A Multilingual Expectations Model for Contextual Utterances in Mixed-Initiative Spoken Dialogue

Hartwig Holzapfel, Alex Waibel

InterACT Research, Interactive Systems Labs
Universität Karlsruhe
hartwig,waibel@ira.uka.de

Abstract

This paper describes a model of generating expectations that are used to improve speech recognition and to resolve elliptical expressions in dialogue context. The algorithm is domain and language independent and part of the dialogue manager. We use the expectation model to weight a speech recognizer's grammar rules in dialogue context which improves recognition rates significantly as shown in the evaluation.

We explain what types of expectations the system can generate and give a classification of system actions based on speech act theory, explain the resolution of elliptical expressions and their interpretation in context, and evaluate the presented algorithm in a multilingual system with English and German speech recognition.
Index Terms: multilingual dialogue, context, expectation model, grammars, speech recognition.

1. Introduction and Related Work

We have previously presented an approach for context dependent weighting of grammar rules [1] with improvement of recognition results that goes in accordance to other work, e.g. [2] that show improvement by using subgrammars. In this work, we present a new approach to create an expectations model that selects rule weights based on domain an language independent algorithms in the context of information requests by the dialogue system. In addition, the model is capable of resolving elliptical expressions.

To generate expectations and correlated grammar rules, we apply a categorization of system utterances according to speech act theory. We describe what kind of input can be expected and how it is interpreted within the given context. The model generates expectations that depend on the type of speech act that was uttered by the system and the type of requested information. The system also needs to cope with conversation acts that are not directly related to the system's request, such as correcting information of previously given information, which are not covered in this work.

Other work exists that make use of different subgrammars based on the current active dialogue move. For example the information state update (ISU) dialogue manager [3] applies grammar-switching, based on the assumption that dialogues consist of adjacency pairs so that answers follow questions, commands are acknowledged in general, etc. so that the subgrammars can be determined by this mechanism. Most researchers working on context control of a speech recognizer by means of a dialogue manager use different stages and language models: A general n-gram language model which is used at the beginning and in underspecified situations and a specialized language model which can be an n-gram language model or a grammar-based one and is used in specific situations based on the preceding system prompt [4, 5]. In [6], the state-independent n-gram language model is also combined with a state-dependent finite state grammar by comparing the acoustic confidence scores. In this way, perplexity and word error rates can be reduced significantly.

As in [3], our approach makes use of the assumption of adjacency pairs. Our approach extends the context switching model in [3] with a more detailed speech act categorization of system utterances. In addition, it uses information about the requested types required by dialogue goals. Furthermore, our approach can be applied to any semantic grammar. In contrast to generating different subgrammars, our approach uses only a single grammar and punishes or privileges different grammar rules. Speech act theory [7] has become common to model and categorize specific actions in dialogue systems. Beyond speech acts, Traum and Hinkelmann [8] describe conversation acts that cover additional actions in dialogue such as turn taking and grounding. They define four speech act categories, 'turn-taking', 'grounding', 'core speech acts', and 'argumentation'. Different annotation and labeling schemes have been developed for speech acts like DAMSL[1], or SWBD-DAMSL. Our dialogue system uses a specific speech act called 'information request' that models almost any action or utterance that expects an answer from the conversation partner. For our analysis a more detailed classification of information requests is required, e.g. as used in CLARITY [9]. The CLARITY annotation scheme is based on DAMSL and SWBD-DAMSL but provides more details especially for information requests. The categories for speech acts used in our system (as system utterances) are similar to those use in CLARITY. They are described later in this paper.

2. Dialogue System Components

For dialogue management we use the TAPAS dialogue framework for multimodal and multilingual dialogue systems [10]. For speech recognition, we use the Janus Recognition Toolkit (JRTk) with the Ibis single pass-decoder [11]. We use the option of Ibis to decode with context free grammars (CFG) instead of statistical n-gram language models (LM). These context free grammars are generated by the dialogue manager that uses the same grammars for language understanding. In the same way, the dialogue manager can be used in combination with other speech recognizers that can decode with context free grammars, by providing grammars in SOUP, PHOENIX, JSGF, and Microsoft SAPI formats.

