows advanced QA functionality in a multimodal setting. One of the main contributions is the ontology-based integration of verbal and non-verbal system input (fusion) and output (system reaction).

The paper is organized as follows: we begin with an example interaction sequence, in section 3, we explain the dialog system architecture. In section 4, the ontological knowledge representation and web service access is described. Section 5 then gives a description of the underlying language parsing and discourse processing steps, and their integration. Conclusions about the success of the system so far and future plans are outlined in section 6.

2 Multimodal interaction sequence example

The following interaction sequence is typical for the SMARTWEB dialog system.

---

(1) **U:** "When was Germany world champion?"

(2) **S:** "In the following 4 years: 1954 (in Switzerland), 1974 (in Germany), 1990 (in Italy), 2003 (in USA)"

(3) **U:** "And Brazil?"

(4) **S:** "In the following 5 years: 1958 (in Sweden), 1962 (in Chile), 1970 (in Mexico), 1994 (in USA), 2002 (in Japan)" + [*team picture, MPEG-7 annotated*]

(5) **U:** Pointing gesture on player *Aldair* + "How many goals did this player score?"

(6) **S:** "Aldair scored none in the championship 2002."

(7) **U:** "What can I do in my spare time on Saturday?"

(8) **S:** "Where?"

(9) **U:** "In Berlin."

(10) **S:** *The cinema program, festivals, and concerts in Berlin are listed.*

The first and second enumeration questions are answered by deductive reasoning within the ontological knowledge base modeled in OWL [Krotzsch *et al.*, 2006] representing the static but very rich implicit knowledge that can be retrieved. The second example beginning with (7) evokes a dynamically composed web service lookup. It is important to note that the query representation is the same for all the access methods to the Semantic Web (cf. section 5.1) and is defined by foundational and domain-specific ontologies. In case that the GPS co-ordinates were accessible from the mobile device, the clarification question would have been omitted.

3 Architecture approach

A flexible dialog system platform is required in order to allow for true multi-session operation with multiple concurrent users of the server-side system as well as to support

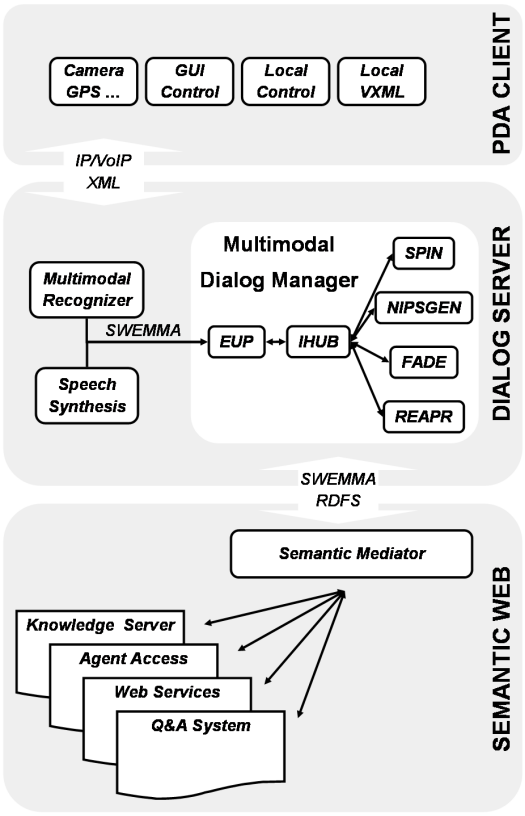


Figure 1: SMARTWEB handheld architecture.

audio transfer and other data connections between the mobile device and a remote dialog server. This types of systems have been developed, like the Galaxy Communicator [Cheyer and Martin, 2001] (cf. also [Seneff *et al.*, 1999; Thorisson *et al.*, 2004; Herzog *et al.*, 2004; Bontcheva *et al.*, 2004]), and commercial platforms from major vendors like VoiceGenie, Kirusa, IBM, and Microsoft use X+V1, HTML+SALT2, or derivatives for speech-based interaction on mobile devices. For our purposes these platforms are too limited. To implement new interaction metaphors and to use Semantic Web based data structures for both dialog system internal and external communication, we developed a platform designed for Semantic Web data structures for NLP components and backend knowledge server communication. The basic architecture is shown in figure 1.

It consists of three basic processing blocks: the PDA client, the dialog server which comprises the dialog manager, and the Semantic Web access system.

