Metalogue: a multiperspective multimodal dialogue system with metacognitive abilities for highly adaptive and flexible dialogue management

Jan Alexandersson and Andrey Girenko
German Research Centre for Artificial Intelligence
GmbH (DFKI), Saarbrücken, Germany
jan.alexandersson@dfki.de, andrey.girenko@dfki.de

Volha Petukhova and Dietrich Klakow
Department of Spoken Language Systems
Universität des Saarlandes, Saarbrücken, Germany
o.petukhova@lsv.uni-saarland.de,
dietrich.klakow@lsv.uni-saarland.de

Niels Taatgen
Department of Artificial Intelligence
University of Groningen, Netherlands
niels@ai.rug.nl

Nick Campbell
School of Linguistic, Speech and Communication Sciences
Trinity College Dublin, Ireland
nick@tcd.ie

Alexander Stricker
Charamel GmbH, Cologne, Germany
stricker@charamel.com

Dimitris Spiliotopoulos
Interaction Design Lab
University of Peloponnese, Greece
dspiliotopoulos@gmail.com

Dimitris Koryzis
Hellenic Parliament, Athens, Greece
dkoryzis@parliament.gr

Marcus Specht
Research Centre for Learning, Teaching and Technology
Open University of the Netherlands (OUNL)
Marcus.Specht@ou.nl

Maria Aretoulaki
DialogConnection Ltd, Manchester, UK
maria@dialogconnection.com

Michael Gardner
School of Computer Science and Electronic Engineering
University of Essex, Colchester, UK
mgardner@essex.ac.uk

***Abstract*—This poster paper presents a high-level description of the Metalogue project that is developing a multi-modal dialogue system that is able to implement interactive behaviors that seem natural to users and is flexible enough to exploit the full potential of multimodal interaction. We provide an outline of the initial work undertaken to define a an open architecture for the integrated Metalogue system. This system includes components that are necessary for the implementation of the processing stages for a variety of application domains: initialization, training, information gathering, orchestration, multimodality, dialogue management, speech recognition, speech synthesis and user modelling.**

Keywords-multi-modal dialogue, meta-cognition, systems architecture, natural language, human-computer interface

I. Introduction

In today's information society the use of multi-modal natural language based dialogue can offer an attractive human-machine interface within a range of different environments, from smart houses through to work spaces. Such interfaces offer a mode of interaction that has certain similarities with natural human communication by using a number of different input and output modalities that people normally employ in communication (eg. speech, gesture, facial expressions, pointing devices, etc). Some of these interfaces can also incorporate multi-modality into virtual and augmented environments through the use of embodied conversational agents.

The Metalogue project is a new project that is being funded by the European Union Framework 7 programme. The goal of *Metalogue* is to produce a multimodal dialogue system that is able to implement an interactive behaviour that seems natural to users and is flexible enough to exploit the full potential of multimodal interaction. It will be achieved by understanding, controlling and manipulating the system's own and users' cognitive processes. A new dialogue manager will be developed that will incorporate a cognitive model based on metacognitive skills that will enable planning and deployment of appropriate dialogue strategies. The system will be able to monitor both its own and users' interactive performance, reason about the dialogue progress, guess the users' knowledge and intentions, and thereby adapt and regulate the dialogue behaviour over time.

The metacognitive capabilities of the *Metalogue* system will be based on a state-of-the-art approach incorporating multitasking and transfer of knowledge among skills. The models will be implemented in ACT-R, [5] providing a general framework of metacognitive skills.

Metalogue research focuses on educational and coaching situations where negotiation skills play a key role in the decision-making processes. Components and algorithms will be developed, tested and integrated into a

978-1-4799-2947-4/14 $31.00 © 2014 IEEE
DOI 10.1109/IE.2014.67

prototype platform, which will provide learners with a rich and interactive environment that will help to develop metacognitive skills, support motivation, and stimulate creativity and responsibility in the decision making and argumentation process. The *Metalogue* system will produce virtual agents capable of engaging in natural interaction through combinations of English, German and Greek, using gestures, mimicry and body language.

It will be deployed in two use-case scenarios: in social educational contexts for training young entrepreneurs and active citizens (Youth Parliament), and in a business education context for training call-centre employees to successfully handle their customers.

This poster presentation will present the initial work to develop an innovative computational architecture for the Metalogue systems, which is described in the next section. In addition we will also provide a number of detailed scenarios that will illustrate the intended applications for the final system. We hope to use this as an opportunity to gather feedback on the approach being taken which will be used to refine the work of the project as we enter the developmental and research phases.