The dialogue system uses semantic grammars to interpret spoken (or typed) input. The integration of the grammars (natural lan-

[1] http://www.cs.rochester.edu/research/cisd/resources/damsl/

guage understanding) into the system is shown in figure 1. Typed features structures (TFS) are used to represent semantic input and discourse information. The processing of the dialogue algorithms and the discourse representation are language independent. This allows using general discourse and dialogue algorithms, including algorithms that define the dialogue context and expectation model on a semantic level.

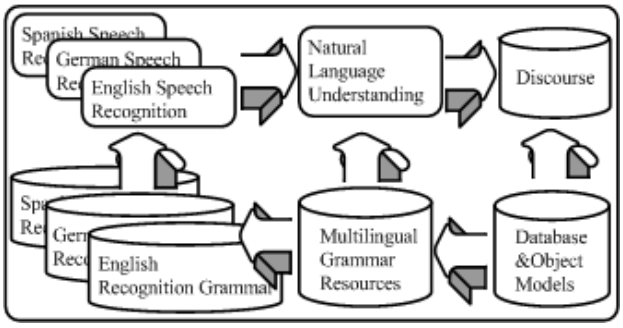

Figure 1: Flow diagram visualizing the integration of recognition and understanding components into the dialogue system.

3. Contextual Model and Expected Information

The dialogue manager maintains a context model to interpret spoken input within dialogue context and to organize speech recognition grammars. Here, we describe a rule based approach applying speech act theory to generate expectations for the next user input.

3.1. Expected Information

The expectations context serves two purposes. First, the expected information is used to increase the weights of grammar rules that relate to the expected information. Increasing grammar weights is not only used to improve speech recognition, but implicitly also selects the most relevant semantic representation for ambiguous semantic parses of the user utterance. Second, the expected information is used to resolve elliptical expressions and to correctly integrate spoken input into the discourse representation.

The expectations context is (non-exclusively) influenced by the type of information that is requested from the user and the type of speech act that the system performs to request the desired information. Other parts that influence the expectations context is information that is intended to continue an active dialogue goal. Of course, all other actions by the user (speech acts, dialogue acts) are still allowed and possible, but their expectations remains the same and is not specifically changed.

The following list shows the order of importance of expected information (i) direct response to question (ii) indirect response to question that implicitly answers the question (iii) response to question in combination with repeating information (iv) repairing previously given information (v) giving information for one of the active discourse segments.

3.2. Speech Acts

A dialogue move is selected by the dialogue strategy based on the purpose that it serves. A move can request new information, generate clarification questions, give information, or generate confirmations. Each purpose leads to a different response by the user. The dialogue moves that generate questions describe the type of question by assigning one of the speech act categories as shown in figure 2. The description of the speech act category is language independent, and the question is generated in the desired language corresponding to the given speech act.

The speech acts shown in figure 2 all inherit from a general node 'info-request'. The 'info-request' is the most general element to describe a question (or any other kind of action) that expects an answer relating to this question. The top level node corresponds to the speech act category describing an information request in DAMSL. However, for our purposes the DAMSL tagging scheme is not detailed enough, so we extended the scheme to the following speech acts. 'qst_yesno' expects 'yes' or 'no' as answer; 'qst_wh' is a category for all 'wh'-questions such as who, what, when, where, and questions asking for numbers, which represent the subcategories of 'qst_wh'; 'qst_or' is a question, where the user can select one of the presented alternatives, e.g. "do you want x or y?"; 'qst_open' is an open question where the user is free to answer, and no explicit expectation can be generated based on the speech act. Here, only the type of requested information determines an expectations context. The last type 'qst_open' cannot restrict the expectations, whereas all others can. Core and Allen [12] use a category that combines different actions that influence the addressee's future action. This category contains 'open option' and 'directive'. Subtypes of 'directive' are 'info-request' and 'action-directive'. Our approach goes in-line with this description and refines the information request category to do more detailed analyses.

