Ontology driven voice-based interaction in mobile environment

Jiri Kopsa[1], Zdenek Mikovec[1], Pavel Slavik[1]

[1] Czech Technical University in Prague
Karlovo namesti 13, Prague 2, Czech Republic
j.kopsa@fee.ctup.cz, {xmikovec, slavik}@fel.cvut.cz

Abstract. In this paper we are presenting a prototype of a voice user interface that allows users to interact with the mobile knowledge management system, which is being developed at Czech Technical University in Prague. The primary application domain of the system is facility and construction site management. The knowledge handled by the system consists of a domain ontology, site plans and their semantic descriptions (application ontology). The site plans are stored in the SVG format; the semantic descriptions along with the domain ontology are stored in the OWL format. The knowledge (described by the corresponding ontology) is used for efficient control of the voice based interaction with the knowledge management system. The implementation of the voice user interface is based on the existing VoiceXML platform.

1 Introduction

The graphical information (e.g. construction site plans stored in SVG [7]) that is handled by our system needs to be accessed by workers with mobile devices while doing construction site inspection. However, their ability to interact with a mobile device in a common way (i.e. using touch screen and stylus) is in many cases restricted. They usually need to use the mobile device with hands and eyes free to perform other tasks like taking samples of materials, etc. We have designed a voice user interface to support this type of use cases. The graphical information is described in textual form - as an ontology (OWL [10]), which can be perceived as an oriented graph. In this case the voice based interaction is the solution to the above given problem. An ontology example is presented in the following figure (Figure 1).

The ontology consists of domain, application and linguistic ontologies. The domain ontology defines abstract terms, classes of objects and their relations (existing on the specific class of construction sites), whereas the application ontology specifies objects describing a particular construction site or facility.

In figure 1 the nodes and edges marked with dashed line represent the domain ontology and the remaining nodes and edges represent the application ontology. For example the room node has more general meaning than the main hall and office objects (the dashed edges represent links between abstract and concrete entity).

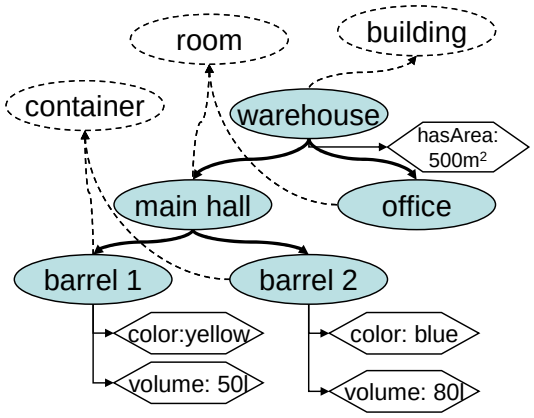

Figure 1. Application ontology example

The user can retrieve information from the ontology in user-initiative conversation with the system. The user forms questions to get information needed and the system answers. New information like material samples taken can be also created using simple commands.

The language that covers all possible questions about the whole ontology is in general very large and the speech recognition is rather problematic. Moreover, the question understanding is significantly affected by the noisy environment of workers in mobile environment.

1.1 Related Work (State of the Art)

The success of question understanding is often improved by restricting the language to a specific application domain [5][9]. In a broader sense contextual information is used in most of the voice applications to cope with the restriction process. The contextual information is obtained from various sources (sensors of user gestures, environment sensors) and used to restrict the language as well as to resolve the semantic ambiguity of the natural language as described in [2][3][4]. This approach does not provide the system with sufficiently detailed context information. There exist several approaches like [6] which try to build up more general multimodal interfaces for multiple applications. These approaches are based on the ontology description of the applications which should be automatically connected with the generic multimodal interface.

Our application is unique in the way that the contextual information is integrated with the application ontology. All necessary contextual information in sufficient detail is available to our system. Because of a link to the linguistic ontology it is possible to access the information in application ontology in natural language (in our case voice based communication). This approach is even more interesting when we take into account the future possibility to generate the ontology description automatically from the natural language [8].

1.2 Our Approach

The large size of the language needed to describe complex ontology is the main issue. We assume that since the worker is in a particular location with specific tasks in mind, he will probably use only a subset of that language.

