

Spatial Cognition and Natural Language Interfaces in Mobile Personal Assistants

Christian Kray
DFKI GmbH
Stuhlsatzenhausenweg 3
D-66123, Saarbrücken
Germany

Robert Porzel
EML GmbH
Schloss-Wolfsbrunnenweg 33
D-69118 Heidelberg
Germany

June 14, 2000

Abstract

One of the main points, in which mobile systems differ from their stationary counterparts, is the fact that their location changes constantly and, thus, constitutes an important contextual factor. As the user moves around in the real world, her interaction methods with the mobile system will also change compared to the stationary settings, e.g. due to the lack of keyboards and high resolution displays. In this paper, we argue that there is a common set of services that can be derived from these two points, and that most mobile personal assistants should incorporate in order to exhibit a truly user-friendly behavior. We analyze these services, and show how they rely on spatial cognition and the presence of a natural language interface. We then present the *Deep Map* system and how it achieves the goal of providing the services identified beforehand.

1 Introduction

Within the Deep Map [6] research framework a first prototype of a mobile tourist information system was created and successfully tested in a real world scenario (a tourist visiting the castle of Heidelberg, Germany). During the development of the actual system, preliminary studies were conducted about what people would ask of a mobile tourist guide, and how they would word it. Together with the observations of the field test, these results helped us to identify two features that are very central to the success of such a system.

On the one hand, the relative importance of the individual interface modalities for mobile systems differs from stationary applications: the relative importance of the natural language interface increases while that of traditional graphical user interfaces decreases. This is a result of the fact that mobile users do not wish to have their visual attention distracted while driving or walking, or simply do not want to focus their visual attention on anything but their actual task. While augmented reality systems provide a means of overlaying reality with additional information, the resulting systems are sometimes just barely wearable [1]. Furthermore, especially head-mounted displays are often perceived as being obtrusive and annoying. While this applies certainly much less to PDAs (or arm-mounted LCDs), they are nevertheless distracting the user's visual attention. In contrast to these, headphones or mobile phones are seen as being rather convenient than obtrusive, and they do primarily allocate the auditory sense.

On the other hand the importance of spatial information, location awareness, spatial reasoning and spatial language¹ becomes paramount. Not only do many queries to a mobile information system contain (implicit or explicit) spatial references such as *How do I get to X?*, *What is that?*, or *Where is the next Y?*, but some queries do even only make sense in the actual locational context. Consider for an example a question such as *Can I still get tickets for tonight's show?* asked in front

¹The term *spatial language* is employed here in a general sense, meaning all text types such as spatial instructions, spatial descriptions as well as spatial references and localizations.

of a theater. If the theater was not mentioned recently, any system that does not take into account the current location will probably fail to answer appropriately.

In this paper, we argue that these two features - a natural language interface and spatial reasoning - are interrelated, and that their presence in a mobile personal assistant is very beneficial for the overall user experience. The following section looks more closely at these two points and why they are needed in many mobile systems. In section 3, we present the approach we took to address the resulting requirements when we built the Deep Map system. Section 4 concludes the paper, and gives an outlook on future research.

2 Talking about space

Despite the fact that the type and scope of information offered by mobile assistant systems will continue to differ greatly from application to application, we think that there are some core services that most users will come to expect from almost every mobile system. These services include taking into account the user's current position, telling her where she is, instructing her how to get to desired locations, and providing her with information about her surroundings. Since the user is always *located* in space, many of her queries arise from her current surroundings or events therein.

Furthermore, it is foreseeable that large efforts for collecting the necessary information underway within the geo-computing communities coupled with the standardization efforts of the Open GIS Consortium [7], will lead to electronically available spatial information, accessible in standardized formats. Ultimately, the consolidation of services and geo-coded information will converge to a standard profile of basic functionality for mobile information systems. Therefore, the issues arising in the context of the services mentioned above need to be addressed by most mobile systems.

