Interaction Design and Implementation for Multimodal Mobile Semantic Web Interfaces

Daniel Sonntag

German Research Center for Artificial Intelligence
66123 Saarbrücken, Germany
sonntag@dfki.de

Abstract. Multimodal mobile interfaces to the Semantic Web offer access to complex knowledge and information structures. In SmartWeb we try to build multimodal interfaces to answer very specific closed and open domain questions by natural language dialogue and multimedia presentations. Advanced user interactions such as pointing gestures are also supported. We present the interaction design and its implementation for Semantic Web related knowledge structures, i.e., ontology instances and relations, and follow the principle of no presentation without representation for information content, its presentation, and interaction possibilities. In particular, we address the challenge to display interactive text and image results obtained from multiple homogeneous and heterogeneous information sources.

Keywords: Multimodal Dialogue Systems, Interaction Design.

1 Introduction

In dialogue systems dialogue managers are accountable for resolving linguistic phenomena such as anaphoric references and ellipses, handling topic and focus, keeping track of dialogue state and interaction, and planning next dialogue moves. In multimodal dialogue systems multimodal dialogue acts and multimodal interaction have to be taken into account, too. Not surprisingly interaction and presentation behaviour plays a major role in multimodal dialogue systems, since the desired intention is presented in different modalities, or the information content comprises of related multimedia artefacts. In SmartWeb [1] we try to build multimodal interfaces to the Semantic Web, to answer very specific closed and open domain questions by natural language dialogue and display multimedia presentations. Our primary goal in SmartWeb is to support multimodal interactions and multimedia presentations on portable mobile devices, such as a PDA, to answer natural language questions about football events, players, matches, and so on. The application scenario we report on in this contribution is a visit to a football game at the Football World Championship 2006. See figure 1 for an example question and the multimedia result.

M.J. Smith, G. Salvendy (Eds.): Human Interface, Part II, HCII 2007, LNCS 4558, pp. 645–654, 2007.

Fig. 1. On the left side: the speech recognition result of the user query *Who got a red card in the final ?* and its semantic interpretation. On the right side: the multimodal answer of the question answering process is a factoid answer snippet (synthesised) with additional supportive image evidence.

Question Answering (QA) in mobile domain means for us, for first, lack of computational power, for second, a small screen, and for third, very specific and interesting interaction possibilities such as pointing gestures on screen and the use of device buttons. In the scenario we envision the primary input modality is speech and the multimedia output is presented on the PDA screen, with additional speech synthesis of the most important information. In most cases, the multimedia results are text snippets, text documents, images, videos, and graphics. In order to support more advanced interaction possibilities, the user should be able to navigate through the semantically-represented answers, point to individual media items or other screen elements, and pose follow-up questions by speech. The following example illustrates some typical dialogue based interactions we support:

(1) *User:* "When was Germany world champion?"
(2) *System*: "In the following 4 years: 1954 (in Switzerland), 1974 (in Germany), 1990 (in Italy), 2003 (in USA)"
(3) *User*: "And Brazil?"
(4) *System:* "In the following 5 years: 1958 (in Sweden), 1962 (in Chile), 1970 (in Mexico), 1994 (in USA), 2002 (in Japan)" + [team picture, MPEG7-annotated]
(5) *User:* *Pointing gesture on player* Aldair + "How many goals did this player score?"

(6) *System:* "Aldair scored none in the championship 2002."
(7) *User:* "What can I do in my spare time on Saturday?"
(8) *System*: "Where?"
(9) *User:* "In Saarbruecken"
(10) *System: The cinema program, festivals, and concerts in Saarbruecken are listed.*