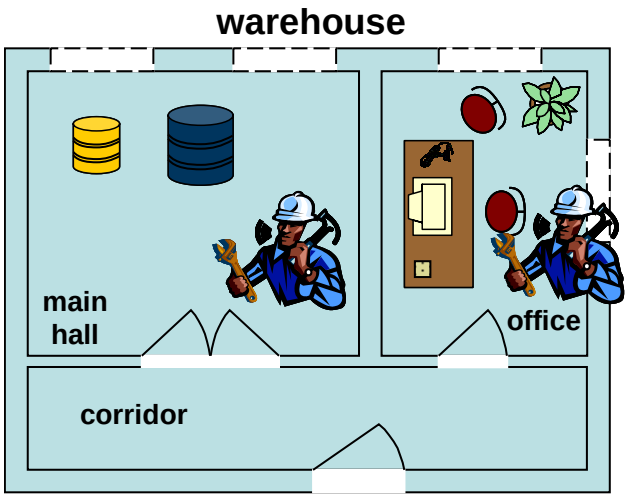

Figure 2. Use-case of construction site inspection: The worker location constrains the conversation context

Our approach is to use intensively such contextual information (contained in the ontology) to make the understanding process efficient. We are seamlessly constraining the language to understand the questions according to the current user context and conversation history. In every particular moment of the conversation only a relevant subset of the language can be used. Union of these particular languages represents the general language large enough to cover the general conversation about the given topic. The main challenge is to use the contextual information and to constrain the language in a way that the voice user interface is highly usable.

The figure 2 explains the presented approach. When the workers are in the main hall of the warehouse, they will most probably ask about the barrels that are contained in that room or about objects that are related to them. The probability that they will ask about the objects in the office is significantly lower in the given context.

2 Solution description

2.1 Ontologies

The ontology consists of objects, their attributes and relationships between them. The ontology is represented as an oriented graph with objects as nodes and properties as edges. An edge and its two nodes represent a fact in the form “Subject – Predicate – Object”.

For example, the edge between the nodes “warehouse” and “main hall” (Figure 1) represents the fact:

`Warehouse contains main hall.`

Object properties may be also transitive, e.g. the following fact is also represented in the graph:

`Warehouse contains barrel 1.`

From this example we can see how the ontology is structured and interpreted.

2.2 Conversation

As outlined in the introduction, the user can formulate questions to ask about the facts stored in the ontology. The question represents a query on the ontology (represented by a graph). There are usually several forms of the same query (the use of synonyms). The query is specified with two key items – source set and target - and several other less important attributes.

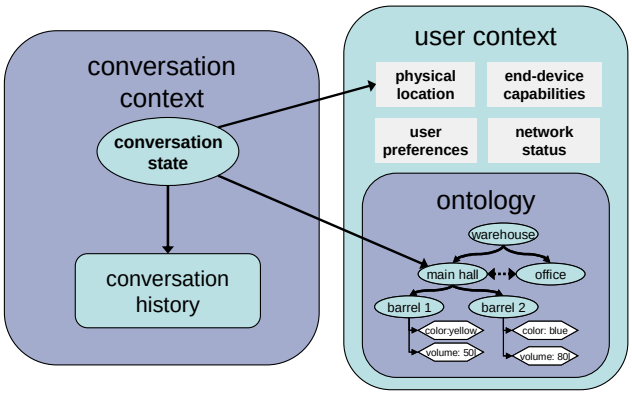

Figure 3. Conversation Context

The source set represents a node or set of nodes whose properties the users are interested in. The target specifies the properties, which is the user interested in. The two

synonymic questions in the following example represent the same query. Its source set contains one object – *main hall* – and target is the property *contains*.

H *(Human)*: What is contained in the main hall?

OR

H: List the contents of the main hall!

If the user question is successfully recognized, the corresponding query is executed and the result of its execution is formed into an answer. The system would respond in the following way:

C *(Computer)*: It contains two barrels.

In the process of the query execution the conversation context is updated and a new language based on it is generated. The conversation context holds the state of the conversation (current user context, pointer to the objects in the ontology and conversation history - see figure 3). Except the grammar production, the conversation context is also used to resolve question ambiguities. For example, the identifier *main hall* used in previous user questions may be ambiguous since there may be a number of objects with the name “main hall”. However, it can be resolved according to the current user location contained in the user context.

The conversation history is a queue that contains references to recently discussed objects of the ontology. It is used to resolve ambiguities of the natural language. For example, the conversation may continue in this way:

H: What is its area?

C: Its area is 500 m^2.

The pronoun *its* is resolved to object *main hall* by searching the conversation history. The notion of the conversation context is shown on the figure 3.

The user can also create objects by forming special commands:

H: Add new sample 123 to barrel one.

