SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 176–177.
Los Angeles, California, 21–23 September 2011.

Dialog Acts from the Processing Perspective in Task Oriented Dialog Systems

Markus Berg
University of Kiel &
University of Wismar
mail@mmberg.net

Bernhard Thalheim
University of Kiel
Technical Faculty

Antje Düsterhöft
University of Wismar
Faculty of Engineering

1 Introduction

The formulation *"I'd like to know what time it is"* has the same aim as *"What's the time?"*. Thus, we can easily see that different formulations can have the same intention. Consequently we learn that it is not possible to infer a one-to-one relationship between form and function. When developing a dialogue system, the main interest is *what* the user expects from the system, and not *how* he formulates his concern. So we propose a backend-oriented scheme for the description of dialog utterances. This scheme applies for three basic types of mixed-initiative systems that often have to be modeled: control systems (e.g. for controlling the lights in a room by speech), question-answering systems and information-seeking/booking-dialogues (i.e. the system asks questions in order to gain information that is necessary to fulfil the user's request). All of these systems have a task-related information exchange in common. Thus we don't classify by initiative (they are all mixed initiative) but by purpose and call those systems, according to (McTear, 2004, p.45), "task-oriented" dialog systems.

2 Modeling of Dialogs

While most dialog models start with linguistic aspects, we specify the model bottom-up. We have a backend and we know what it is able to do. Then we can find out how to address these functions, i.e. what linguistic form triggers which function.

2.1 Backend Functions

In the introduction we have mentioned three basic system types. This leads to three different categories of user aims:

- the user gives a command in order to make the system realize the request
- the user asks a question in order to retrieve an information
- the user gives information in order to enable the system to provide him with information

We now introduce appropriate functions that model these capabilities: `do`, `getInfo` and `setInfo`. The following examples are annotated with these basic functions and by this means indirectly describe the users aim, or the *intended perlocutionary effect*.

- Could you please switch on the light? → `do`
- Play some music → `do`
- How is the weather in London? → `getInfo`
- I'd like to start on May 4th → `setInfo`

2.2 Utterance Role and Speaker

We already observed that *form is not function*. Thus we should avoid the terms *question* and *answer* as they extremely relate to the form. So we replace them by the introduction of the terms *concern* and *reply*. A concern comprises all types of utterances that have the aim of causing a system reaction. This can be a regular question, a command, a request or just a wish. We summarize both a command and a question under the same category as they both constitute a form of system request. A reply is any possible response which satisfies the concern, i.e. an answer or an acknowledgement. Furthermore we introduce the *speaker* of an utterance leading to four base units (in combination with the utterance role): *user concern* (UC), *user reply* (UR), *system concern* (SC) and *system reply* (SR). After analysing several dialogs, we realized that the combination of SC and UR equals a UC: The system concern *"Tell me your destination"* and the user reply *"San Francisco"* equals the user concern *"I'd like to go to San Francisco"*. So if the system initiates a question, by answering it, the user states his own concern. For the

dialog manager it is important to know of the utterance role in order to infer the next dialog step. For the backend itself it does not matter if the request was a UC or a UR in consequence of a SC.

2.3 Selection of Dialog Acts

In order to model the user's intention we use dialog acts. While many dialogue act schemata suffer from the fact that form is mixed with function, we apply Bunt's second-level general-purpose functions (Bunt and others, 2010): *information seeking functions*, *information providing functions*, *commissives* and *directives*. In the types of dialog system described in this paper we don't need commissives, as we do not concentrate on human-human-like conversations. Of course the system could produce utterances like *"I will look for that"*, but from the backend processing perspective we don't need to understand promises, invitations, oaths or threats. From the *directives* we only use the *instructions*-category and rename it to *action requesting* in order to delimit from the form (instructions are often associated with imperatives). These dialog acts can now be related to our backend functions: an *information-seeking* dialog act will initiate the `getInfo` function, an *information-providing* act initiates the `setInfo` function and an *action-requesting* act leads to the *do* function. Apart from these acts, we also have to do with what in *DIT++* (Bunt and others, 2010) is called *social obligations*, like greeting/return greeting. These acts often don't need any backend access, which means that they bypass it. Moreover they form symmetric adjacency pairs as the reaction always belongs to the same dialog act category as the request. Hence we name them *copy* dialog acts.