In order to allow these interaction possibilities, we developed a new dialogue system framework complementing other approaches to dialogue system architectures (e.g. [2–4]). Standardised interface descriptions (EMMA[1], SSML[2], RDF[3], OWL-S[4], WSDL[5], SOAP[6], MPEG7[7]) are used between the modules and components. Basically, we want to allow the user to send requests to various information services that are linked by a Semantic Web framework. The *dialogue server* [5] integrates the dialogue components, the speech interpretation component [6], the modality fusion and discourse component [7], the reaction and presentation component [8], and the natural language generation module [9] are the most important for interaction and presentation. The text is organised as follows: the core user interface, mainly its interaction and presentation part, is described in section 2. Consistent and homogeneous presentation for multiple answer streams is the result of ontological representation of common sense interaction and presentation knowledge patterns in interactive mobile domains. Compared to previous existing approaches to media representation and media allocation, we base all decision processes and presentation items on ontological structures as part of a discourse ontology, or upper-level ontologies, or domain ontologies. We conclude by describing the implementation process in short and the improvements we build in while implementing the interaction storyboard (section 3). Lack of computational power for task scheduling on mobile device processors remains an open problem for the synchronisation of speech and text output, and advanced graphics on the device. Content-planning is heavily influenced by previously selected media and modalities for presentation or interaction. We motivate the use of ontologies and Semantic Web data structures [10] for multimodal interaction design and implementation—spinning the Semantic Web for mobile dialogue system applications.

2 Core User Interface

The user interacts with the dialogue system by using a mobile PDA (T-Mobile's MDA pro) as core user interface; we explain the graphical screen interface in the following: In ontology-based interaction design for mobile devices [11], software development for the mobile context [12] has to be tailored toward ontology-driven

[1] http://www.w3.org/TR/EMMAreqs
[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/

discourse development for the mobile context. On the PDA screen (640 x 480 pixels) we defined three regions for the graphical user interface (figure 2), an upper region, a middle region, and a lower region. The title bar with status indications and corresponding icons are displayed in the upper region. Here we also display a direct feedback to the users input and the recognised words are presented with a direct editing function (cf. figure 4). We also display whether the microphone is on/off and the progress bar. In the middle region, we display the paraphrased question (cf. figure 3) in a semantic template-based or generated sentence-based form before the request is being sent to the Semantic Web access systems (see [13] for further information). In the lower section, we display result navigation buttons and a status bar. This means, for example, to display icons for result media types, a legend of symbols and colours when dynamic button allocation is used, or system control and system status information.

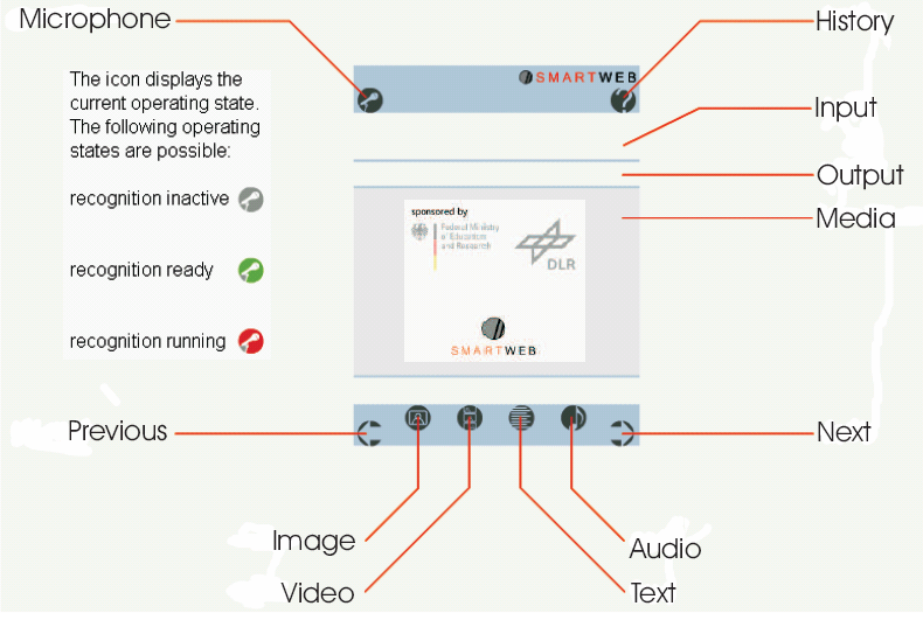

Fig. 2. Main GUI Widgets. Visit `http://smartweb.dfki.de/Intro_Demo/start.html` for an interactive demonstration.