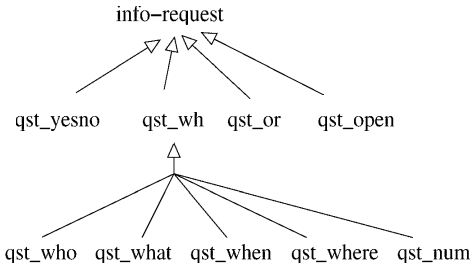

Figure 2: Categorization with inheritance model for subtypes of the 'question' speech act category, which are used by the system to generate information requests.

3.3. Target Types

Independent of the speech act that is used for the system's question, a specific type of information is requested. We call this a target. A target is a piece of information that is described by its semantic type, a reference to the dialogue goal that defines the frame for the target and the path to the desired information within the specified goal. A path refers to a specific node in a typed feature structure. As already briefly described, dialogue goals, as well as the discourse representation are modeled with typed feature structures. The dialogue goal, that defines the structure of the expected information, and the existing information in discourse that relates to the referenced dialogue goal, define the context for the targeted information. Information required by the dialogue goal, which is

not given in discourse is expected to be delivered by the user. Information that is already given in discourse is either expected to be repeated/confirmed or to be repaired. This relates to the listing in section 3.1.

As already mentioned, the target references to some specific information in a dialogue goal. When asking for this piece of information, we expect to be able to extract this from the user's answer. The answer can be elliptic, giving directly the desired information, such as 'two' in reply to asking 'how many persons?'. Or, the answer can be embedded within a complete sentence. The construction algorithm for generating expectations based on the target information first picks the target type and then walking up in the TFS path, picks all parents recursively. This results in a list of TFS nodes describing the targeted information within more or less context of the dialogue goal.

A small example illustrates the algorithm. Figure 3 shows the required information for a dialogue goal. When executed, it instructs the robot to serve a cup of coffee, with the options of adding milk or sugar. The path '$OBJ|MILK$' references the type 'att_milk' with its sub-feature. To get information about the type 'att_milk', the system generates an information request. The target is defined by the dialogue goal and the path '$OBJ|MILK$' that references the type 'att_milk'. The expected response can be 'yes' or 'no', which both directly respond to the given question. The answer 'yes' is converted to the following TFS and is then unified with the discourse representation with the prefix path '$OBJ|MILK$'.

```
{att_milk BOOL [base:boolean]}
```

Note that the expectation model also covers formulations like 'with milk please' or 'I would like my coffee with milk and sugar'. The answer 'with milk' is first converted to the above TFS and is then unified with the discourse with the prefix path '$OBJ|MILK$'. The same works for the response 'with milk and sugar' which describes a more complex construct than 'with milk', but matches the expected information as well.

```
[act_bring
OBJ [obj_coffee
     MILK  [ att_milk
             BOOL [base:boolean] ]
     SUGAR [ att_sugar
             BOOL [base:boolean] ]
    ]
]
```

Figure 3: A TFS describing the precondition of the 'make-coffee' goal.

3.4. Generating Subgrammars

After generating a list of possible TFS nodes that semantically represent possible answers, grammar rules are selected that can be converted to the desired semantic representation. The algorithm to find these grammar rules is constructive and uses induction to search all conversions of grammar nodes to a given semantic representation, where the semantic type of the grammar node matches the desired semantic type.

