Talking with HIGGINS: Research challenges in a spoken dialogue system

Gabriel Skantze, Jens Edlund, and Rolf Carlson

Department for Speech Music and Hearing, KTH
Lindstedtsv. 24, 100 44 Stockholm, Sweden
{gabriel, edlund, rolf}@speech.kth.se

Abstract. This paper presents the current status of the research in the Higgins project and provides background for a demonstration of the spoken dialogue system implemented within the project. The project represents the latest development in the ongoing dialogue systems research at KTH. The practical goal of the project is to build collaborative conversational dialogue systems in which research issues such as error handling techniques can be tested empirically.

1 Introduction

This paper presents the current status of the research in the HIGGINS project and the spoken dialogue system implemented within the project. The project represents the latest development in the ongoing dialogue systems research at KTH (for an overview, see [1]). The practical goal of the project is to build collaborative conversational dialogue systems in which research issues such as error handling techniques can be tested empirically.

The initial HIGGINS domain – pedestrian city navigation and guiding – is similar to the now classic MapTask [2] domain and to a number of guide systems such as REAL [3]. In this domain, a user gives the system a destination and the system guides the user verbally. For simulation purposes, a 3D model of a city is used (see fig. 1). The system does not have access to the users' positions, but must rely on their descriptions of their surroundings. Since the user is moving, the system must continually update its

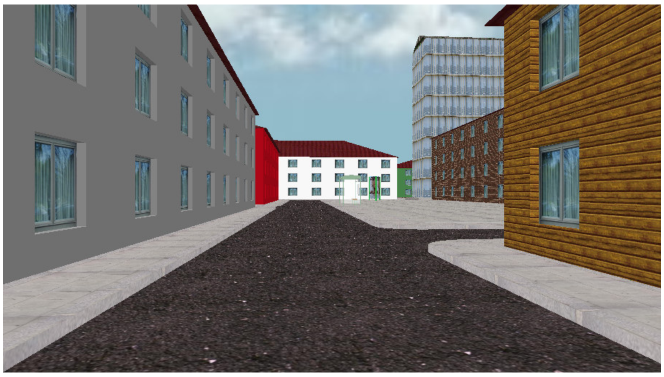

Fig. 1. The 3D simulation that is used for user tests.

model of the user's position and provide new, possibly amended instructions until the destination is reached. The surroundings the user and system talk about contain complex landmarks and relations that are challenging to interpret and represent semantically. Compared to simpler domains, the users' utterances (e.g. natural language descriptions of the surrounding environment) tend to be longer, more disfluent and filled with pauses. The domain is implemented in Swedish.

The second major domain to be implemented in HIGGINS was the KTH Connector [4] – a telephony based personal secretary whose task it is to mediate communication between callers and (potentially occupied) callees[1]. Again, the users and the system can reason about complex concepts that are challenging to model – notably relations in time. The KTH Connector is implemented in English.

Finally, several toy domains have been implemented to highlight and test particular HIGGINS features, notably a voice controlled chess board and the language training game *Is it blue?* [5].

2 The HIGGINS spoken dialogue system

The HIGGINS spoken dialogue system has a distributed architecture with modules communicating over sockets. Each module has a clearly defined input and output, and can be implemented in any language, running on any platform. All messages and resources are encoded in XML.

The complex semantics used in the HIGGINS domains call for deep semantic structures, and a main focus of the project to date has been developing and testing robust and "error aware" modules for interpretation: the semantic interpreter PICKERING [6] and the discourse modeller GALATEA [7], both implemented in Oz[2]. PICKERING is designed to work with continuous incremental input from a probabilistic speech recogniser. It allows insertions and non-agreement inside phrases, and combines partial results to return a limited list of semantically distinct solutions. The semantic structures that PICKERING produces are represented as rooted unordered trees of semantic concepts. Nodes in the tree may represent for example attribute-value pairs, objects, relations and properties.

GALATEA provides the next step in the interpretation. Whereas PICKERING builds a model of the semantics of an utterance, but does not consider context outside the utterance, GALATEA takes the communicative acts that PICKERING finds in an utterance and does a context aware interpretation of them, resulting in a dynamic model of the discourse.

In HIGGINS, the traditional aspects of dialogue management are not implemented in a single module. Instead, the processing is divided so that GALATEA models the discourse, and the discourse model is then sent to an action manager, which consults the discourse model and a domain database to make decisions and generate communicative system acts. These acts are sent back to GALATEA, as well as to a generator. Thus, GALATEA models communicative acts both from the user and the system.

[1] The KTH CONNECTOR is part of the EU-funded project CHIL – a project investigating automatic tracking and support of human interactions.