2.1 Taking into account the user's current position

While the current position of the user is of lesser importance in the case of stationary systems, it becomes an important contextual factor for mobile systems. Not only does it place constraints on technical aspects such as bandwidth, network availability, etc. but first of all, it provides additional information for responding to the user's queries. Processes that can benefit from this include the resolution of (deictic) anaphora ("What's that?") and the understanding of the user's input ("What's the name of that building?"). On the other hand, if the current position is ignored by the system, this will sharply decrease its ease of use for the user, either by generating unnecessary questions (to gather missing information otherwise inferred from the current location) or by misinterpreting the user's query (e.g. by considering only previously mentioned objects for anaphora resolution and generation). Generally, this will lead to a behavior that differs greatly from a human guide.

2.2 Telling the user where she is

A point closely related to locational awareness, is the ability to inform the user about his/her current location. Even though currently existing systems already provide this service (e.g. GPS handhelds, car navigation systems) they mostly provide metrical information such as coordinates and longitude/latitude data. In order to really help the user to build a correct mental representation of her current position, additional factors need to be taken into account: proper reference objects must be selected, as well as spatial relations to them. In addition, the right level of granularity must be computed: does the user want to be informed about the city he's currently driving through, or about the name of the street?

The use of verbal communication of such positional information offers some distinct advantages over graphical representations such as maps (in addition to the ones mentioned in section 1). On the one hand, it is possible to transmit just the information that is required to solve the actual task ("You are right behind the castle") instead of providing additional unnecessary (and potentially confusing) information (such as large maps with a cross indicating the user's position). On the other hand, the granularity and exactness of the actual positional statement is easily encoded using verbal communication (e.g. by means of linguistic hedges [5]), which is not the case for graphical representations².

²While there are ways to express e.g. vagueness graphically (such as blurring, or abstraction), there is no generally accepted semantic to them, making them harder to understand.

2.3 Instructing the user how to get to desired locations

While these arguments also hold for route directions, there are several points specific to this kind of spatial statements. First of all, route directions usually describe several steps that need to be taken in order to arrive at a destination. Therefore, potential interruptions or deviations need to be taken care of. Furthermore, most routes are too complex to be described in a single sentence; they must be subdivided into segments that can be described and understood easily. In order to handle these aspects, several requirements must be fulfilled. This does not only include keeping track of the user’s position and movements, but also being able to localize diverse parts of the route (such as the beginning and the end of it or one of its segments). If an interruption occurs, the user should be informed about the state and destination of the route she is resuming. In case of a deviation, a modified route must be computed and communicated to the user.

2.4 Providing the user with information about her surroundings

Another service that is closely tied to locational awareness and that is very fundamental to a truly mobile system, is the provision of information about the user’s current surroundings. On one hand this includes the localization of objects in the real world with respect to a known landmark or the user’s current position. This functionality is very similar to the one described in 2.2³, so that the same arguments apply here as well. On the other hand, the user may wish to have an object identified (e.g. by its name) or described more thoroughly. Although this is not a task involving spatial cognition per se, the intended object needs to be found, which in turn is a spatial task.

3 The Talking Map system

The goal of the Deep Map project [6] is to create a mobile tourist guide with a intuitive and multi-modal interface. The current system is based on a multi-agent architecture (see figure 3), consisting of three layers: the interface, the cognitive, and the knowledge layer. While the topmost layer contains agents that handle the user’s input (such as raw speech) and the system’s output, the actual replies to the user’s queries are computed on the cognitive layer. A community of agents (labeled ‘query and answer translator’ in the figure) cooperate to generate a feasible reply, which is in turn communicated to the user according to a plan generated by the presentation planner. The agents on the cognitive layer rely on the knowledge provided by agents on the knowledge layer, such as a geographical information system (GIS) or diverse databases.