Perceptual feedback allows the user to understand the current processing state of the system, in order to react accordingly. Although we do not use a life-like character like Smartakus [14], system activity and responsiveness can be expressed by liveliness of the presentation elements. Lively metaphors for system states that are used as perceptual states are (1) *Listening/Idle State*, (2) *Recording* (green and red icon), (3) *Understanding* by presenting a semantic query interpretation, (4) *Query Processing* (status bar), (5) *Presentation Planning*, and (6) *Presenting*.

The challenge we address is to give feedback on speech recognition and language understanding accuracy. We experimented with displaying best hypothesis confidences as probability values, or confidence scores in terms of discrete categorical classes (good, medium, bad) to indicate the status of the recognition process. The automatic speech recognition (ASR) component delivers these values as metadata. As informal user tests show, to simply present words with low probability in different colour works best as ASR feedback. Analogously, process metadata of language understanding accuracy is provided in terms of the semantic query interpretation we call semantic paraphrase. The dialogue management component [8] decides on-the-fly whether a verification dialogue is to be initiated, or a confirmation is needed, or the query is being sent without any confirmation. In this way, we try to obtain optimised dialogue prompts in specific data situations which is one of the main duties for dialogue managers. We give priorities to (1) users feedback from multimodal input parsing, (2) offer correction possibilities, (3) provide interface simplicity by progressive disclosure and a special link structure. We regard paraphrasing the natural language understanding output and displaying the proposed query as user feedback as the key element to the system's acceptability (figure 1 and 3). The language understanding module is used to analyse the input and to build up the internal semantic representation of the user utterance. This representation is in the same format as the Semantic Web knowledge bases. The textual semantic paraphrase is then generated from this query description. The ontological RDF/S structures resemble intentionally typed feature structures (TFS) [15] common in formal natural language processing. An example semantic query of the user enquiry: Show me the mascots of the Football World Cup is shown in figure 3.

Fig. 3. Top Left: Textual semantic paraphrase as ontology concepts of winner team search and the MPEG7 image result shown on the PDA. Top Middle: Semantic Paraphrase of Webcam search and Web Service results . Top Right: Paraphrase of the world cup mascot search and the ontological query representation.

Fig. 4. On the left: User input corrections on *Werner (Who) was world champion in 1990 ?* On the right: Concept-icons present feedback of question understanding (a team instance is being asked) and answer presentation in a language-independent, conceptual way. The sun icon additionally complements a textual weather forecast result and conveys weather condition information.

2.1 Correction Possibilities and Focus Attention

One very important question concerning the user interaction model is how to correct invalid user input from automatic speech recognition errors[8], or from errors that occur while interpreting a user utterance. This becomes more serious in the context of composite multimodality, where the dialogue system must understand and represent the multimodal input. According to our interaction storyboard, the user should be able to correct input in the following way: As soon as the speech recognition result and the semantic paraphrase is presented, the user is able to correct the speech recognition result using handwriting recognition or the Qwerty keypad, thereby initiating a new language understanding step (see the left part of figure 4). Focus attention, to direct the user's attention to a special region on the screen, is addressed in the following way: According to [16], multiple attention calling items should be avoided, except for two situations: when they complement each other, or they express the same information. Combining different media as focus (text with graphic) we try to avoid that an user is either distracted or confused by redundant information—whereby redundancy itself is used to express the high importance of the understanding process of a concept which is being asked. We include redundant graphic information which presents easily-to-understand, language-independent information which additionally pleases the eye, at least by colouring a system response dominated by written

[8] The accuracy of our recogniser started server-side is over 90%. Using out-of-vocabulary words, the accuracy, however, decreases drastically.

language. People do accept multimodal presentations which intersperse text and graphic, otherwise comic strips would not have become so popular. Figure 4 shows concept icons for system understanding feedback and presentation of results. In addition, we deal with layout as a rhetorical force, influencing the intentional and attentional state of the discourse participants [17]. As shown in figure 1, the reader is presented the multimodal paraphrase in the central display section, from upper left to lower right. Text is naturally presented from left to right. In addition to the textual speech recognition result, the textual part of the paraphrase is presented first. The concept icons are always placed in the lower right area, in order to exploit the rendering surface and to balance out the general text view from the left.

