Tight Coupling of Speech Recognition and Dialog Management – Dialog-Context Dependent Grammar Weighting for Speech Recognition

Christian Fügen, Hartwig Holzapfel, Alex Waibel

Interactive Systems Labs
Universität Karlsruhe (TH), Germany
{fuegen|hartwig|waibel}@ira.uka.de

Abstract

In this paper we present our current work on a tight coupling of a speech recognizer with a dialog manager and our results by restricting the search space of our grammar based speech recognizer through the information given by the dialog manager.

As a result of the tight coupling the same lingustic knowledge sources can be used in both, speech recognizer and dialog manager. Furthermore, the flexible context-free grammar implementation of our speech decoder Ibis allows weighting of specific rules at run-time to restrict the search space of the recognizer for the next decoding step. These rules are given by the dialog manager depending on the current dialog context.

With this approach we were able to reduce the word error rate of user responses to system questions by 3.3% relative for close talking and 16.0% relative, when using distant speech input. The sentence error rates were reduced by 2.2%, 9.2% respectively.

1. Introduction

In the upcoming field of humanoid and human-friendly robots, the ability of the robot to interact in a simple, unconstrained and natural way with its users is of great importance. Therefore, a user should neither be restricted by a command based speech interface nor should he be forced to wear head-mounted microphones in order to communicate with the robot. Instead, spontaneous, mixed initiative speech dialogs recorded by distant microphones should be possible. To improve the system understanding performance in such difficult situations, we are working on a tight coupling of the speech recognizer and dialog manager.

Most current human-machine interfaces consist of three main components: a speech recognizer, a natural language parser and a dialog manager [1, 2, 3]. The output of a speech recognizer in form of n-best lists or lattices is given to the NL parser, which builds one interpretation out of it and then passes it on to a dialog manager for making a system decision depending on the current dialog context. Our proposed tight coupling makes a separate NL parser superfluous and the human-machine interface more robust, easier to maintain and hence more portable to new domains.

One well-known technique for making human-machine interfaces more robust is dialog context or state dependent language modeling for the speech recognizer. In most cases n-gram language models are used for speech recognition and therefore, e.g., language model interpolation with dialog state specific models [4, 5]. The problem hereby is the data sparseness of the state dependent sub-corpora. Therefore in [5] the use of maximum entropy exponential models are suggested.

Another solution is splitting language modeling into a dialog dependent concept modeling using n-grams and a syntactic modeling consisting of a set of SCFGs [6]. All these approaches are using a seperate rescoring pass with a dialog dependent language model over the speech recognition output (n-best lists, lattices) that was produced with a common language model. In [7] a n-gram based recognizer and a dialog-state specific finite state grammar based recognizer are run in parallel, and the systems decides, whether the input sentence was in or out-of grammar. Differences are also in the way the dialog context is computed. This can be done e.g. by simply using the parser output, by topic detection [8] or with the help of semantic classification trees [9].

In contrast thereto, we are using a more integrated approach, where no separate rescoring is neccessary. The information about the predicted dialog context for the next user utterance is directly integrated into the next decoding step of a context-free grammar based speech recognizer. For predicting the dialog context, the already parsed output of the speech recognizer is used.

2. Speech Recognition

For speech recognition we are using the Ibis decoder [10], which was developed at the University of Karlsruhe as part of our Janus Recognition Toolkit (JRTk) [11]. Besides several other advantages over our old three-pass search such as smaller memory usage and higher recognition speed, Ibis allows us to decode along context-free grammars in addition to the classical statistical n-gram language models.

2.1. Context-Free Grammar Decoding

Using grammars instead of n-gram language models is of advantage especially in small domains, such as our household scenario. In such domains there is normally little domain dependent data available for the training of robust statistical n-gram language models.

Our context-free grammar implementation in Ibis also has several other advantages. Rather than compiling one network or a finite state graph out of the grammar description files, we use a more dynamic approach, where several rule based recursive transition networks (RTNs) are linked together by their non-terminal symbols. During decoding, a rule stack gives us the ability to enter or leave the linked networks. This kind of network organization gives us high flexibilty when used in combination with dialog managers. Furthermore, it gives us the ability to work with real context-free grammars.