On the PDA client, a local Java-based control unit takes care of all I/O, and is connected to the GUI-controller. The local VoiceXML-based dialog system resists on the PDA for interaction during link downtimes.

The dialog server system platform instantiates one dialog server for each call and connects the multimodal recognizer

for speech and gesture recognition. The dialog system instantiates and sends the requests to the *Semantic Mediator*, which provides the umbrella for all different access methods to the Semantic Web we use. It consists of an open domain QA system, a Semantic Web service composer, Semantic Web pages (wrapped by semantic agents), and a knowledge server.

The dialog system consist of different, self-contained processing components. To integrate them we developed a Java-based hub-and-spoke architecture [Reithinger and Sonntag, 2005]. The most important processing modules in the dialog system connected in the IHUB are: a speech interpretation component (SPIN), a modality fusion and discourse component (FADE), a system reaction and presentation component (REAPR), and a natural language generation module (NIPSGEN), all discussed in section 5. An EMMA Unpacker/Packer (EUP) component provides the communication with the dialogue server and Semantic Web subsystem external to the multimodal dialog manager and communicates with the other modules of the dialog server, the multimodal recognizer, and the speech synthesis system.

Processing a user turn, the normal data flows through $SPIN \rightarrow FADE \rightarrow REAPR \rightarrow SemanticMediator \rightarrow REAPR \rightarrow NIPSGEN$. However, the data flow is often more complicated when, for example, misinterpretations and clarifications are involved.

4 Ontology representation and web services

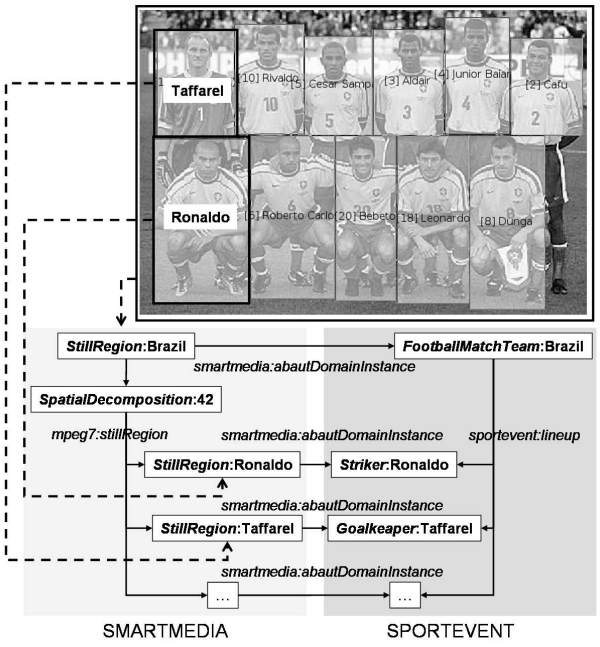


Figure 2: A SMARTMEDIA instance representing the decomposition of the Brazil 1998 world cup football team image.

The ontological infrastructure of SMARTWEB, the SWIntO (SMARTWEB **Int**egrated **O**ntology), is based on an upper model ontology realized by merging well chosen concepts from two established foundational ontologies, DOLCE [Gangemi *et al.*, 2002] and SUMO [Niles and Pease, 2001], in a unique one: the SMARTWEB foundational ontology SMARTSUMO [Cimiano *et al.*, 2004]. Domain specific knowledge (sportevent, navigation) is defined in dedicated ontologies modeled as sub-ontologies of the SMARTSUMO. The SWIntO integrates question answering specific knowledge of a discourse ontology (DISCONTO) and representation of multimodal information of a media ontology (SMARTMEDIA). The data exchange is RDF-based.

We realized a discourse ontology (DISCONTO) with particular attention to the modeling of discourse interactions in QA scenarios. The DISCONTO provides concepts for dialogical interaction with the user as well as more technical request-response concepts for data exchange with the Semantic Web subsystem including answer status which is important in interactive systems. In particular DISCONTO comprises concepts for multimodal dialog management, a dialog act taxonomy, lexical rules for syntactic-semantic mapping, HCI concepts (e.g. pattern language for interaction design [Sonntag, 2005]), and concepts for questions, question focus, semantic answer types [Hovy *et al.*, March 2001], and multimodal results [Sonntag and Romanelli, 2006].