This approach of selecting grammar nodes is language independent. It can thus be applied to multilingual resources, as used by our dialogue system [10]. The approach supports multilingual as well as multiple monolingual speech recognizers. Both use separate language models for different languages [13].

	baseline		improvement	
	WER	SER	WER	SER
Responses (C)	29.11%	30.00%	8.87%	8.89%
Overall (C)	22.74%	31.89%	3.56%	3.88%
Responses (D)	36.77%	39.60%	16.45%	11.86%
Overall (D)	31.41%	45.33%	6.66%	5.15%

Table 1: Set 1: Close (C) and distant (D) talking word and sentence error rates together with their relative improvements

.

3.5. Experimental Results

We compared the speech recognition results of a system which uses the context dependent weighting of rules to one without it, on human-robot dialogs in the domain of a household robot. We evaluated the approach on two different interaction sets. Both sets were recorded with close talk (C) and distant speech (D) microphones.

Set 1 consists of requests for actions by the user (User Commands), responses by the system including clarification requests or queries for missing information where necessary, and user replies (Response Set). It contains eight speakers, all interaction are in English.

Set 2 was recorded in a different setup with different users in multimodal human-robot interaction, where the robot plays the part of a bartender to serve different objects from the table in front of him. The user responded to questions from the robot asking for object properties [14]. The full set contains 314 utterances for English and 171 utterances for German, each including some segmentation errors (e.g. utterance was recognized though nothing was said) and out-of-domain utterances that are not covered by the system. The constrained set excludes out-of-domain utterances and segmentation errors, which results in a set size of 267 utterances for English and 152 utterances for German.

Three categories of weights have been used: unexpected, normal and expected. The weights for these expectation categories that are applied by the speech recognizer have already been trained in previous work [1]. The acoustic model that we have used for English during our experiments was trained on nearly 95hrs of close talking meeting data mixed with 180hrs of Broadcast News data. It is a slimmed down version of a system, which was used in the NISTs RT-04S evaluation [15].

Evaluation details on Set 1 with handcrafted weighting have already been presented in [1], in our experiment the selected rules offer a marginally broader selection of rules that however, did not have any effect in word-error rate, presumably because the handcrafted selection was already very good. Table 1 shows the baseline (no rule weighting) and the relative improvements achieved on Set 1, measured with word error rate (WER) and sentence error rate (SER). The evaluation on the Set 2 is shown in table 2, where the figures for German (close-talk) and English (close-talk and distant-speech) are given. Here, we show the numbers for word-error rate (WER) and semantic concept error rate (CER) for both close-talk and distant-speech on the full set ('all') and a constrained set ('i.d.'). The relative improvements for the numbers are computed in table 3. We have chosen the concept error rate (CER)

	baseline		improved	
	WER	CER	WER	CER
i.d. (C) - English	12.8%	7.8%	10.1%	5.2%
All (C) - English	28.3%	15.9%	26.2%	13.7%
i.d. (D) - English	32.1%	19.9%	29.2%	15.7%
All (D) - English	41.9%	26.4%	39.7%	21.7%
i.d. (C) - German	9.8%	4.6%	9.1%	3.9%
All (C) - German	21.3%	13.1%	20.9%	12.0%

Table 2: Set 2: all utterances and in domain 'i.d.' utterances (parsable input) for close (C) and distance (D) talking conditions for English and close talk for German. Evaluated on word error rates (WER) and semantic concept error rates (CER).

	impr. WER	impr. CER	rel.impr. WER	rel.impr. CER
i.d. (C) - English	2.7%	2.6%	21.1%	33.3%
All (C) - English	2.1%	2.2%	7.4%	13.8%
i.d. (D) - English	2.9%	4.2%	9.0%	21.1%
All (D) - English	2.2%	4.7%	5.3%	17.8%
i.d. (C) - German	0.7%	0.7%	7.1%	15.2%
All (C) - German	0.4%	1.1%	1.9%	8.4%

Table 3: Absolute and relative improvements on Set 2.

since it is useful to measure the effects on a dialogue system. It is more informative than word error rate and also ignores semantically irrelevant errors. It is computed similar to the common word error rate by simply comparing IDs of semantic concepts. The results for the German baseline (without rule weighting) are already very good, probably because the system has been used by people that regularly speak to ASR systems, which is not the case for English. So, only few errors remain in the German set that are not due to segmentation errors or noises. Thus, the improvements are smaller than for English. It is interesting to see that the improvements for the concept-error rate, which is more important for the dialogue system are more significant than the improvement for the word-error rate.

