Semi-Automatic Creation of Resources for Spoken Dialog Systems

Tatjana Scheffler, Roland Roller, and Norbert Reithinger*

DFKI GmbH, Projektbüro Berlin
Alt-Moabit 91c
10559 Berlin, Germany
{firstname.lastname}@dfki.de

Abstract. The increasing number of spoken dialog systems calls for efficient approaches for their development and testing. Our goal is the minimization of hand-crafted resources to maximize the portability of this evaluation environment across spoken dialog systems and domains. In this paper we discuss the user simulation technique which allows us to learn general user strategies from a new corpus. We present this corpus, the VOICE Awards human-machine dialog corpus, and show how it is used to semi-automatically extract the resources and knowledge bases necessary in spoken dialog systems, e.g., the ASR grammar, the dialog classifier, the templates for generation, etc.

1 Introduction

The more spoken dialog systems (SDSs) are put into practice in different domains, the more efficient methods for their development and deployment are urgently needed. The project SpeechEval aims to address this need in two ways: First, by investigating the use of dialog corpora in order to automatically or semi-automatically create the resources necessary for the construction of SDSs. And second, by learning general user behavior from the same corpora, and building a flexible user simulation which can be used to test the overall usability of SDSs during development or after deployment.

Automatic testing of dialog systems is attractive because of its efficiency and cost-effectiveness. However, previous work in this area concentrated on detailed tests of individual subcomponents of the SDS (such as the ASR). In order to judge the overall usability of a system, extended testing by human callers has been necessary – a step that is usually too costly to be undertaken during the prototype stage or repeatedly after changes to the deployed system. SpeechEval intends to fill this gap. Maximum modularity of the system architecture (see [1]) as well as the (semi-)automatic techniques for the creation of the underlying resources for the user simulation (in particular, domain knowledge and user strategies) allow SpeechEval to be easily portable across different SDSs.

* This research was funded by the IBB through the ProFIT framework, grant #10140648, and sponsored by the European Regional Development Fund.

In the following, we first discuss our approach to user simulation. Then we present the VOICE Awards corpus, a new dialog corpus which is the basis of our further work. The rest of the paper describes our finished and ongoing work in extracting knowledge bases for spoken dialog systems from corpora.

2 User Simulation

User simulation is used in the SDS literature for several purposes. First, for training the dialog manager of a SDS during reinforcement learning. In this case, the SDS with the learned strategy is the actual purpose of the research, whereas the user simulation is just a means to that end. Second, user simulation is used for evaluation or testing of the trained policies/dialog managers of the developed SDSs. The two types of purposes of user simulations may call for different methods. A user simulation may be used to test for general soundness of an SDS, specifically searching for errors in the design. In such a case, a random exploration may be called for [2]. A restricted random model may also perform well for learning [3].

In other cases, ideal users may be modelled so that reinforcement learning is able to learn good paths through the system's states to the goal [4]. Often (as in the previous example), a suitable user simulation is hand-crafted by the designer of the dialog system. A good overview of state-of-the-art user models for SDS training is given in [5].

Our goal is to as much as possible avoid hand-crafting the strategy (i.e., user simulation). An optimal strategy is not needed for our user simulation, neither is a random explorative strategy. Instead, the aim should be realistic user behavior. Our goal is to rapidly develop user simulations which show similar behavior (at least asymptotically) to human users in the same situations. The behavior of human callers of spoken dialog systems can be observed in our corpus, the VOICE Awards (VA) Corpus described below in section 3. We therefore define realistic user behavior in our case as user utterances that probabilistically match the ones represented in our corpus. Such probabilistic models are often used for evaluation of learned dialog managers [3].

Our current target approach is very close to the one proposed in [6] for an information state update system. At each state in the dialog, the user model choses the next action based on the transition probabilities observed in the corpus. Since some states have never or only rarely been seen in the corpus, we choose a vector of features as the representation of each dialog state. These features in our case include properties of the dialog history (such as the previous dialog act, the number of errors), the current user characteristics (expert vs. novice, for example), as well as other features such as the ASR confidence score. We estimate from the corpus the amount that each feature in the vector contributes to the choice of the next action. Thus, unseen states can be easily mapped onto the known state space as they lead to similar behavior as closely related seen states would.

The chosen next action is a dialog act type that must be enriched with content based on the goal and user characteristics. General heuristics are used to perform this operation of tying in the user simulation with the domain- and system-specific ontology.

3 A Human-Machine Dialog Corpus

For the development of a new spoken dialog system, the rules and knowledge bases must be specified by a human expert. As an alternative to hand-crafted systems, the strategies in a SDS may be learned automatically from available corpora. Much research has been done in this area recently, especially on dialog strategy optimization by reinforcement learning with (Partially Observable) Markov Decision Processes ((PO)MDPs) (see for example [7] for an overview). This approach works best for learning very specific decisions such as whether or not to ask a confirmation question or how many pieces of information to present to a user [8]. In addition, such systems must have access to large corpora of interactions with the particular system for training. Our goal, however, is to be able to interact with a new SDS in a new domain with little modification. In particular, in real applications we cannot assume the existence of a large specialized corpus of human-machine dialogs from that particular system or domain, as has been done in much of the previous literature. Therefore, we aim to learn general strategies of user behavior as well as other kinds of knowledge bases from a general dialog corpus.