Information exchange between the components of the server-side dialog system is based on the W3C EMMA standard that is used to realize containers for the ontological instances representing, e.g., multimodal input interpretations. SWEMMA is our extension to the EMMA standard which introduces additional *Result* structures in order to represent components output. On the ontological level we modeled an RDF/S-representation of EMMA/SWEMMA.

The SMARTMEDIA is an MPEG7-based media ontology and an extension to [Hunter, 2001; Benitez *et al.*, 2002] that we use to represent output result, offering functionality for multimedia decomposition in space, time and frequency (mpeg7:SegmentDecomposition), file format and coding parameters (mpeg7:MediaFormat), and a link to the Upper Model Ontology (smartmedia:aboutDomainInstance). In order to close the semantic gap between the different levels of media representations, the *smartmedia:aboutDomainInstance* property has been located in the top level class *smartmedia:Segment*. The link to the upper model ontology is inherited to all segments of a media instance decomposition to guarantee deep semantic representations for the *smartmedia* instances referencing the specific media object and for making up segment decompositions.

Figure 2 shows an example of this procedure applied to an image of the Brazilian football team in the final match of the World Cup 1998, as introduced in the interaction example. In the example an instance of the class *mpeg7:StillRegion*, representing the complete image, is decomposed into different *mpeg7:StillRegion* instances representing the segments of the image which show individual players.

The *mpeg7:StillRegion* instance representing the entire picture is then linked to a *sportevent:MatchTeam* instance, and each segment of the picture is linked to a *sportevent:FieldFootballPlayer* instance or sub-instance. These representations offer a framework for gesture and speech fusion when users interact with Semantic Web results such as MPEG7-annotated images, maps with points-of in-

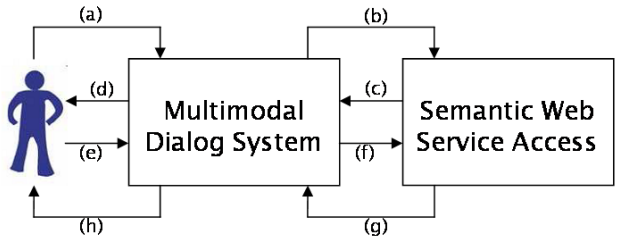
terest, or other interactive graphical media obtained from the ontological knowledge base or multimedia web services.

4.1 Multimodal access to web services

To connect to web services we developed a semantic representation formalism based on OWL-S and a service composition component able to interpret an ontological user query. We extended the OWL-S ontologies to flexibly compose and invoke web services on the fly, gaining sophisticated representation of information gathering services fundamental to SMARTWEB.

Sophisticated data representation is the key for developing a composition engine that exploits the semantics of web service annotation and query representation. The composition engine follows a plan-based approach as explained, e.g., in [Ghallab *et al.*, 2004]. It infers the initial and goal state from the semantic representation of the user query, whereas the set of semantic web services is considered as planning operators. The output gained from automatic web service invocation is represented in terms of instances of the SMARTWEB domain ontologies and enriched by additional media instances, if available. Media objects are represented in terms of the SMARTMEDIA ontology (see above) and are annotated automatically during service execution. This enables the dialog manager for multimodal interaction with web service results.

A key feature of the service composition engine is to detect underspecified user queries, i.e., the lack of required web service input parameters. In these cases the composition engine is able to formulate a clarification request as specified within the discourse ontology (DISCONTO). This points out the missing pieces of information to be forwarded to the dialog manager. Then the composition engine expects a clarification reponse enabling it to replan on the refined ontological user query.



(a) User query: What can I do in my spare time on Saturday?
(b) Ontological user query is sent to web services.
(c) Clarification request (asking for a city) is sent back.
(d) Verbalized clarification request: Where?
(e) User clarification response: In Berlin.
(f) Completed ontological query is sent to web services.
(g) Ontological result of service execution is sent to dialog.
(h) Generated results are multimodally presented to the user.

Figure 3: Data flow for the processing of a clarification request as in the example (7-10) "What can I do in my spare time on Saturday?".

According to the interaction example (7-10) the composition engine searches for a web service demanding for activity event types and gets its description. Normally, the context module incorporated in the dialog manager would complete the query with the venue obtained from a GPS receiver attached to the handheld device. In case of no GPS signal, for instance indoors, the composition engine asks for the missing parameter (cf. figure 3), which makes the composition engine more robust and thus more suitable for interactive scenarios.