II. Metalogue Architecture

The envisaged overall architecture can be viewed as three layers. The low level core layer comprises a system architecture which supports several input and output modalities, such as spoken natural language, facial expressions, body posture and biosensor data, and is designed to be modality-agnostic; a high level layer providing a tool kit for developers (API) to create applications; and finally a validation layer in which the API to is employed to create an application that allows user partners to validate the benefits of the project (see figure 1 below).

The proposed architecture will follow the principles of multi-modal dialogue architectures such as described in [1] and will include the following components:

- Sensor specific input
- Interpretation (Modality specific analyses)
- Fusion
- Discourse and dialogue manager
- Metacognition
- Fission
- Generation
- Rendering (Text to speech, visual)

In addition, components may be associated with knowledge bases such as common-sense ontologies and domain-specific structured databases (see [2] and [3]). Communication between the components will use common standards, but the details of the transport mechanisms and protocols will have yet to be defined.

The system architecture is manifested as the *low-level* layer of the overall METALOGUE architecture. This will include several modalities, such as spoken natural language, facial expressions, body posture and biosensor data. Where appropriate, these modalities are designed to be symmetric: a modality available for user input will have a counterpart for output. For example, the ability to use facial recognition will be complemented by the generation of facial expressions via virtual characters. This approach allows for a natural interaction and is especially useful in multi-party interactions. In such settings, some modalities may be exclusive to two communication partners (for example, speech), but other participants can still interact in parallel, for example, using gestures.

The system architecture is designed to be modality-agnostic as far as practical. Therefore, its aim is to provide APIs that allow developers to connect new devices and interpretation functionality for their sensor data. The only requirement from the architecture is that the interpretation of the sensor data provides results in a uniform semantically enriched format that can be mapped to DiaML [4] communicative acts.

The cognitive, learning and interaction models are closely integrated and direct the operation of a dialogue manager component. The dialogue manager acts over a shared information state that incorporates and manipulates information from all models. The dialogue manager is plan-based and works on the level of sub-goals. These take the form of tasks that are generated by the learning model based on the current situational context and the progress of the learning objectives. From this, the dialogue manager devises intentions and character-specific goals for the virtual characters and dynamically plans how the intentions can be achieved by taking into account the available interaction patterns and conditions or requirements arising from the current situational context.

The result of the planning phase is a set of actions for each virtual character. These actions are then rendered in a virtual environment using a combination of generated speech output and gestures, facial expressions and body postures that can be performed by virtual characters in that environment. The data, including user inputs and rendered outputs, is available to all components in a semantically enriched representation. This is necessary to enable the processing of some interaction phenomena. For example, a virtual character's behaviour must be present and accessible to the interpretation components if they are to be able to resolve a subsequent user input that references that behaviour, whether it is a linguistic reference (for example, an elliptic expression), a deictic reference or a reference of some other kind.

The system architecture is designed to allow conversational interactions in real time and use processing shortcuts where time-critical reactions are desired. For example, if a spoken input is detected, immediate attention feedback via gaze is generated by the virtual characters, bypassing slower full semantic processing of the input. This feature is very important in creating and preserving the perception of immersion and responsiveness for the user. In a further step, the system will feature incremental processing, which is both a more natural and accurate way of achieving such behaviour. Rather than waiting for a complete utterance from the user (such as a sentence), input (such as a word or gesture) is processed in real time, updating system understanding as new information is expressed.

The *high-level* layer will include an easy-to-use JavaScript API that allows developers to create new applications based on the system architecture. The scripting mechanism and the way to insert data will be simplified to allow its use by non-experts. The JavaScript API will be supplemented by a graphical toolset that allows users to create new dialogues and expand existing ones, although such a toolset cannot sensibly anticipate the full range of possibilities that will be available on all different target platforms and configurations. A comprehensive graphical workbench would be a very large undertaking beyond the scope of the project. Therefore, the workflow for creating new use-cases will consist of creating a high-level application skeleton with the toolset and subsequently refining it using the API functions.

This layer of the architecture acts as an abstraction feature for the low level core and hence depends heavily on the design and implementation decisions for the core. As a consequence, its development and further specifications will start once the low level is mature enough for this API to be relevant.

The *validation* layer will make use of the architecture to create an application that allows user partners to validate its benefits. The project is designed to go through several incremental iterations that will be used to gradually improve and refine the models for learning, interaction and cognition, the underlying knowledge representation and the general usability of the system. Thus, it is crucial that demonstrator systems are deployed early and accessible to public users. In this way, data from user interactions can be collected and evaluated to gain insight with regard to any shortcomings and provide suggestions for improvements for the subsequent development cycles. To this end, the validation layer will provide a means to record and collect interaction data together with meta-data such as error rates, task completion and measures of user satisfaction.