C: A sample 123 related to barrel one was created!

This is aimed to fulfill the use-case of taking samples - the worker wears gloves and takes samples with tools. In the same time the worker creates voice annotation describing the process of taking samples. The physical sample and the annotation are interlinked via the unique identification (in our case 123).

The range of allowed questions and commands is quite broad, e.g. the user may also specify a condition that must be met for all objects that are included in the answer:

H: List barrels contained in the main hall and manufactured by Liquids Ltd.

The detail specification of the whole language along with its semantics, query execution and answers formatting processes can be found in [1].

2.3 Towards the Natural Language

The language that can be recognized by a speech recognizer – the input language - is defined with the context free grammar based on the ontology and the conversation context. The fundamental role in this process have natural language annotations that define natural language attributes of the ontology objects (see figure 4). The domain ontology as well as the application ontology is annotated with natural language annotations whose types and classes are defined within linguistic ontology. These annotations are used to assign natural language identifiers and grammatical classes (e.g. grammatical gender) to ontology objects, to define synonymic verbs in various voices and tenses to object properties, etc. From these data, the clauses of the input language as well as the answers are constructed.

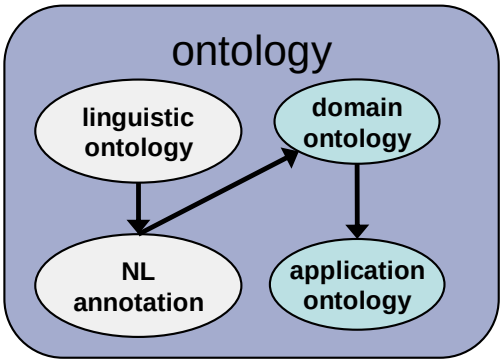

Figure 4. Structure of ontology description

We formally describe the input grammar as a function of three arguments:

```
G = F (O, UC, CH)
```

where `O` means ontology, `UC` means user context and `CH` conversation means history.

2.4 System Architecture

We have used the existing VoiceXML server platform for speech recognition and synthesis. It is configured to request VoiceXML documents from our system, which implements the interaction logic, conversation state management and ontology data retrieval. The mobile devices are connected to our system with Voice-over-IP clients.