In the interaction example (7-10) the composition planner considers the *T-Info EventService* appropriate for answering the query. This service requires both date and location for looking up events. While the date is already mentioned in the initial user query, the location is being asked from the user by clarification request. After the location information (dialogue step (9) in the example:*In Berlin*) is obtained from the user, the composition engine invokes in turn two T-Info (DTAG) web services[8] offered by Deutsche Telekom AG (see also [Ankolekar *et al.*, 2006]): first the *T-Info EventService* as already mentioned above, and then the *T-Info MapService* for calculating an interactive map showing the venue as point-of-interest. Text-based event details, additional image material, and the location map are semantically represented (the map in MPEG7) and returned to the dialog engine.

5 Semantic parsing and discourse processing

Semantic parsing and other discourse processing steps are reflected on the interaction device as advanced user perceptual feedback functionality. The following screenshot illustrates the two most important processing steps for system-user interaction, the feedback on the natural language understanding step and the presentation of multimodal results. The semantic parser produces a semantic query (illustrated on the left in figure 4), which is presented to the user in nested attribute-value form. The web service results (illustrated on the right in figure 4) for the interaction example (7-10) are presented in a multimodal way, combining text, image, and speech: *5 Veranstaltungen* (five events).

Figure 4: Semantic query (illustrated on the left) and web service results (illustrated on the right).

[8] http://services.t-info.de/soap.index.jsp

5.1 Language understanding with SPIN and text generation with NIPSGEN

The parsing module is based on the semantic parser SPIN [Engel, 2005]. A syntactic analysis of the input utterance is not performed, but the ontology instances are created directly from word level. The typical advantages of a semantic parsing approach are that processing is faster and more robust against speech recognition errors and disfluencies produced by the user and the rules are easier to write and maintain. Also, multilingual dialog systems are easier to realize as a syntactic analysis is not required for each supported language. A disadvantage is that the complexity of the possible utterances is somewhat limited, but this is acceptable for most dialog systems.

One outstanding feature of the parser is the possibility for order-independent matching, i.e., the order of elements in the input stream is ignored if order-independent matching is active. This simplifies the processing of free-word order languages like German and increases the robustness. Order-independent matching can have an huge impact on performance as parsing in general becomes an NP-complete task [Huynh, 1983]. To ensure fast processing notwithstanding, several off-line optimizations, like rule ordering, have been implemented which increase the performance for rule sets that are typical for dialog systems. The average processing time is about 50ms per utterance, which ensures direct feedback to user inputs.

The knowledge base of the parser consists of 544 rules and 2250 lexicon entries currently. To give an impression how the rules look like, four rules are provided as examples to process the utterance *When was Brazil world champion*. The first one transforms the word *Brazil* to the ontology instance `Country`:

```
Brazil → Country(name:BRAZIL)
```

The second one transforms countries to teams as each country can stand for a team in our domain:

```
$C=Country() → Team(origin:$C)
```

The third one processes *when* generating an instance of the type `TimePoint` which is marked as questioned:

```
when →
TimePoint(variable:QEVariable(focus:text))
```

The fourth rule processes the verbal phrase <*TimePoint*> *was* <*Team*> *world champion*

```
$TP=TimePoint() was $TM=Team() world
   champion →
QEPattern(patternArg:Tournament(
            winner:$TM, happensAt:$TP))
```

The text generation module uses the same SPIN parser that is used in the language understanding module together with a TAG grammar which is modelled similar to the XTAG grammar[9]. The input of the generation module are instances of SWIntO representing the search results. Then these results are verbalized in different ways, e.g., as heading, as row of a table or as text which is synthesized. A processing option indicates the current purpose.

[9] http://www.cis.upenn.edu/∼xtag/

The input is transformed to an utterance in four steps:

1. An intermediate representation is built up on a phrase level. The required rules are domain dependent.
2. A set of domain independent rules transforms the intermediate representation to a derivation tree for the TAG-grammar.
3. The actual syntax tree is constructed using the derivation tree. After the tree has been built up, the features of the tree nodes are unified.
4. The correct inflections for all lexical leafs are looked up in the lexicon. Traversing the lexical leafs from left to right produces the result text.

In the SMARTWEB system currently 179 domain dependent generation rules and 38 domain independent rules are used.