The goal of the design of the natural language processing system within the Deep Map project is to create flexible and intelligent natural language processing modules, that are able to adapt to factors such as the context, the user and the subject matter at hand. The resulting natural language processing system, called *Talking Map*, enables users to gain intuitive access to underlying information technology services without prior knowledge of the services’ internal modalities. Thus, users can employ complex and heterogeneous systems and information sources through a single natural language interface and processing system. The Talking Map system consists of several agents (mainly located on the interface and cognitive layer of Deep Map), which communicate using a shared ontology of roughly two hundred classes and concepts. Currently, Talking Map has four main sub-systems that cooperate to react to utterances of the user:

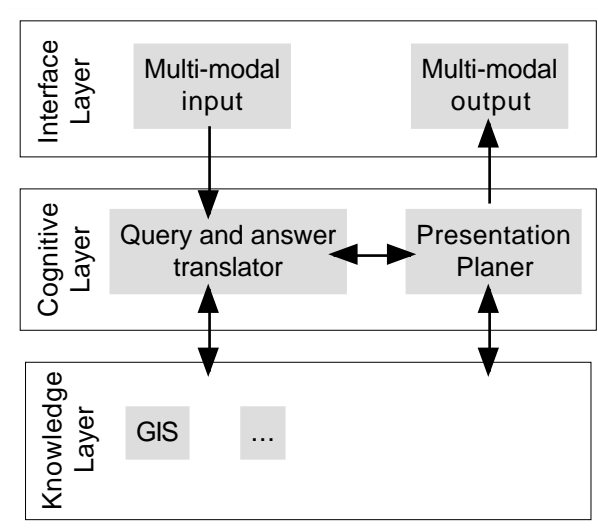


Figure 1: Architecture of Deep Map

- **The language recognition and parsing agents** recognize what the user said and transform the utterance into an internal representation, which is augmented with linguistic infor-

³Instead of the user’s current position, it is the position of an arbitrary object that is described.

mation, and a language independent type feature structure (TFS) is produced. The TFS represents the semantic content that is extractable from the recognized text.

- **The cognitive processing and dialog agents** take the output of the former, and further enrich it with extra-linguistic information. They then proceed to compute an appropriate reply to the user's input, and produce language independent verbalization plans.
- **The natural language generation and articulation agents** analyze these preverbal messages, generate linguistic surface structures from them, and supervise the synthesis of these surface structures into speech.

3.1 Understanding the user's query and replying to it

Every time the user speaks into the system's microphone, the utterance is preprocessed and acoustically analyzed by the spontaneous continuous speech recognizer, i. e. the Janus Recognition Toolkit [2]. The most likely hypothesis for the user's utterance is passed to a syntactic parser whose output is transformed into a semantically annotated (language independent) type feature structure. If the content of the TFS is a query related to spatial concepts, the structure is passed to the content planning agent in charge, i. e. SPACE⁴.

SPACE in turn first checks the content of the message for missing information such as omissions and anaphoric expressions. It then tries to infer the information based on the last object mentioned, the user's current position and view direction, and partial information (such as the type of an object in case that no name was mentioned).

Once the user's goal has been identified, SPACE tries to generate an appropriate reply to the query if it is related to space (localization, route directions, object identification, etc.). In this task, it relies on other agents such as a geographic information system or the global positioning system (GPS) to provide basic information (such as the user's location or the names and positions of landmarks) that is needed to reason about an educated response to the user's input. This information is processed concurrently by the sub-components (semi-autonomous micro agents) of SPACE. There are four main subsystems, which interact to evaluate the user's input and try to compute an appropriate response.

- The Utterance Manager analyzes the TFS generated by WHATIS, and adds the request to the appropriate agent's job list.
- The Position Manager keeps track of the user's current location.
- The Route Manager handles tasks such as route progression, interruption, and reaction to deviation from the route.
- The Core performs the basic computations such as selecting a frame of reference and choosing the best reference object. It is also responsible for determining spatial (path) relations, and for disambiguation of referential statements, such as "*what's that tower*", by recourse to pragmatic and contextual features, e.g. the analysis of position-dependent visibility and proximity factors. The methods employed are based on a decompositional approach for the computation of spatial relations (see [Gapp 94], [Kray & Blocher 99]).