III. Summary

This short poster paper has provided a high-level description of the Metalogue project and an overview of the initial systems architecture that will be used as the basis for the development of the final system. We are currently developing a number of scenarios that will illustrate the possible uses for the system and intend to present these along with the systems architecture in order to solicit feedback from the research community. It is envisaged that the interactive behaviors and naturalistic multi-modal dialogue supported by the Metalogue system will be an important component in the development of future intelligent environment applications.

Acknowledgment

The Metalogue project full title is ‘Multi- perspective multimodal dialogue system with meta-cognitive abilities’ and is funded by the European Union Framework 7 ICT programme, under grant agreement 611073. The work in this paper represents activities carried out by all members of the Metalogue consortium. Further information about the project can be found at http://www.metalogue.eu.

References

[1] Herzog, G., & Reithinger, N. (2006). The SmartKom Architecture: A Framework for Multimodal Dialogue Systems. *SmartKom: Foundations of Dialogue Systems*, 55–70. doi:10.1007/3-540-36678-4.

[2] Matuszek, C., & Cabral, J. (2006). An Introduction to the Syntax and Content of Cyc. *AAAI Spring Symposium:* …. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.1357&rep=rep1&type=pdf.

[3] Fellbaum, C. (1999). *WordNet*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal1285/full

[4] Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., … Traum, D. R. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *LREC* (pp. 430–437).

[5] Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press. ISBN 0-19-532425-0.

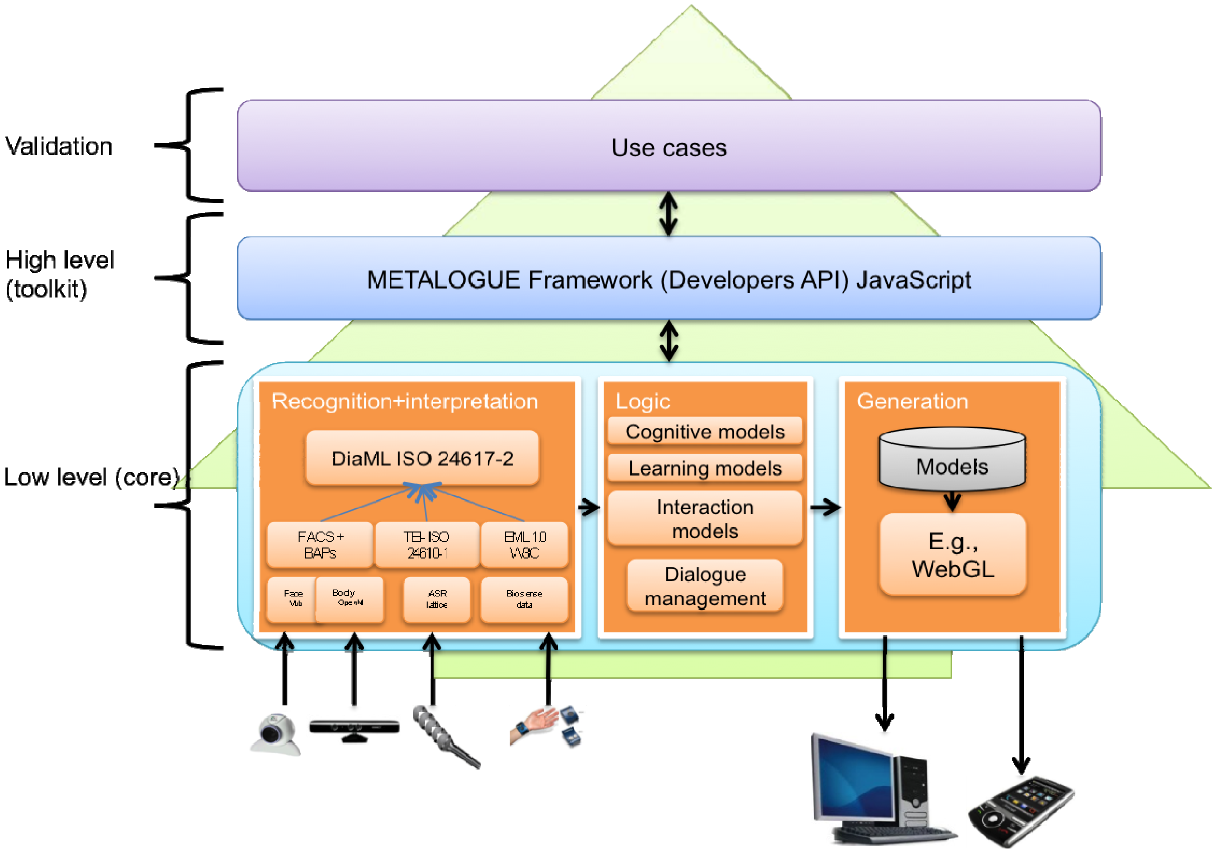

Figure 1. Metalogue System Architecture