5.2 Multimodal discourse processing with FADE

An important aspect of SMARTWEB is its context-aware processing strategy. All recognized user actions are processed with respect to their situational and discourse context. A user is thus not required to pose separate and unconnected questions. In fact, she might refer directly to the situation, e.g., *"How do I get to Berlin from here?"*, where *here* is resolved to GPS information, or to previous contributions (as in the elliptical expression *"And in 2002?"* in the context of a previously posed question *"Who won the Fifa World Cup in 1990?"*). The interpretation of user contributions with respect to their discourse context is performed by a component called *Fusion and Discourse Engine*—FADE [Pfleger, 2005][10]. The task of FADE is to integrate the verbal and nonverbal user contributions into a coherent multimodal representation to be enriched by contextual information, e.g., resolution of referring and elliptical expressions.

The basic architecture of FADE consists of two interweaved processing layers: (1) a production rule system—PATE—that is responsible for the reactive interpretation of perceived monomodal events, and (2) a discourse modeler—DiM—that is responsible for maintaining a coherent representation of the ongoing discourse and for the resolution of referring and elliptical expressions.

In the following two subsections we will briefly discuss some context-related phenomena that can be resolved by FADE.

Resolution of referring expressions

A key feature of the SMARTWEB system is that the system is capable of dealing with a broad range of referring expressions as they occur in natural dialogs. This means the user can employ deictic references that are accompanied by a pointing gesture (such as in *"How often did this team [pointing gesture] win the World Cup?"*) but also—if the context provides enough disambiguating information—without any accompanying gestures (e.g., if the previous question is uttered in the context of a previous request like *"When was Germany World Cup champion for the last time?"*).

[10] The situational context is maintained by another component called *SitCom* that is not discussed in this paper.

Moreover, the user is also able to utter time deictic references as in *"What's the weather going to be like tomorrow?"* or *"What's the weather going to be like next Saturday?"*.

Another feature supported by FADE is the resolution of *cross modal* spatial references, i.e., a spoken reference to visually displayed information. The user can refer, for example, to an object that is currently displayed on the screen. If a picture of the German football team is displayed, the system is able to resolve references like *"this team"* even when the team has not yet been mentioned verbally. MPEG7-annotated images (see section 4) even permit spatial references to objects displayed within pictures, e.g., as in *"What's the name of the guy to the right of Ronaldo?"* or *"What's the name of the third player in the top row?"*.

Resolution of elliptical expression

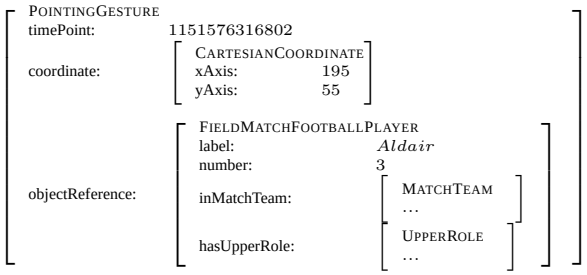
Humans tend to keep their contributions as short and efficient as possible. This is in particular the case for follow-up questions or answers to questions. Here, people often make use of elliptical expressions, e.g., when they ask a follow-up question *"And the day after tomorrow?"* in the context of a previous question *"What's the weather going to be like tomorrow?"*. But even for normal question-answer pairs people tend to omit everything that has already been conveyed by the question (User: *"Berlin"* in the context of a clarification question of the system like *"Where do you want to start?"*; see section 4.1).

Elliptical expressions are processed in SMARTWEB as follows: First, SPIN generates an ontological query that contains a semantic representation of the elliptical expression, e.g., in case of the aforementioned example "Berlin". This analysis would only comprise an ontological instance representing the city Berlin. FADE in turn, then tries to integrate the elliptical expression with the previous system utterance, if this was a question. Otherwise it tries to integrate the elliptical expression with the previous user request. If the resolution succeeded, the resulting interpretation either describes the answer to the previous clarification question, or it describes a new question.

5.3 Reaction and presentation planning for the Semantic Web

Integral part of dialog management is the reaction and presentation module (REAPR). It manages the dialogical interaction for the supported dialog phenomena such as flexible turn-taking, incremental processing, and multimodal fusion of system output. REAPR is based on a finite-state-automaton and information space (IS). Our new approach differs from other IS approaches (e.g. [Matheson *et al.*, 2000]) by generating IS features from the ontological instances generated during dialog processing [Sonntag, 2006]. [11]