Once the computation is concluded, SPACE encodes the results in a preverbal message (PVM), which consists of language independent information chunks that describe the content plan for the next utterance/communicative act of the system. Currently, there are four types of PVMs that are generated by SPACE depending on the user's original request: spatial instructions, spatial localizations, object identifications and object descriptions.

3.2 Route instruction

This type of PVMs is used for route instructions, i. e. giving the user dynamic directions⁵ to desired locations. It consists of an ID for the current route (`rid`) and one for the actual segment that is described (`preid`).⁶ In addition, the overall goal (`global-goal`) and the route state is included

⁴SPACE is an acronym standing for *Spatial Cognition Engine*

⁵*Dynamic directions* are not start-to-finish directions, requiring the user to memorize the entire route, but are split into adequate segments, which are communicated to the user incrementally. That means after starting with an instruction for the first segment, the following ones are uttered once the user has reached the end of the previous one.

⁶This can also be used to visualize the route and the progress the user made so far.

(*route-state*) which can be: *start*, *continue*, *resume*, *interrupt*, *cancel*, *done*, *undefined* or *unknown*. Furthermore, the actual segment is described in terms of spatial relations: there is a localization for its start and end point (*start & end*), a localization of its trajectory (*2p-relation*), and a path relation describing its course (*np-relation*). The quality of each relation is rated depending on its applicability to the current situation. Finally, the PVM contains a turning angle that encodes a turning action the user must perform at the beginning of the segment (*reorient*), and metric information about the length of the segment (*metric*).

In case the natural language generation module, called DIRECT⁷, receives a PVM encoding a route instruction, as shown in figure 2, a spatial instruction will be generated, by means of a template-based grammar. A full-fledged spatial instruction can consist of four phrases at most:

- a confirmation phrase, e. g. *in order to get to the station ...*
- a reorientation phrase, e. g. *... you have to turn to the right ...*
- an instruction phrase, e. g. *... and walk towards the bridge...*
- a termination phrase, e. g. *until you see the broken tower on your right*

Depending on the route state DIRECT will generate more or less elaborate confirmation phrases - in case of *start* or *resume* - or none at all - in case of *cancel*. Reorientation phrases will be generated unless the user is already facing in the right direction. Instruction phrases will always be generated unless the user has reached the desired location, i. e. *route-state* is *done*. Termination phrases are generated only if the user model indicates that such specificity seems necessary. The choice of reference objects

```
(preverbal-message
:rid 17
:route-state "resume"
:preid 5
:pvm-type "path"
:global-goal (:oid 5 :name "station")
:start (next-to .78 (:oid 78 :name "ticket office"))
:end (in 1.0 (:oid 13 :name "broken tower"))
:reorient 85.0
:2p-relation (:left .45 (:oid 9 :name "castle moat"))
:np-relation (:approach .78 (:oid 4 :name "bridge")))
:metric 54.25
```

Figure 2: A sample PVM

(RO) depends on the the specific degrees of applicability (DA) for the given spatial relations. If the path relation (*np-relation*) provides a satisfactory DA, it is preferred over *2p-relations* due to its vast expressive capabilities. Otherwise, less expressive relations, e. g. *2p-relations*, are chosen that feature a higher DA. The linguistic means for referring to the chosen RO depends on the information from the user model, e. g., whether the object has been introduced before and is known by the user, or not.

3.3 Localization

Whenever the user asks for the location of an object, e. g., *where is the powder tower*, a PVM of this type is generated. It consists of the localized object (LO), a spatial relation and an appropriate reference object describing the location of the LO in question. Furthermore the metric distance and an angle indicating the direction are computed and included in this type of PVM as well.

