SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies
13 September 2016, San Francisco, USA

`flexdiam` – flexible dialogue management for problem-aware, incremental spoken interaction for all user groups (Demo paper)

Ramin Yaghoubzadeh, Stefan Kopp

Social Cognitive Systems Group, CITEC, Bielefeld University, Germany
ryaghoubzadeh@uni-bielefeld.de, skopp@uni-bielefeld.de

Abstract

The dialogue management framework `flexdiam` was designed to afford people across a wide spectrum of cognitive capabilities access to a spoken-dialogue controlled assistive system, aiming for a conversational speech style combined with incremental feedback and information update. The architecture is able to incorporate uncertainty and natural repair mechanisms in order to fix problems quickly in an interactive process – with flexibility with respect to individual users' capabilities. It was designed and evaluated in a user-centered approach in cooperation with a large health care provider. We present the architecture and showcase the resulting autonomous prototype for schedule management and accessible communication.

Index Terms: human-computer interaction, conversational spoken dialogue, user models, incremental processing, flexible grounding, assistive systems

1. Introduction and outline

Making spoken human-machine interaction both easy and effortless, and also robust in presence of contradictory pieces of information, is one of the central challenges in providing universal accessibility over this modality. Two of the user groups that would benefit most from this are, on the one hand, older adults, who may be reluctant or lack the capacities to interact with technology using more widely supported modalities, but also people with cognitive impairments, for whom accessing even well-designed classical interfaces can be a challenging task. Spoken interaction is overall reported as the preferred modality by older adults with little technological experience [1]. While speech recognition for these user groups can present specific difficulties [2], the available technology for word recognition has improved in the last few years to a degree that it is now feasible. Given robust – and engaging – spoken interaction, these user groups could benefit from easily accessible and understandable interfaces to technological solutions that help them to maintain an autonomous lifestyle.

In our cooperation with the large health and social care provider v. Bodelschwinghsche Stiftungen Bethel, we have explored the paradigm of a spoken-language controlled virtual assistant for schedule management, to aid in maintaining a client's day structure. Initially, in Wizard-of-Oz explorations, we established that both user groups are, in general, capable of conducting such interactions in a brief and effortless conversational style. We also found that the approach was subjectively judged as pleasant, effective and appropriate.

Building on our existing architecture for incremental dialogue processing, we created a dialogue management framework that aims to address several issues critical to making autonomous interactions with these user groups work robustly, the central requirements being:

- being aware, and addressing interactively, ambiguities in user input,
- being able to react rapidly and give feedback before problems can cascade,
- presenting and negotiating information in a way that supports individual capabilities, and
- allowing the user to feel in charge and being served well.

The resulting architecture was used to build a dialogue system that is able to provide basic schedule management and access to video communication with a conversational, incremental spoken interface represented by an embodied assistant, which we are presenting here. A subset of this functionality, namely completing a weekly schedule if events, was evaluated with older adults and people with cognitive impairments, leading to comparable performance and subjective ratings as the earlier WOz system.

2. Architecture overview

We present the architecture in an abridged account here, please refer to our previous work [3] for more details on the internal mechanics. `flexdiam` builds on our general architecture for incremental processing, IPAACA [4]. This this architecture, based on an abstract model by Schlangen et al. [5], information is represented as so-called 'Incremental Units' (IUs), which are globally exchanged information packages that can form functional networks. It is designed to be used to represent data in both the input (and interpretation) channels and processing, and also in output planning and realization (cf. Fig. 1, left).

The temporal structure of dialogue is represented in the `TimeBoard`, which stores all past, ongoing, and projected future events in thematically grouped tiers (Fig. 2). It serves as the interface between input processing, dialogue management proper, and behavior planning and realization. Events are most often either a single IU or a specific sequence of IUs. A set of interval relations on sets of tiers is used to determine higher-level events.

Data other than events with temporal extent, i.e. knowledge and propositional information, are represented via a structure termed `VariableContext` (Fig. 1, right), a blackboard satisfying two requirements: firstly, all information may reside there in the form of distributions. Moreover, all changes are stored as time-stamped deltas, enabling both rollbacks and for analysis between two points in time. Task and discourse states are represented a forest of structures called `Issues`, terminology adapted from Larsson [6], that represent (attributed) common current topics or current questions that have to be resolved cooperatively. In `flexdiam`, they are independent agents that