Since the dialog ontology is a model for multimodal interaction, multimodal MPEG7 result representations, multimodal result presentations, dialog state, and (agent) communication with the backend knowlege servers, large information spaces can be extracted from the ontological instances describing the system and user turns in terms of special dialog acts - to ensure accurate dialog management capabilities. REAPR decides, for example, if a semantic query is accepted for transfer to the Semantic Mediator. The IS approach to dialog modeling comprises, apart from dialog moves and update strategies, a description of informational components (e.g. common ground) and their formal representations. Since in REAPR the formal dialog specification consists of ontological structures as Semantic Web data structures, a formal well-defined complement to previous formal logic-based operators and Discourse Representation Structures (DRS) is provided. However, the ontological structures resemble typed feature structures (TFS) [Carpenter, 1992] we use for illustration further down. During interaction, many message transfer processes take place, mainly for query recognition and query processing, all of which are based on Semantic Web ontological structures, and REAPR is involved in many of them. Here we give an example of ontological representations of user pointing gestures (dialog step (5) in the interaction example) which are obtained from the PDA and transformed into ontology-structures to be used by the input fusion module. The following figure shows the ontological representation of a pointing gesture as TFS.



It is important to mention that dialog reaction behaviour within SMARTWEB is governed by the general QA scenario, which means that almost all dialog and system moves relate to questions, follow-up questions, clarifications, or answers. As these dialog moves can be regarded as adjacency pairs, the dialog behaves according to some finite state grammar for QA, which makes up the automaton part (FSA) in REAPR. The finite state approach enhances robustness and portability and allows to demonstrate dialog management capabilities even before more complex IS states are available to be integrated into the reaction and presentation decision process. The dialog component integration process is described in the next section.

5.4 Dialog component integration

In this section we will focus on issues of interest pertaining to the system integration. In the first instance dialog component integration is an integration on a conceptual level. All dialog manager components communicate via ontology instances. This assumes the representation of all relevant concepts in the foundational and domain ontologies – which is

[11] The IS state is traditionally divided into global and local variables which make up the knowledge state at a given time point. Ontological structures that change over time vastly enhance the representation capabilities of dialog management structures, or other structures like queries from which relevant features can also be extracted.

hard to provide at the beginning of the integration. In our experience, using ontologies in information gathering dialog systems for knowledge retrieval from ontologies and web services in combination with advanced dialogical interaction is an iterative ontology engineering process, which requires very disciplined ontology updates, since changes and extensions must be incorporated into all relevant components. The additional modeling effort pays off when regarding the strength of this Semantic Web technology for larger scale projects.

We first built up an initial discourse ontology for request-response concepts for data exchange with the Semantic Web sub-system. In addition, an ontological dialog act taxonomy has been specified, to be used by the semantic parsing and discourse processing modules. A great challenge is the mapping between semantic queries and the ontology instances in the knowledge base. In our system, the discourse (understanding) specific concepts have been linked up with the foundational ontology and, e.g., the sportevent ontology, and the semantic parser only builds up interpretations with SWIntO concepts. Although this limits the space of possible interpretations according to the expressivity of the foundational and domain ontologies, the robustness of the system is increased. We completely circumvent the problem of concept and relation similarity matching between conventional syntactic/semantic parsers and backend retrieval systems.

Regarding web services we transform the output from the web services, in particular maps with points of interest, into instances of the SMARTWEB domain ontologies for the same reasons of semantic integration. As already noted, ontological representations offer a framework for gesture and speech fusion when users interact with Semantic Web results such as MPEG7-annotated images and maps. Challenges in multimodal fusion and reaction planning can be addressed by using more structured representations of the displayed content, especially for pointing gestures, which contain references to player instances after integration. We extended this to pointing gesture representations on multiple levels in the course of development, to include representations of the interaction context, the modalities and display patterns used, and so on.

The primary aim is to generate structured input spaces for more context-relevant reaction planning to ensure naturalness in system-user interactions to a large degree. Currently, we experiment with the MDA's camera input indicating whether the user is looking at the device, to combine it with other indicators to a measure of user focus. The challenge of integrating and fusing multiple input modalities can be reduced by ontological representations, which exist at well-defined timepoints, and are also accessible to other components such as the semantic parser, or the reaction and presentation module.

6 Conclusions

We presented a mobile system for multimodal interaction with an ontological knowledge base and web services in a dialog-based QA scenario. The interface and content representations are based on W3C standards such as EMMA and RDF. The world knowledge shared in all knowledge-intensive components is based on the existing ontologies SUMO and DOLCE, for which we added additional concepts for QA and multimodal interaction in a discourse ontology branch.