In most cases, we work with non-statistical semantic grammars, i.e. each transition to the next word has the same language model score, whereby terminals are grouped to non-terminal symbols by their semantical meaning.

3. Dialog Management

Our dialog manager Tapas is a recently developed collection of dialog processing and dialog management tools. The dialog management algorithms are based on the approaches of the language and the domain independent dialog manager Ariadne [12]. Tapas inherits the features from Ariadne and offers a tighter integration with the speech recognizer. As Ariadne, Tapas is specifically tailored for rapid prototyping because only the domain and language dependent components have to be implemented for new applications, whereas the general concepts are already available and can be reused. Furthermore, possibilities to evaluate the dialog state and general input and output mechanisms are already implemented which can then be applied in the actual application. For the domain-dependent part, we have developed different kinds of resources: An ontology, a specification of the dialog goals, a data base, a context-free grammar and generation templates.

The user utterance is parsed by means of a context-free grammar which is enhanced by information from the ontology defining all the objects, tasks and properties about which the user can talk. After parsing, the parse tree is converted into a semantic representation with conversion rules. The dialog manager uses typed feature structures [13] to represent semantic input and discourse information. The semantic input is combined with the already existing discourse information. Dialog goal representations define the information that must be available in the discourse to execute a dialog goal. More information to disambiguate between different dialog goals and to fullfill the requirements of the goal is obtained by executing generation templates. These templates are selected by the dialog strategy. They can query the user for the required information or execute any other dialog-controlled action.

4. Tight Coupling of the Speech Recognizer and Dialog Manager

Our goals for a tight coupling between a speech recognizer and a dialog manager are to share as much information between these two components as possible, to improve speech recognition and hence system understanding performance. Therefore, the implementations of Ibis and Tapas allow us to share the linguistic knowledge sources, i.e. context-free grammars, which gives us the ability to use the results of one component directly for improving the performance of the other component in the next step.

Due to the fact, that Ibis uses linked RTNs for its internal grammar representation, the original grammar structure can be directly accessed, which has several advantages:

- Ibis can also be used as a parser for natural language processing. Therefore, a separate parser is superfluous. The recognized and parsed output can be directly given to Tapas.

- Rules can be activated/deactivated or weighted (penalized) during run-time, which can be used, e.g., to restrict the decoding process to sub-grammar parts only.

In our current human-machine interface implementation all the entry rules of the used grammars are divided into two sets, a ResponseSet and a QuerySet. The ResponseSet consists only of rules, which are less likely to be used at the beginning of a dialog, i.e. consists mainly of rules which cover all responses to clarification questions. The QuerySet contains all the rules which are most likely used at the beginning of the dialog. Depending on the current dialog context, Tapas decides which rules are most likely used by the user for its next query/response and gives this information to Ibis. At the beginning of the dialog, all rules of the ResponseSet are penalized, whereas during the dialog the specified set of rules out of the ResponseSet given by Tapas are preferred over all others. It should be emphasized that still other user inputs can be recognized by Ibis which is conform to a mixed initiative dialog system.

To improve the understanding performance, Ibis gives an n-best list of parses to Tapas, instead of the first best parse only. In the future we want to expand this by using parsed, i.e. semantically annotated lattices together with confidences. To predict the next user input we are using the missing discourse information to complete a dialog goal, together with the information if the dialog is finalized or not.

5. Experimental Results

We compared the speech recognition results, i.e. the word error rates (WER) and sentence error rates (SER), in the domain of a household robot for a system which uses the context dependent weighting of rules to one without it. We used both, close and distant talking microphones, to meassure the difference in performance gain by introducing the new methods. As mentioned above, the preferred speech input for human-robot interfaces is distant speech.