4. Conclusions

We have presented our work on building an expectations context model to improve speech recognition and to facilitate resolution of elliptical expressions. Experiments have shown improvement in word accuracy, sentence and semantic concept recognition rates over the baseline system(s); on Set 1 especially for distant speech, on Set 2 for both distant and close talk. Set 2 shows improvements also for languages other than English. The presented approach to select preferred grammar rules is domain and language independent, since it uses only speech act theory and ontological information, it is generic and can be combined with any application.

5. Acknowledgements

This work was supported in part by the German Research Foundation (DFG) as part of the Collaborative Research Center 588 "Humanoid Robots - Learning and Cooperating Multimodal Robots".

6. References

[1] C. Fügen, H. Holzapfel, and A. Waibel, "Tight coupling of speech recognition and dialog management - dialog-context grammar weighting for speech recognition," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, 2004.

[2] A. Stent, J. Dowding, J. M. Gawron, E. O. Bratt, and R. Moore, "The commandtalk spoken dialogue system," in *Proceedings of the 37th Annual Meeting of ACL*, 1999.

[3] O. Lemon, "Context-sensitive speech recognition in isu-dialogue systems: Results for the grammar-switching approach," in *Catalog '04. Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue*, 2004.

[4] W. Xu and A. Rudnicky, "Language modeling for dialog system," in *Proc. of the Int. Conf. of Speech and Signal Processing (ICSLP'00)*, 2000.

[5] E. Fosler-Lusier and H.K. J. Kuo, "Using semantic information for rapid development of language models within asr dialogue systems," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'01)*, 2001.

[6] R. A. Solsona, E. Fosler-Lussier, H.-K. J. Kuo, A. Potamianos, and I. Zitouni, "Adaptive language models for spoken dialogue systems," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'02)*, 2002.

[7] D.R. Traum, "Speech acts for dialogue agents," *Foundations of Rational Agency*, pp. 169–201, 1999.

[8] D.R. Traum and E.A. Hinkelman, "Conversation acts in task-oriented spoken dialogue," *Computational Intelligence*, vol. 8(3), pp. 575–599, 1992.

[9] L. Levin, A. Thyme-Gobbel, A. Lavie, K. Ries, and K. Zechner, "A discourse coding scheme for conversational spanish," in *Proc. of Int. Conf. on Speech and Language Processing (ICSLP 98)*, 1998.

[10] H. Holzapfel, "Towards development of multilingual spoken dialogue systems," in *Proc. of the 2nd Language & Technology Conference (L&T'05)*, 2005.

[11] H. Soltau, F. Metze, C. Fuegen, and A. Waibel, "A one pass- decoder based on polymorphic linguistic context assignment," in *Proc. of the Automatic Speech Recognition and Understanding Workshop (ASRU-2001)*, 2001.

[12] M.G. Core and J.F. Allen, "Coding dialogs with the damsl annotation scheme," in *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA, November 1997.

[13] T. Schultz, S. Stüker, H. Soltau, F. Metze, and C. Fügen, "Efficient handling of multilingual language models," in *Proc. of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2003.

[14] T. Prommer, H. Holzapfel, and A. Waibel, "Rapid simulation-driven reinforcement learning of multimodal dialog strategies in human-robot interaction," in *Proc. of Interspeech (ICSLP)*, 2006.

[15] F. Metze, Q. Jin, C. Fügen, Y. Pan, and T. Schultz, "Issues in meeting transcription the meeting transcription system," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP 2004)*, 2004.