2.4 Description of Dialog Utterances

We now have described two different classification approaches: the distinction into *concern* and *reply* as well as the differentiation between *information-seeking*, *information-providing*, *action-requesting* and *copy* acts. The attempt to integrate both into a common taxonomy fails as we have to do with different, independent dimensions. While the first approach describes the *role r* of an utterance, the second approach describes its *primary illocution i* and its derived *intended action a*. The *role* is important to enable the system to differentiate between *"I'd like to go to New York"* as a *concern* or as a *reply* to the question *"Where do you want to go?"*. Moreover an utterance is described by the *speaker s* and the *domain d* of the utterance, i.e. task oriented, dialog handling or social. It is further characterized by its *form f* (roughly equivalent with the secondary illocution) and the *range R* (only in case of inf.-seeking acts) of the resulting answer. Because range and action can be inferred from the primary illocution, we only have five independent attributes. Thus an utterance can be described by the following quintuple: $U = (s, r, i, d, f)$ where $s \in \{user, system\}$, $r \in \{concern, reply\}$, $i \in \{inf.seeking, inf.prov., act.req., copy\}$, $d \in \{task1, ..., taskN, dialog, social\}$ and $f = (sentence\ type, mode, verb, style, \dots)$. So the sentence *"Could you please close the window?"* can be described as: $U = (user, concern, action\ requesting, smart\ room, (question, subjunctive, close, formal))$. This would result in a `do` backend call and no range beacuse requests don't expect an answer.

3 Conclusion

In this paper we have discussed dialog acts from the processing perspective in mixed-initiative task-oriented dialogs. For the system it is most important to recognize what the user wants in order to be able to accomplish his needs. There is no need for the backend to know whether he formulated a request as an instruction or as a question. We identified the *role* of an utterance and three classes of backend functions which build the basis for the top level of a backend-motivated and formulation independent taxonomy of illocutionary acts. It comprises *information-seeking*, *information-providing*, *action-requesting* and *copy* acts. An extension of these attributes results in a quintuple for the description of utterances in a dialog which is a compact way of representing the user's aim and the intended system reaction in task-oriented mixed-initiative dialogs.

References

Harry Bunt et al. 2010. Towards an ISO standard for Dialogue Act Annotation. In Nicoletta Calzolari et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta. European Language Resources Association (ELRA).

Michael F McTear. 2004. *Spoken Dialogue Technology: Toward the Conversational User Interface*. Springer.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 178–179.
Los Angeles, California, 21–23 September 2011.

Unveiling the Information State with a Bayesian Model of the Listener

Hendrik Buschmeier and **Stefan Kopp**
Sociable Agents Group, CITEC and Faculty of Technology, Bielefeld University
PO-Box 10 01 31, 33501 Bielefeld, Germany
{hbuschme, skopp}@techfak.uni-bielefeld.de

Abstract

Attentive speaker agents – artificial conversational agents that can attend to and adapt to listener feedback – need to attribute a mental 'listener state' to the user and keep track of the grounding status of their own utterances. We propose a joint model of listener state and information state, represented as a dynamic Bayesian network, that can capture the influences between dialogue context, user feedback, the mental listener state and the information state, providing an estimation of grounding.

1 Introduction

Listeners providing communicative feedback reveal – not always deliberately – their mental state of processing to speakers (Allwood et al., 1992). Producing a backchannel (e.g., a quick 'yeah' or a nod) at appropriate places in the dialogue signals that they attend to and perceive what the speaker is saying. Looking puzzled or producing a hesitant 'yeah', on the other hand, might show that they have difficulties understanding what the speaker wants to express. Speakers attend to these signals, use them as information for grounding (Clark and Schaefer, 1989), and take them into account when producing their ongoing and subsequent communicative actions.

To be able to do this, speakers need to interpret a listener's feedback signal in its context and infer what the listener indicates, displays, or signals. Using this information, speakers can refine the model they have of their interlocutor and conjecture about the grounding status of dialogue moves in the information state that caused the listener to produce this feedback signal.