The resulting localization phrase will employ as much as possible of the information encoded in the PVM, i. e., verbalizing both the metric distance as well as the direction in which the object lies, e. g., *the powder tower is roughly 140 meters to your right* or *about 20 meters left of the castle*. At the moment all spatial relations are realized with a listener-centric frame of reference.⁸

3.4 Object identification

Preverbal messages of this type are created in case the user asked a question such as *what's that building ahead of me*. The main ingredient of these PVMs is a reference to the corresponding object in the database.

The resulting phrase will be either a simple phrase of the type *you are looking at the ...* or if short and fitting information on the object can be extracted from the database it is verbalized as such, supplying the tourist with additional concise information about the object.

⁷DIRECT is an acronym standing for *Dynamic Instruction and Referencing Communication Tools*.

⁸A *listener-centric* frame of reference can either employ the listener as origo and reference object or use the listener as origo but chose a third object as reference object, see [8].

3.5 Object description

This type is generated in case the user asked for more information on a certain object, e. g., *tell me more about that tower over there*. Although this PVM is very similar to the one for object identification as it also consists mainly of a reference to the corresponding information in the database, the intention behind it goes beyond pure identification. Consequently, the user receives a richer reply than in the identification case: Not only will the texts be longer, but they will be, in most cases, displayed graphically along with other multi-media information, rather than synthesized.⁹

4 Conclusion and Future Work

Based on our experiences with a mobile tourist information system, we have identified some basic functional requirements, which we think apply to most mobile systems. We have argued that a natural language interface and a spatial cognition component are necessary to meet those requirements, and to provide the user with a intuitive interface. We have presented the Talking Map system, that was built to offer all of the services described in this paper, and we showed how these services are handled within the system. In the future, we plan to extend the preverbal messages in several ways, e. g. by incorporating regions and by unifying the input and output representations. We will also add new agents to the agent community (such as dialog and context handling agents) in order to further improve the reasoning power of the system.

Acknowledgments

Deep Map is a research project funded by the Klaus Tschira Foundation. The research presented here is conducted in collaboration with DFKI (Saarbrücken) and the University of Heidelberg.

References

- [1] Feiner et. al: A Touring Machine: Prototyping 3D Mobile Augmented Reality Systems for Exploring the Urban Environment. Proc. of ISWC '97 (International Symposium on Wearable Computing). Cambridge, MA (1997) 74-81
- [2] Finke et. al: The Karlsruhe-Verbmobil Speech Recognition Engine. Proceedings of ICASSP 1997. Munich, Germany (1997)
- [3] Gapp: Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space. Proc. of AAAI '94. Seattle, WA (1994).
- [4] Kray C. and Blocher A.: Modeling the Basic Meanings of Path Relations. Proceedings of the 16th IJCAI, Morgan Kaufmann. San Francisco, CA (1999) 384-389
- [5] Lakoff, G.: Hedges: A Study of Meaning Criteria and the Logic of Fuzzy Concepts. Journal of Philosophical Logic **2** (1973) 458-508
- [6] Malaka, R. and Zipf, A.: Deep Map: A mobile Tourist System GeoBit 28, (1999)
- [7] McKee, L. and W. Kuhn: The objectives of the Open GIS Consortium. In D. Fritsch et al. (Eds), Proceedings of the 46th Photogrammetric Week, Wichmann Verlag. Stuttgart, Germany (1999)
- [8] Porzel et. al: A cognitively motivated NLG system. Spatial Language. Olivier P. (ed). Kluwer Academic Publishers (in print)
- [9] The RAGS project.: Towards a reference architecture for natural language generation systems. ITRI Technical Report **ITRI-99-14** University of Brighton (1999)

⁹In all cases, the system will not only generate and synthesize texts: in the case of instructions, identifications and descriptions, the system will also generate graphical output in various forms, e. g. pictures, slide shows, maps, animations, etc.