We presented the development of the second demonstrator of the SMARTWEB system which was successfully demonstrated in the context of the Football World Cup 2006 in Germany. The SWIntO ontology now comprises 2308 concept classes, 1036 slots and 90522 instances.[12] For inference and retrieval the ontology constitutes 78385 data instances after deductions.[13] The answer times are in a 1 to 15 seconds time frame for about 90% of all questions. In general, questions without images and videos as answers can be processed much faster. The web service composer addresses 25 external services from traveling (navigation, train connections, maps, hotels), event information, points of interest (POIs), product information (books, movies), webcam images, and weather information.

The SMARTWEB architecture supports advanced QA functionalities such as flexible control flow to allow for clarification questions of web services when needed, long- and short-term memory provided by distributed dialog management in the fusion and discourse module and in the reaction and presentation module, as well as semantic interpretations provided by the speech interpretation module. This can be naturally combined with dialog system strategies for error recoveries, clarifications with the user, and multimodal interactions. Support for inferential, i.e., deductive reasoning, complements the requirements for advanced QA in terms of information- and knowledge retrieval. Integrated approaches as presented here rely on ontological structures and deeper understanding of questions, not at least to provide a foundation for result provenance explanation and justification. Our future plans on the final six month agenda include dialog management adaptations via machine learning and collaborative filtering of redundant results in our multi-user enviroment, and incremental presentation of results.

7 Acknowledgments

The research presented here is sponsored by the German Ministry of Research and Technology (BMBF) under grant 01IMD01A (SmartWeb). We thank our student assistants and the project partners. The responsibility for this papers lies with the authors.

References

[Allen *et al.*, 2000] James Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. An Architecture for a Generic Dialogue Shell. *Natural Language Engineering*, 6(3):1–16, 2000.

[Ankolekar *et al.*, 2006] Anupriya Ankolekar, Pascal Hitzler, Holger Lewen, Daniel Oberle, and Rudi Studer. Integrating semantic web services for mobile access. In *Proceedings of 3rd European Semantic Web Conference (ESWC 2006)*, 2006.

[12] The SWIntO can be downloaded at the SMARTWEB homepage for research purposes.

[13] The original data instance set was 175293 instances, but evoked processing times up to two minutes for single questions by what interactivity was no longer guaranteed.

[Benitez *et al.*, 2002] Ana B. Benitez, Hawley Rising, Corinne Jorgensen, Ricardo Leonardi, Alesandro Bugatti, Koiti Hasida, Rajiv Mehrotra, A. Murat Tekalp, Ahmet Ekin, and Toby Walker. Semantics of Multimedia in MPEG-7. In *IEEE International Conference on Image Processing (ICIP)*, 2002.

[Bontcheva *et al.*, 2004] Kalina Bontcheva, Valentin Tablan, Diana Maynard, and Hamish Cunningham. Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, 10, 2004. Special issue on Software Architecture for Language Engineering.

[Carpenter, 1992] B. Carpenter. The logic of typed feature structures, 1992.

[Cheyer and Martin, 2001] Adam J. Cheyer and David L. Martin. The Open Agent Architecture. *Autonomous Agents and Multi-Agent Systems*, 4(1–2):143–148, 2001.

[Cimiano *et al.*, 2004] Philipp Cimiano, Andreas Eberhart, Pascal Hitzler, Daniel Oberle, Steffen Staab, and Rudi Studer. The smartweb foundational ontology. Technical report, (AIFB), University of Karlsruhe, Karlsruhe, Germany, 2004. SmartWeb Project.

[Engel, 2005] Ralf Engel. Robust and efficient semantic parsing of free word order languages in spoken dialogue systems. In *Proceedings of 9th Conference on Speech Communication and technology*, Lisboa, 2005.

[Fensel *et al.*, 2003] Dieter Fensel, James A. Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press, 2003.

[Gangemi *et al.*, 2002] Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schcneider. Sweetening Ontologies with DOLCE. In *In 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, volume 2473 of Lecture Notes in Computer Science, page 166 ff, Sigünza, Spain, Oct. 1–4 2002.

[Ghallab *et al.*, 2004] Malik Ghallab, Dana Nau, and Paolo Traverso. *Automated planning*. Elsevier Kaufmann, Amsterdam, 2004.

[Herzog *et al.*, 2004] Gerd Herzog, Alassane Ndiaye, Stefan Merten, Heinz Kirchmann, Tilman Becker, and Peter Poller. Large-scale Software Integration for Spoken Language and Multimodal Dialog Systems. *Natural Language Engineering*, 10, 2004. Special issue on Software Architecture for Language Engineering.