2.2 Navilink Structures

Having received a result list of multimodal items, we address the question, what kind of response is appropriate to be presented. Following the last example, the names of the mascots can be synthesised, and additional textual and image material can be presented on screen. The problem we address is how to display extensive text and image information on the small PDA screen. We implemented a hyperlink structure (*Navilinks*) to provide an interactive result summary and additional multimedia material through navigation. Navilink structures are generated automatically from the ontological descriptions (figure 5) obtained as results from the ontology servers. The constraints coded into the presentation ontology, augmented with result filtering and aggregation mechanisms, provide the automatic layouts and interaction possibilities shown in figure 5.

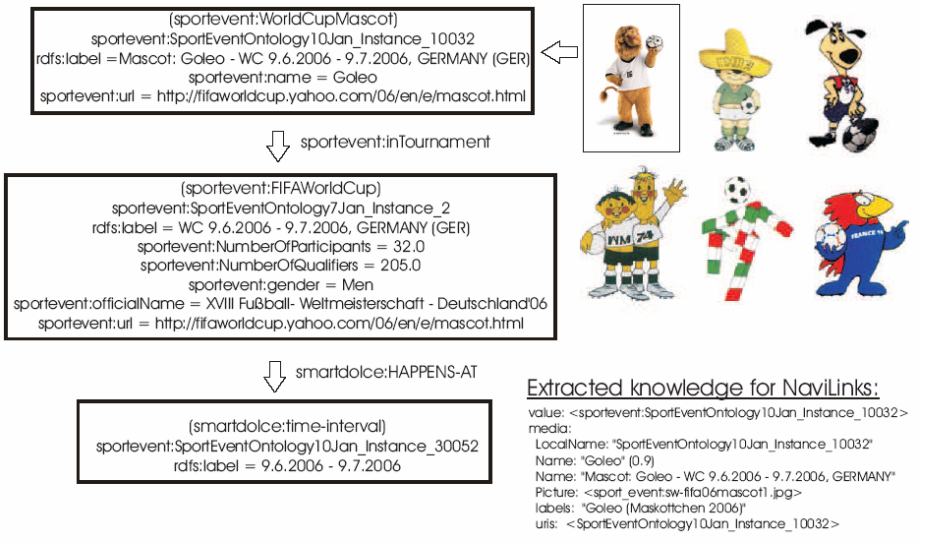

Fig. 5. On the left: the fragment of the ontology results for the 2006 World Cup mascot *Goleo*. The extracted knowledge for result presentation and link generation is shown on the right.

In the QA scenario a summary of the QA process should be displayed on the screen for fast visual inspection at first sight (*11 mascots*). Lateral navigation for displaying numerous parallel results must be applicable in this presentation situation, i.e., presenting the result of enumeration questions. Limiting the number of vertical layers of presentation elements (hierarchical links) by surface links such as *Help* and *Home* buttons which are active on each result page, is a common practice we adopted for our needs. The Navilink generator similar to hyperlink generation for desktops [18] takes the extracted knowledge for Navilinks and produces the level-1 hierarchical link structure shown in the middle region of the left screenshots in figure 6. The user is able to focus on the special result instance *Goleo* through (a), (b), and (c). The most suitable word within a text that contains enough content information as a Navilink headword is chosen by special natural language parsing and generation rules [9]. Pointing gesture on the screen allow users to activate Navilinks in red colour, or other textual material which is displayed on the screen for navigation and linguistic discourse reference. Pointing gestures can be used to complement language input as shown in the interaction example (utterance 5). GUI interaction is additionally leveraged by meta dialogue which the user utters as speech commands, such as accept, reject, going back to previous turns and question (history of system-user turns is available), or starting over the dialogue session, which deletes all active discourse referents.

Fig. 6. Links can be followed by pointing gestures to focus on specific data items

3 Conclusions and Future Work

We presented the interaction design and implementation for multimodal mobile Semantic Web interfaces. With an interaction ontology we provide a semantic-based data integration framework for symmetric multimodal question processing. We evaluated in our mobile application scenario that

- audio-repetition of user queries is useless, only implicit or non-intrusive feedback in textual form should be given.
- extra input correction modes with bigger fonts are valuable for handwriting recognition applications on small devices.
- editing the semantic paraphrase remains a challenge. The current implementation only allows to edit the ASR output.
- building integrated ontological representation is very time-consuming (about 30% of implementation effort) and experts are needed.
- co-ordination of short textual answer and its synthesis is impressive on PDAs if it works. Unfortunately, due to the PDA processor charge, the synthesis start is delayed for about 3 seconds for some questions, which remains an open problem. Synthesising augmented speech, where the synthesis is a complete sentence, instead of a short answer, compensates the effect.
- it is beneficial to follow the principle: no multimodal presentation is generated without representation. By rigorously following this design principle we can refer to all presentation elements at input processing as a side-effect of profound data structure design. Ontologies provide semantic-based data integration frameworks for multimedia.