10.21437/SLPAT.2016-15

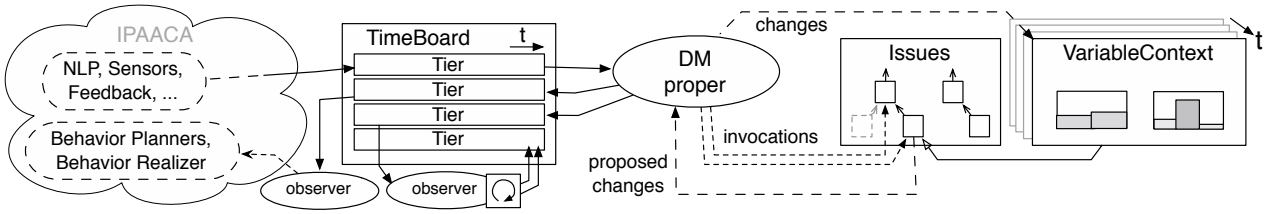

Figure 1: *Architecture overview. The cloud on the left represents external input/output modules that are not part of flexdiam proper, but connected via the common middleware IPAACA. Data structures and processing are described in the text.*

encapsulate the structure of the task addressed so far, localized planning, as well as situated interpretative capability and situated capability for abstract actions (multimodal dialogue contributions and side effects). The *dialogue manager proper* relays information hierarchically though the Issue forest (see Fig. 3 for an example of this in the interpretation process).

In line with the general notion of temporal variability and uncertainty, all operations that do not have immediate effect are treated as asynchronously performed operations that can fail.

3. Input and output

As mentioned above, all input and output components are connected to flexdiam using the IPAACA middleware.

Speech input can be delivered by several components, alternatively or concurrently (there are bridges for Windows ASR, Dragon NaturallySpeaking Client SDK and an experimental one for Google's ASR). A parser component is used to preprocess all ASR hypotheses, identifying the points of deviation in hypotheses, performing an early classification of portions of an utterance using pattern matching, and offering an interface for triggering external NLU accessories, such as POS taggers. Other input modalities accessible over IPAACA include two types of eye tracker, touch screens, keyboard and mouse input.

Output is realized by emitting request IUs that realizer components can listen for and handle. The virtual agent is controlled by the ASAPrealizer [7], which accepts action descriptions in the Behavior Markup Language. Speech generation is realized using a CereVoice [8] TTS component, which is driven by ASAPrealizer. There is a separate controller for GUI elements that can either be addressed directly or in a speech-synchronized manner by ASAPrealizer. Language output is not generated directly in flexdiam, but relayed to an associated dedicated NLG component that can offer multiple alternative realizations for an abstract request (though currently, flexdiam always chooses the first one to appear).

Fig. 4 depicts the typical interface setup, in an interaction scene between an older subject and the virtual agent "Billie". Subjects interacted using the table microphone and touchscreen (red 'panic button' in the corner).

4. Experiments

A basic dialogue system constructed with flexdiam has been subjected to small-scale evaluations with both older adults (n=6) [3] and people with cognitive impairments (n=5). The task for participants was to enter a freely chosen set of appointments into their fictional calendar, the same domain as an earlier Wizard-of-Oz experiment [9], in which we showed that people with cognitive impairments in particular benefit from a much more explicit information grounding strategy compared to controls when their ability to detect system errors is observed. We also found inter-group differences in preferred verbalizations (e.g. more frequent first-person requests in older adults vs. more frequent neutral dictation in people with cognitive impairments) [10].

Figure 4: *flexdiam driving a virtual agent, "Billie", in an autonomous interaction study with an older adult (anonymized).*

For the interactions with the autonomous prototype, we provided some ideas for events on a paper sheet with textual and iconic representations. Subjects were instructed to stick to the task and be to the point, but not primed as to how to phrase their requests or replies. In general, participants were able to enter appointments successfully. Some leeway was given by participants if the agent paraphrased only (a relevant) part of their event descriptions – a simple heuristic approach was used to extract candidate topics from the free-form utterances.

The system in that state was configured to always yield the floor and let the user talk at their leisure. One subject from each group used very verbose interaction styles and attempted to provide a lot of tangential information, despite a clarifying instructive intervention that could be inserted after an initial free practice phase. The current focus of development is hence on subtle and acceptable approaches to pre-emptive floor management.

Subjective ratings of the autonomous system in terms of effectiveness and usability did not differ significantly from the earlier WOz experiment that targeted the same interface and task domain [3].