[Hovy *et al.*, March 2001] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. Towards semantic-based answer pinpointing. In *Proceedings of Human Language Technologies Conference, San Diego CA*, pages 339–345, March 2001.

[Hunter, 2001] Jane Hunter. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In *Proceedings of the International Semantic Web Working Symposium (SWWS)*, 2001.

[Huynh, 1983] Dung T. Huynh. Communicative grammars: The complexity of uniform word problems. *Information and Control*, 57(1):21–39, 1983.

[Krotzsch *et al.*, 2006] Markus Krotzsch, Pascal Hitzler, Denny Vrandecic, and Michael Sintek. How to reason with OWL in a logic programming system. In *Proceedings of RuleML'06*, 2006.

[Matheson *et al.*, 2000] C. Matheson, M. Poesio, and D. Traum. Modelling grounding and discourse obligations using update rules. In *Proceedings of NAACL 2000*, May 2000.

[Niles and Pease, 2001] Ian Niles and Adam Pease. Towards a Standard Upper Ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine, October 17–19 2001.

[Oviatt, 1999] Sharon Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.

[Pfleger, 2005] Norbert Pfleger. Fade - an integrated approach to multimodal fusion and discourse processing. In *Proceedings of the Dotoral Spotlight at ICMI 2005*, Trento, Italy, 2005.

[Reithinger and Sonntag, 2005] Norbert Reithinger and Daniel Sonntag. An integration framework for a mobile multimodal dialogue system accessing the semantic web. In *Proc. of Interspeech'05*, Lisbon, Portugal, 2005.

[Reithinger *et al.*, 2005a] Norbert Reithinger, Simon Bergweiler, Ralf Engel, Gerd Herzog, Norbert Pfeger, Massimo Romanelli, and Daniel Sonntag. A Look Under the Hood Design and Development of the First SmartWeb System Demonstrator. In *Proceedings of 7th International Conference on Multimodal Interfaces (ICMI 2005)*, Trento, Italy, October 04-06 2005.

[Reithinger *et al.*, 2005b] Norbert Reithinger, Dirk Fedeler, Ashwani Kumar, Christoph Lauer, Elsa Pecourt, and Laurent Romary. MIAMM - A Multimodal Dialogue System Using Haptics. In Jan van Kuppevelt, Laila Dybkjaer, and Niels Ole Bernsen, editors, *Advances in Natural Multimodal Dialogue Systems*. Springer, 2005.

[Seneff *et al.*, 1999] Stephanie Seneff, Raymond Lau, and Joseph Polifroni. Organization, Communication, and Control in the Galaxy-II Conversational System. In *Proc. of Eurospeech'99*, pages 1271–1274, Budapest, Hungary, 1999.

[Sonntag and Romanelli, 2006] Daniel Sonntag and Massimo Romanelli. A multimodal result ontology for integrated semantic web dialogue applications. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy, May 24–26 2006.

[Sonntag, 2005] Daniel Sonntag. Towards interaction ontologies for mobile devices accessing the semantic web - pattern languages for open domain information providing multimodal dialogue systems. In *Proceedings of the workshop on Artificial Intelligence in Mobile Systems (AIMS). 2005 at MobileHCI*, Salzburg, 2005.

[Sonntag, 2006] Daniel Sonntag. Towards combining finite-state, ontologies, and data driven approaches to dialogue management for multimodal question answering. In *Proceedings of the 5th Slovenian First International Language Technology Conference (IS-LTC 2006)*, 2006.

[Thorisson *et al.*, 2004] Kristinn R. Thorisson, Christopher Pennock, Thos List, and John DiPirro. Artificial intelligence in computer graphics: A constructionist approach. *Computer Graphics*, pages 26–30, February 2004.

[Wahlster, 2000] Wolfgang Wahlster, editor. *VERBMOBIL: Foundations of Speech-to-Speech Translation*. Springer, 2000.

[Wahlster, 2003] Wolfgang Wahlster. SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In R. Krahl and D. Günther, editors, *Proc. of the Human Computer Interaction Status Conference 2003*, pages 47–62, Berlin, Germany, 2003. DLR.

[Wahlster, 2004] Wolfgang Wahlster. SmartWeb: Mobile Applications of the Semantic Web. In Peter Dadam and Manfred Reichert, editors, *GI Jahrestagung 2004*, pages 26–27. Springer, 2004.