In the context of enabling virtual conversational agents to attend to and adapt to user feedback, we proposed that such an 'attentive speaker agent' maintains an 'attributed listener state' (ALS) of its user (Buschmeier and Kopp, 2011). The ALS is the part of the agent's interlocutor model that is particularly relevant when processing communicative listener feedback since it represents the agent's knowledge about the user's current ability to perceive and understand the agent's actions.

Here we propose a more sophisticated approach to ALS that integrates with the agent's information state and is modelled as a (dynamic) Bayesian network, giving the agent degrees of belief in the user's mental state as well as the grounding status of the current dialogue move.

2 Model

Figure 1 shows a schema of the model. Each step in time (t, t+1) corresponds to one dialogue move of the agent. The attributed listener state contains three nodes C, P and U that model whether the user is in contact with the agent and perceives and understands the agent's utterance. Allwood et al. (1992) propose that these functions of feedback relate to each other: being in contact is, for instance, a prerequisite for perception, which in turn is a prerequisite for understanding. We can easily capture these relations in terms of influences in our Bayesian network model. Evidence that contact is established increases the degree of belief in the user being able to perceive what is said, which in turn increases the degree of belief in her understanding the utterance.

Variables influencing the nodes in the ALS are hidden in the boxes 'Context' and 'User FB'. Important

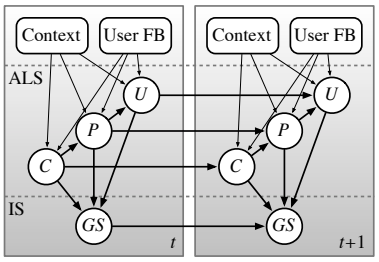

Figure 1: Attributed listener state (ALS) and information state (IS) modelled as one dynamic Bayesian network. User feedback and dialogue context influence the degrees of belief in contact, perception and understanding (C, P, U) in the ALS. These determine the grounding status (GS) of the current dialogue move kept in the information state.

contextual factors for perception might, for example, be whether noise is present in the environment, or the occurrence probability of the agent's utterance calculated by an n-gram language model. The type of the user's feedback function as well as certain features of the feedback signal obviously also influence the ALS nodes. Presence of feedback signalling 'understanding' increases the agent's degree of belief in U and should certainly influence P and C as well. Similarly, the influence of a prosodically flat 'yeah' on U should be smaller than an enthusiastic one.

Our model also enables the agent to relate different kinds of user feedback to the grounding status of the dialogue move it refers to. This is modelled with a node GS in the information state part of the model. If we have evidence from feedback signals that the user understood the agent's utterance, the degree of belief in the dialogue move being in the common ground should be high. If, in contrast, the agent only got feedback of the communicative function 'perception', the degree of belief in the dialogue move being grounded should be lower. Nevertheless, depending on the context (for example, the dialogue move is simple and there is no apparent reason for the user not to understand it) the degree of grounding can still be high enough to take it as being grounded.

Finally, our model is a *dynamic* Bayesian network since the previous dialogue move influences the current one. If the previous move has a high degree of being grounded this should increase belief in the current move being grounded as well. Similar assumptions can also be made about the values of C, P and U in the ALS.

3 Discussion and Conclusion

We presented first steps towards a joint model of attributed listener state and information state for artificial conversational agents. Modelled as a dynamic Bayesian network, it can easily capture the influences between dialogue context, user feedback, the mental listener state the agent attributes to the user and the grounding status of the agent's dialogue moves.

This is an improvement on our previous model of listener state (Buschmeier and Kopp, 2011), since dialogue context and features of feedback signals can be taken into account during state estimation. In contrast to state of the art models of (degrees of) grounding (Traum, 1994; Roque and Traum, 2008) the model presented here allows for continuous instead of discrete grounding values, based on the user's feedback signals and the dialogue context.

Several issues, however, have not yet been addressed. It is, for instance, still unclear how exactly the timing of feedback signals will be handled. Furthermore, although a simple hand-crafted prototype looks promising, the question how such a network can be learnt is open as well.

Acknowledgements This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence in 'Cognitive Interaction Technology' (CITEC).

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Hendrik Buschmeier and Stefan Kopp. 2011. Towards conversational agents that attend to and adapt to communicative user feedback. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, pages 169–182, Reykjavik, Iceland.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Antonio Roque and David R. Traum. 2008. Degrees of grounding based on evidence of understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 54–63, Columbus, OH.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester, Rochester, NY.