5. Demo system

The demo system showcases flexdiam in a schedule planning scenario controlled by spoken language, enabling the user to go through their (fictional) week, modifying events, decid-

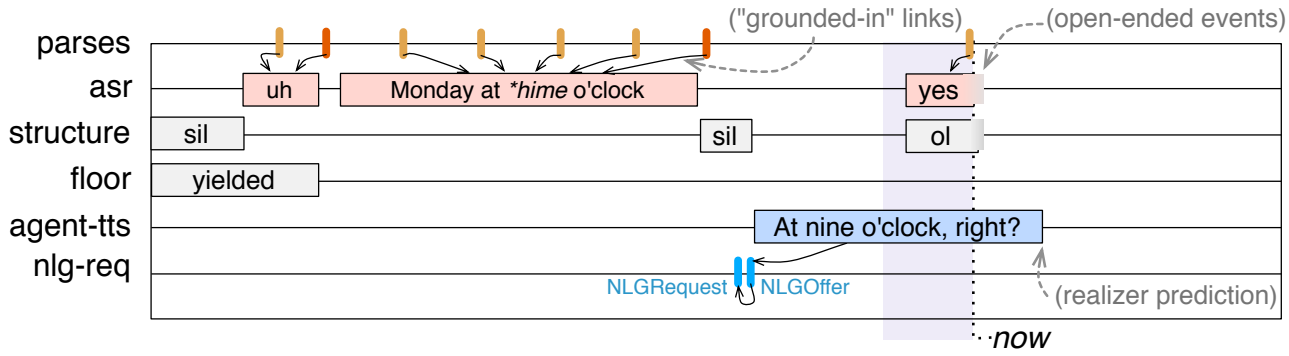

Figure 2: *Temporal representation of dialogue events on the TimeBoard. In a situation where the user (red) wanted to enter a new appointment, they produced an utterance that was mispronounced, leading to ambiguities. The DM posted a clarification question (blue), its predicted end time is shown extending beyond the time marked 'now'. In the current situation, the reply by the user has already started (producing an overlap). In the default configuration, the system would yield the floor to the user immediately if an overlap over a threshold length is encountered.*

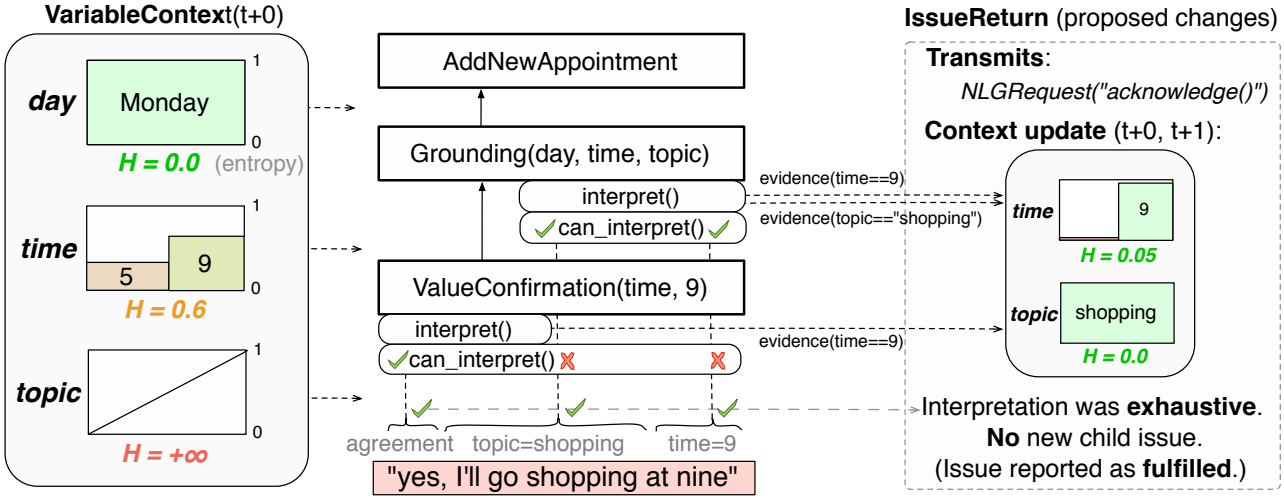

Figure 3: *Current context (left) and relevant subset of the Issue forest (center) for a situation just past the one in Fig. 2: the user has completed their utterance. Interpretation is performed locally first (in the ValueConfirmation issue), then deferred to its parent (Grounding). Both contribute to the evidence that leads to a proposed Context update (right). Since the requirements for a confirmation question were met, the bottommost Issue reports itself as fulfilled. The mechanism during incremental interpretation is identical.*

ing about events offerd by third parties, as well as an interface to an encrypted video telephony application that can be triggered from inside the dialogue situation. Different modes of information grounding can be selected (e.g. concise summaries vs. fine-grained confirmation requests), these correspond to user models for different cognitive abilities; and a live view into the information update process is possible. The demo setup includes a computer, a desk microphone, and an eye tracker. A recording of a previous version is accessible online at purl.org/net/ramin/slpat2016/.

6. Outlook

In the ongoing project with our health care partner, larger-scale evaluations of various aspects of the system are underway, later extending to prolonged experimental deployment in private home environments of interested participants. Current and future steps of development include proactive floor management and improved reference resolution mechanisms in the framework. We are planning to release documentation and a source code package for flexdiam in the future under a permissive license.