An accompanying evaluation with real users of the whole SmartWeb-system is planned. Extensions are editing functions via concurrent pen and voice, to extend symmetric multimodality to symmetric multimodal query correction. More flexibility and robustness in speech recognition and understanding is much appreciated, since users formulate questions in various styles and phrase order, even in closed domains, such as football. Even closed-domain applications have to deal with many different linguistic surface forms. Our future investigations will explore more fine-grained co-ordination of multimodal presentations in mobile environments, and graph-like visualisation of ontological result structures (Semantic Navigation).

Acknowledgements. The research presented here is sponsored by the German Ministry of Research and Technology (BMBF) under grant 01IMD01A (SmartWeb). We thank our student assistants, the project partners, and Ralf Engel and Norbert Reithinger for valuable comments on earlier versions. The responsibility for this papers lies with the author.

References

1. Wahlster, W.: SmartWeb: Mobile Applications of the Semantic Web. In: Dadam, P., Reichert, M. (eds.) GI Jahrestagung 2004, pp. 26–27. Springer, Heidelberg (2004)
2. Herzog, G., Ndiaye, A., Merten, S., Kirchmann, H., Becker, T., Poller, P.: Largescale Software Integration for Spoken Language and Multimodal Dialog Systems.Natural Language Engineering 10, Special issue on Software Architecture for Language Engineering (2004)
3. Wahlster, W.: SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In: Krahl, R., Günther, D. (eds.) Proceedings of the Human Computer Interaction Status Conference 2003, Berlin, Germany, DLR, pp. 47–62 (2003)

4. Wahlster, W. (ed.): SmartKom: Foundations of Multimodal Dialogue Systems. Springer, Berlin (2006)
5. Reithinger, N., Sonntag, D.: An integration framework for a mobile multimodal dialogue system accessing the Semantic Web. In: Proceedings of Interspeech'05, Lisboa, Portugal (2005)
6. Engel, R.: 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)
7. Pfleger, N.: Fade - an integrated approach to multimodal fusion and discourse processing. In: Proceedings of the Dotoral Spotlight at ICMI 2005, Trento, Italy (2005)
8. Sonntag, D.: 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)
9. Engel, R.: Spin: A semantic parser for spoken dialog systems. In: Proceedings of the 5th Slovenian First International Language Technology Conference (IS-LTC 2006)
10. Fensel, D., Hendler, J.A., Lieberman, H., Wahlster, W.: Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, Cambridge (2005)
11. Sonntag, D.: 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)
12. Longoria, R. (ed.): Designing software for the Mobile Context. A Practitioner's Guide. Springer, Heidelberg (2004)
13. Sonntag, D., Engel, R., Herzog, G., Pfalzgraf, A., Pfleger, N., Romanelli, M., Reithinger, N.: Smartweb handheld - multimodal interaction with ontological knowledge bases and semantic web services. In: Proceedings of the International Workshop on AI for Human Computing (AI4HC) in conjunction with IJCAI (2007)
14. Poller, P., Reithinger, N.: A state model for the realization of visual perceptive feedback in Smartkom. In: Proceedings of Interspeech 2004, pp. 265–268 (2004)
15. Carpenter, B.: The logic of typed feature structures (1992)
16. Arens, Y., Hovy, E., Vossers, M.: On the knowledge underlying multimedia presentations, pp. 280–306 (1993)
17. Wahlster, W.: Planning multimodal discourse. In: Proceedings of the 31st annual meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics, pp. 95–96 (1993)
18. Stock, O.: Human-computer interaction through natural language and hypermedia in Alfresco. SIGCHI Bulletin 28(3), 102–107 (1996)