Therefore, we collected a set of dialogs of different speakers which consists of spontaneous speech queries and responses to clarification questions from the robot. For the distant data only one microphone at a distance of about 2-3m from the speaker was used, which means that no array processing could be done to improve the speech recognition results. Given the parsed transcripts of the pre-recorded dialogs, the dialog manager was used to compute the preferred rules for the next user response depending on the dialog context. The weighting parameters for the grammar sets were optimized on a cross validation set. The details of the evaluation set and the grammar size is given in Table 1.

Speakers	8
Sentences	646
User queries	346
User responses	300
Duration	~21 min
CFG Rules	142
Entry Rules	47
ResponseSet size	32

Table 1: Details of the evaluation set and the context-free grammar.

5.1. Acoustic Model

The acoustic model that we have used for 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 NIST's RT-04S evaluation [14]. It is a fully continous system consisting of 6000 codebooks with 185k Gaussians over a 42-dimensional feature space based on MFCCs after LDA and global STC transforms with utterance based CMS. Incremental constrained MLLR is used in decoding to compensate for different channels effects.

5.2. Results

In Table 2 the baseline results are reported. It can be seen that the recognition results for the user responses are worse than for the user queries. Especially for the distant condition the WERs for the user responses are about 50% worse than for the user queries. The sentence error rates do not vary as much as the word error rates. The speech recognizer runs for both, the close and the distant talking condition in less than 0.5xRT on a 1.7GHz Pentium M, so that there is enough room for other components of the human-machine interface.

	WER	SER
User queries (C)	20.21%	34.10%
User responses (C)	30.28%	30.67%
Overall (C)	23.52%	32.51%
User queries (D)	30.53%	51.15%
User responses (D)	43.77%	43.62%
Overall (D)	34.86%	47.66%

Table 2: Close (C) and distance (D) talking word and sentence error rates (baselines).

When using our context dependent grammar weighting as described above it can be seen in Table 3 that there is an overall reduction of the WER of 3.3% for the close and 9.9% for the distant talking condition. Whereas there is a smaller gain for the user queries, the user responses are recognizd much better. It can also be seen, that the relative improvement increases for the distant condition.

			improvement	
	WER	SER	WER	SER
Queries (C)	19.63%	33.53%	2.87%	1.67%
Responses (C)	29.11%	30.00%	3.86%	2.18%
Overall (C)	22.74%	31.89%	3.32%	1.91%
Queries (D)	28.81%	50.29%	5.63%	1.68%
Responses (D)	36.77%	39.60%	15.99%	9.22%
Overall (D)	31.41%	45.33%	9.90%	4.89%

Table 3: Close (C) and distant (D) talking word and sentence error rates together with their relative improvements compared to Table 2.

To analyse the influence of mis predicted grammar rules which leads to a wrong grammar weighting, we replaced in $1/3$ of the test set the correct prediction by a different one. As can be seen in Table 4 the relative improvement also goes down by nearly $1/5$, but we were not otherwise penalized for the mis predictions.

6. Conclusions

We described a tight coupling of the speech recognizer and the dialog manager for implementing a human-machine interface. The tight coupling allows us to work

			improvement	
	WER	SER	WER	SER
Queries (C)	19.63%	33.53%	2.87%	1.67%
Responses (C)	29.93%	31.33%	1.16%	-2.15%
Overall (C)	23.01%	32.51%	2.17%	0.00%
Queries (D)	29.10%	50.29%	4.68%	1.68%
Responses (D)	38.20%	40.27%	12.73%	7.68%
Overall (D)	32.07%	45.64%	8.00%	4.24%

Table 4: Close (C) and distant (D) talking word and sentence error rates together with their relative improvements compared to Table 2, with mis-predicted grammar rules

with the same semantic, non-statistical grammars in both, the speech recognizer and the dialog manager, and makes a separate parser superfluous. As a result of this tight coupling and due to our flexible context-free grammar implementation in our speech decoder the dialog manager has the ability to control the search space of the recognizer depending on the dialog context. This was done by weighting specific entry rules in the semantic, non-statistical grammars. This information can be directly used for the next decoding step, so that a separate rescoring pass is not neccessary.