7. Acknowledgements

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) in the project 'KOMPASS' (FKZ 16SV7271K) and by the Deutsche Forschungsgemeinschaft (DFG) in the Cluster of Excellence 'Cognitive Interaction Technology' (CITEC).

```
a1  AGNT  Do you have another appointment?
    SUBJ                                  Yes. Then, I have yet another appointment ... on Friday

a2  AGNT  So, on Friday, right?        OK. At what time does it start?
    SUBJ                       Right.                                  Then I'll pick 3 PM again,

a3  AGNT                   So, at 3 PM, right?                 So, at 3 [interrupt]   Good.
    SUBJ  have ice cream.                      [hoarsely] Yah             Yes.

a4  AGNT  So, at that time, there is "Have ice cream", right?         Okay. Then I'll enter it as follows...
    SUBJ                                                  Right.


b1  AGNT Do you have another appointment?        Then tell me the next appointment, please.
    SUBJ                                  Yes.                                          On (.) Wednesday.

b2  AGNT So, on Wednesday, right?             So, at 4 PM, right?       Good. What have[interrupt]
    SUBJ                           Yes. 4 PM.                        Yes.                Bingo.

b3  AGNT So, at that time there is "Hiking", right?       Fine. What[interrupt]
    SUBJ                                          No.                      BIN-GO (-) Game.

b4  AGNT So, at that time there is "Game", right?            Okay. Then I'll enter it as follows...
    SUBJ                                       (1.5) Yes.


b1  AGNT  Do you have[interrupt]
    SUBJ    Yes yesyes I understand (1.5) I'd like to discuss this with the people uhm directly whether a

b2  AGNT                                                          So you have "be arranged", right?
    SUBJ  bowling meeting in the evening (1.5) Could be arranged                                  Bowling.

b3  AGNT [glitch]So you have "be arranged", right?
    SUBJ                                         Yes, arrange (-) to discuss (2.0) a meeting for bowling

b4  AGNT  (2.0) Good. So you have bowling, right?
    SUBJ                                          [chuckling] Yes, that is good. Bowling.


c1  AGNT Then tell me the next appointment, please.
    SUBJ                                         I have uhm (-) today shopping *thr 3 PM 3 PM *appoin

c2  AGNT
    SUBJ  appointment with <Name> (.) and then I also(?) later go shopping later *thr 3 PM with <Name>

c3  AGNT
    SUBJ  (.) and (-) then I also go shopping (-) later
```

Figure 5: Examples of observed interaction styles (autonomous study): **Top:** older adult, brief but casual style; **second from top:** older adult, brief style; settling on alternative / partial event description; **second from bottom:** older adult, more verbose style; renegotiation; **bottom:** person with noticeable cognitive impairment, verbose turns, exacerbated by dysfluent and unclear articulation.

8. References

[1] GUIDE Consortium, *User Interaction & Application Requirements - Deliverable D2.1*, 2011.

[2] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, pp. 99—-112, 2010.

[3] R. Yaghoubzadeh, K. Pitsch, and S. Kopp, "Adaptive grounding and dialogue management for autonomous conversational assistants for elderly users," in *Proceedings of the 15th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 9238, 2015, pp. 28–38.

[4] D. Schlangen, T. Baumann, H. Buschmeier, O. Buß, S. Kopp, G. Skantze, and R. Yaghoubzadeh, "Middleware for incremental processing in conversational agents," in *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2010, pp. 51–54.

[5] D. Schlangen and G. Skantze, "A general, abstract model of incremental dialogue processing," in *EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009, pp. 710–718.

[6] S. Larsson, "Issue-based dialogue management," Ph.D. dissertation, University of Gothenburg, Sweden, 2002.

[7] H. van Welbergen, D. Reidsma, and S. Kopp, "An incremental multimodal realizer for behavior co-articulation and coordination," in *Proceedings of the 12th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 7502, 2012, pp. 175–188.

[8] M. P. Aylett and C. J. Pidcock, *The CereVoice Characterful Speech Synthesiser SDK*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 413–414.

[9] R. Yaghoubzadeh, M. Kramer, K. Pitsch, and S. Kopp, "Virtual agents as daily assistants for elderly or cognitively impaired people," in *Proceedings of the 13th International Conference on Intelligent Virtual Agents*, ser. LNCS (LNAI), vol. 8108, 2013, pp. 79–91.

[10] I. Grishkova, R. Yaghoubzadeh, S. Kopp, and C. Vorwerg, "How do human interlocutors talk to virtual assistants? A speech act analysis of dialogues of cognitively impaired people and elderly people with a virtual assistant," *Cognitive Processing*, vol. 15, p. 40, 2014.