[2] http://www.mozart-oz.com

3 Research challenges

The Higgins domains are chosen to give rise to a number of research challenges. Here, we will discuss a few of our current foci: concept-level error handling, interaction control, and how well the Higgins techniques adapt to other domains.

Concept-level error handling

A challenging research issue is how to handle errors (due to imperfect speech recognition) for individual concepts in the deep semantic structures, using for example elliptical clarification requests, in order to make the dialogue natural and efficient. Pickering can assign confidence scores to individual concepts in the semantic result, based on the word confidence scores from the speech recogniser. Galatea may then store these scores in the discourse model as a measure of *concept grounding status*. Dialogue (1) illustrates different ways of using this information:

(1) **U** I have a red building on my left
Sa Red?
Sb Red, hm. How many storeys does it have?
Sc How many storeys does the red building have?

If the grounding status for a concept is low, the system may use an elliptical clarification request (as in Sa), display understanding (as in Sb) or modify the way it refers to objects (as in Sc), in order to "boost" this grounding status. Since Galatea handles anaphora and ellipsis resolution, it can accurately update the grounding status for the concepts that are involved.

As elliptical clarification requests lack syntax, the prosodic realisation becomes more important. An example is "red" in Sa and Sb, which should be pronounced differently. We are currently performing studies of how prosodic features affect the interpretation of synthesised fragments in such situations [8, 9].

Another error handling method is *late error detection,* i.e. to find possible errors in the discourse at later stages in the dialogue. If, for example, the system finds that there is no possible location for the user given its current beliefs, it may try to remove concepts with low grounding status.

As seen above, the discourse model allows for different error handling strategies. Currently, we are collecting data of users talking to the system. Based on analysis of this data, we investigate methods for choosing strategy.

Interaction control

Another interesting challenge is how the system should best handle interaction control. Dialogue (2) shows a typical problem in spoken dialogue systems, where the voice activity detection in the automatic speech recogniser has erroneously told the system that the speaker is done talking.

(2) **U** to my left I see a /SIL/ yellow building
S what do you see to your left

Silence detection alone, as used by most systems today, is not sufficient to deal with this kind of situation. Methods involving both syntactic and semantic completeness (e.g. [6]) as well as prosody [10] have been shown to improve the situation, and we are currently investigating their use in HIGGINS.

Generalisability

An important question is to what extent the techniques developed within the HIGGINS project apply to other domains. One of the reasons for dividing dialogue management into a discourse modeller and an action manager is that it allows the discourse modeller to be fairly generic (GALATEA is simply configured using XML), while the action manager is highly domain dependent. The action planner may have to be reimplemented for each new domain, but this is facilitated by the facts that it can be implemented in any programming language, and that much of the work typically done by a dialogue manager (e.g. ellipsis and anaphora resolution) is already dealt with by GALATEA.

As mentioned above, there are currently several domains implemented in HIGGINS. More domains will be tested in an attempt to find out how general the methods are, and how easy (or difficult) they are to use.

Acknowledgements

This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations and was supported by the EU project CHIL (IP506909).

References

1. Gustafson, J. (2002): Developing Multimodal Spoken Dialogue Systems. Empirical Studies of Spoken Human-Computer Interaction. TRITA-TMH 2002:8, ISSN 1104-5787.
2. Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R (1991): The HCRC Map Task corpus. Language and Speech 34(4) 351-366
3. Baus, J., Kray, C., Krüger, A., & Wahlster, V (1991): A resource-adaptive mobile navigation system. In proc. of the International Workshop on Information Presentation and Natural Multimodal Dialog.
4. Edlund, G. & Hjalmarsson, A. (2004): Applications of Distributed Dialogue Systems: the KTH Connector. In proc. of the ISCA Tutorial and Research Workshop on Applied Spoken Language Interaction in Distributed Environments (ASIDE 2005) Aalborg, Denmark
5. Hjalmarsson, A.. & Wik, P. (2005): Is it blue?, Term paper, Course in NLP, GSLT, Sweden
6. Skantze, G. & Edlund, J. (2004). Robust interpretation in the Higgins spoken dialogue system. In proc. of ITRW on Robustness Issues in Conversational Interaction 2004.
7. Skantze, G: Galatea – a discourse modeller supporting concept-level error handling in spoken dialogue systems. In proc. of SigDial 2005
8. Edlund, J., House, D., & Skantze, G (2005): The effects of prosodic features on the interpretation of clarification ellipses. In proc. of Interspeech 2005, Lisbon, Portugal
9. Wallers, Å., Edlund, J., & Skantze (under review): Small sounds of great importance. Submitted to Perception and Interactive Technologies (PIT06), Kloster Irsee, Germany
10. Edlund, J & Heldner, M (2005). Exploring Prosody in Interaction Control. Phonetica, 62(2-4).