In the domain of a household robot we compared the speech recognition results of a system which uses the context dependent weighting to one without it for close and for distant speech input, which can be seen as the more preferred speech input for human-robot interfaces. We were able to reduce the overall word error rate of user queries and responses by 3.3% relative for close talking and 9.9% relative, when using distant speech input. The sentence error rates were reduced by 1.9%, 4.9% respectively. It can also be seen that the gain is larger for the user respones than for the user queries.

7. Acknowledgements

We would like to thank Petra Gieselmann for her help during the data collection, Carolin Anne Linnenbank for transcribing the collected data and Sebastian Stüker for training the acoustic model for the speech recognizer.

Part of this work was funded by the German Research Agency (DFG) under Sonderforschungsbereich 588 - Humanoid Robots[1] and by the European Union under IST project No. IST-2000-28323 (FAME[2]: "Facilitating Agents for Multi-cultural Exchange").

8. References

[1] Y. Gao, H. Erdogan, Y. Li, V. Goel, and M. Picheny, "Recent advances in speech recognition system for ibm darpa communicator," in *Proceedings of the Eurospeech*, Aalborg, Denmark, September 2001.

[2] B. Pellom, W. Ward, and S. Pradhan, "The cu communicator: An architecture for dialogue systems," in *Proceedings of the ICSLP*, Beijing, China, October 2000.

[3] S. Seneff, R. Lau, and J. Polifroni, "Organization, communication, and control in the galaxy-ii conversational system," in *Proceedings of the Eurospeech*, Budapest, Hungary, September 1999.

[4] W. Xu and A. Rudnicky, "Language modeling for dialog systems," in *Proceedings of the ICSLP*, Beijing, China, October 2000.

[5] K. Visweswariah and H. Printz, "Language models conditioned on dialog state," in *Proceedings of the Eurospeech*, Aalborg, Denmark, September 2001.

[6] K. Hacioglu and W. Ward, "Dialog-context dependent language modeling combining n-grams and stochastic context-free grammars," in *Proceedings of the ICASSP*, Salt Lake City, Utah, May 2001.

[7] R. A. Solsona, E. Fosler-Lussier, H.-K. J. Kuo, A. Potamianos, and I. Zitouni, "Adaptive language models for spoken dialogue systems," in *Proceedings of the ICASSP*, Orlando, Florida, May 2002.

[8] I. R. Lane, T. Kawahara, and T. Matsui, "Language model switching based on topic detection for dialog speech processing," in *Proceedings of the ICASSP*, Hong Kong, April 2003.

[9] Y. Estève, F. Béchet, and R. de Mori, "Dynamic selection of language models in a dialogue system," in *Proceedings of the ICSLP*, Beijing, China, October 2000.

[10] H. Soltau, F. Metze, C. Fügen, and A. x Waibel, "A One Pass-Decoder Based on Polymorphic Linguist ic Context Assignment," in *Proceedings of the ASRU*, Madonna di Campiglio, Trento, Italy, December 2001.

[11] M. Finke, P. Geutner, H. Hild, T. K. mp, K. Ries, and M. Westphal, "The Karlsruhe-VERBMOBIL Speech Recognition Engine," in *Proceedings of the ICASSP*, Munich, Germany, 1997.

[12] M. Denecke, "Rapid prototyping for spoken dialogue systems," in *Proceedings of the 19th International Conference on Computational Linguistics*, Taiwan, 2002.

[13] B. Carpenter, *The Logic of Typed Feature Structures*. Cambridge University Press, 1992.

[14] F. Metze, Q. Jin, C. Fügen, Y. Pan, and T. Schultz, "Issues in Meeting Transcription – The Meeting Transcription System," in *submitted to ICSLP*, Jeju Island, Korea, October 2004.

[1] http://www.sfb588.uni-karlsruhe.de

[2] http://isl.ira.uka.de/fame