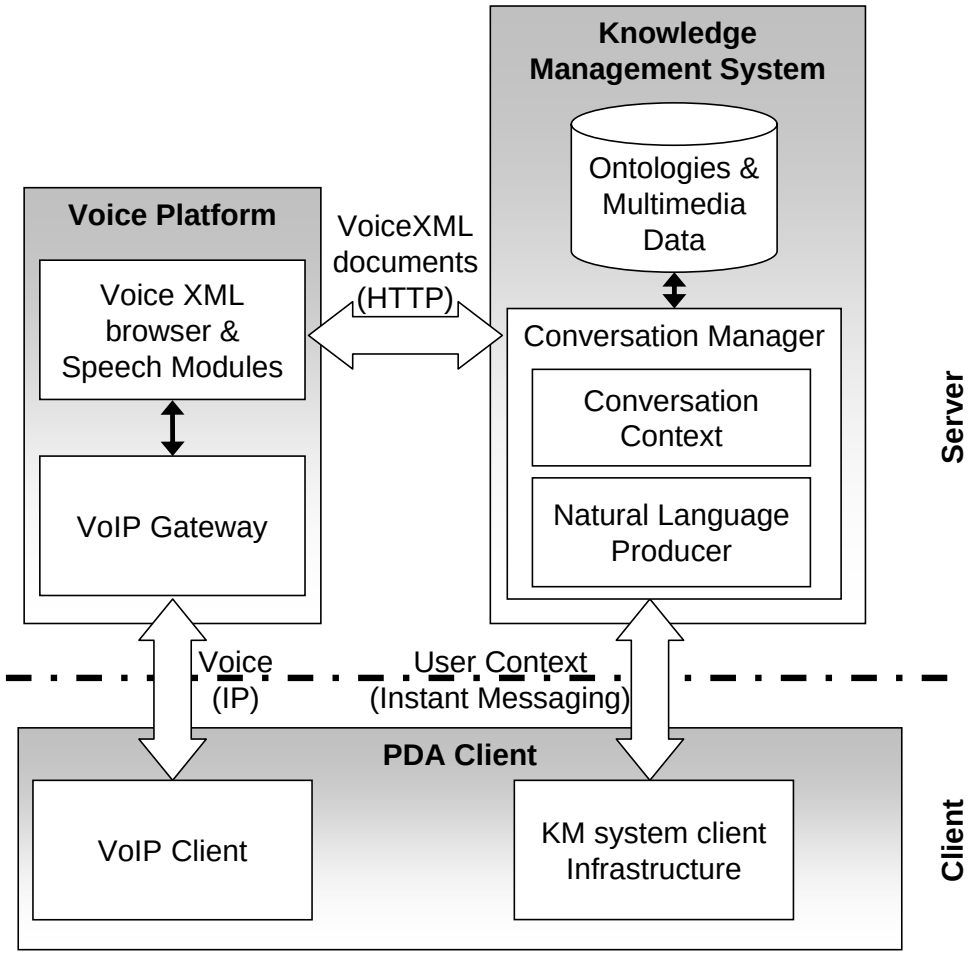

Figure 5. System Architecture

3 Testing

In this paper we are presenting an ongoing work, we are still in the process of development. We have implemented a prototype for the testing purposes. Our usability tests were focused on two aspects: first to determine the size of the input language where the speech recognition system will still be usable, second to test our hypothesis that during the conversation we are able to dynamically reduce the input language (in respect to the current user context and detailed application ontology description) in such a way, that the users will not be restricted in their questions (see chapters 1.2 and 2.3).

During the test we have determined the size of the input language where the speech recognition system was at least 90% successful in recognizing the user queries. The conversation context given by the user context, application ontology and conversation history was continuously changing and based on its current state the input language was dynamically generated to allow the natural language conversation with the user.

4 Conclusion

We have designed, implemented and tested a way of presenting graphical information with voice user interface. A crucial role has the usage of ontologies for storing data including natural language attributes and the usage of contextual information to improve the speech recognition rate and to resolve natural language ambiguities.

The hypothesis that the user needs changes accordingly to the conversation context determined by the application ontology and user context was proven and dynamically generated input language of the voice user interface matched the user needs during the communication.

However, there are still some issues that need to be addressed. We are currently in the process of finding a point with the optimal user experience in a compromise between the size of the input language and the recognition rate. We are also exploring recognition process improvements with an audio signal filtering and second phase semantics-driven selection of n-best results of the speech recognizer.

The goal is to develop voice-based user interface, which will be usable in the real environment of specific class of application. For this purpose we plan to perform a second set of usability tests that would simulate real work of a construction site inspector to find out whether input language being restricted in time really suites the needs of a real user.

Acknowledgement

The research is running in the framework of MUMMY project (Mobile knowledge management -- using multimedia-rich portals for context-aware information processing with pocket-sized computers in facility management and at construction site) and is funded by Information Society DG of European Commission (IST-2001-37365). See http://www.mummy-project.org/.

References

[1] Kopsa, J. *Voice User Interface for Multimodal Data*. Master Thesis, CTU in Prague, Prague, 2005.

[2] Leong, L. H., Kobayashi, S., Koshizuka, N., Sakamura, K., CASIS: *A Context-Aware Speech Interface System*. ACM IUI 2005, pages 231-238, 2005.

[3] Oviatt, S. *Advances in Robust Multimodal Interface Design*. IEEE Computer Graphics and Applications, pages 62-68, September 2003.

[4] Pfleger, N. *Context based multimodal fusion*, ACM ICMI'04, 2004.

[5] Ramakrishnan, I. V., Stent, A., Yang, G. *HearSay: Enabling Audio Browsing on Hypertext Content.* ACM WWW2004, pages 80-89, 2004.

[6] Reithinger, N., Alexandersson, J., Becker, T., Blocher, A., Engel, R., Löckelt, M., Müller, J., Pfleger, N., Poller, P., Streit, M., Tschernomas, V. SmartKom - adaptive and flexible multimodal access to multiple applications. ACM ICMI'03, pages 101-108, 2003.

[7] *Scalable Vector Graphics (SVG) 1.1 Specification*, http://www.w3.org/TR/2003/REC-SVG11-20030114/

[8] Shamsfardand, M., Barforoush, A., A.: *Learning ontologies from natural language texts.* International Journal of Human-Computer Studies 60, Elsevier, page 17-63, 2004

[9] Zue, V., JUPITER: *A telephone-based conversational interface for weather information*. IEEE Trans. on Speech and Audio Processing, Vol. 8, No.1 pp 100-112, 2000.

[10] *Web Ontology Language (OWL) Specification*, http://www.w3.org/2004/OWL/