Since we could not identify an appropriate human-machine dialog corpus in German, we are currently in the process of compiling and annotating the VOICE Awards (VA) corpus, which is based on the "VOICE Awards" contest. The annual competition "VOICE Awards"[1] is an evaluation of commercially deployed spoken dialog systems from the German speaking area. Since 2004, the best German SDSs are entered in this benchmarking evaluation, where they are tested by lay and expert users. We are constructing an annotated corpus of the available audio recordings from this competition, including the years 2005–2008 (recording of 2009 data is in progress).

The corpus represents a large breadth of dialog systems and constitutes a cut through the state-of-the-art in commercially deployed German SDSs. Altogether, there are more than 120 dialog systems from different domains in the corpus, with over 1500 dialogs. In each year of the competition, several lay users were asked to call the dialog systems to be tested and perform a given task in each of them. The task was pre-determined by the competition organizers according to the developers' system descriptions. After completing the task, the users filled out satisfaction surveys which comprised the bulk of the evaluation for the award. In addition, two experts interacted with each system and performed more intensive tests, specifically to judge the system's reaction to barge-ins, nonsensical input, etc. Table 1 contains a list of some of the domains represented by the dialog systems included in the VOICE Awards corpus.

[1] http://www.voiceaward.de/

Table 1. Some domains of SDSs included in the VOICE Awards corpus.

public transit schedule information
banking
hotel booking
flight info confirmation
phone provider customer service
movie ticket reservation
package tracking
product purchasing

Audio data for the VA corpus is available in separate .wav files for each dialog. The transcription of the corpus, using the open source Transcriber tool, is more than 50% complete. With the transcription, a rough segmentation into turns and dialog act segments is being performed. Since more fine-grained manual timing information is very difficult and time-consuming to obtain, it is planned to retrieve word-level timing by running a speech recognizer in forced alignment mode after the transcription is completed. As a basis of our statistical analyses, the entire corpus is currently being hand-annotated with several layers of information: (1) dialog acts, (2) sources of miscommunication, (3) repetitions, and (4) task success. Since the lack of space prohibits a detailed discussion, the annotation schemes are simply listed in table 2. We are using a modified tool from the NITE XML Toolkit (NXT) that has been adapted to our needs to perform these annotations in a single step.

We have performed an evaluation of the annotation scheme on part of the available data. Two annotators independently segmented and classified the data from 4 systems (69 dialogs). This test showed very good inter-annotator agreement of Cohen's $\kappa = 0.89$, as shown in table 3. The confusion matrix between the annotators further reveals that most mismatches concern only very few dialog act types (e.g., *alternative_question* and *instruction*), suggesting that revisiting the annotation scheme for these categories could further improve the agreement.

The result will be a large corpus of human-SDS-dialogs from many different domains, covering the entire breadth of the current state-of-the-art in commercially deployed German-language SDSs. In the next section, we describe how we are using this corpus and its annotations to derive resources for the rapid development of spoken dialog systems.

4 Corpus-Assisted Creation of SDS Resources

ASR Grammar In order to improve the coverage of an SDS's speech recognition, the recognizer's grammar must be augmented by adding both domain specific terminology as well as terms and phrases that are important in the scenario of spoken dialog systems in general. Different strategies will be used to extract both kinds of vocabulary from the VA Corpus as well as other sources.

Table 2. Hand-annotation schemes of the VOICE Awards corpus.

dialog acts	**errors**	**repetition**	**task_success**
hello	not_understand	repeat_prompt	task_completed
bye	misunderstand	repeat_answer	subtask_completed
thank	state_error		system_abort
sorry	bad_input		user_abort
open_question	no_input		escalated
request_info	self_correct		abort_subtask
alternative_question	system_command		other_failure
yes_no_question	other_error		
explicit_confirm			
implicit_confirm			
instruction			
repeat_please			
request_instruction			
provide_info			
accept			
reject			
noise			
other_da			

Table 3. Inter-annotator agreement for the dialog act (DA) dimension.

total # DAs	**agree on segmentation**	**agree on seg & type**
2375	1917	1740
	.81	.73
Chance agreement	0.16	(matching segments only)
Cohen's kappa	0.89	(matching segments only)

For the extraction of domain specific terminology, we have categorized the systems in the corpus along two dimensions into 24 topic domains (see table 1) and 8 interaction types (e.g., game, number entry, shopping, etc.). A simple chi-square test is used to determine whether a certain word i is significant for a domain j. Using a stop-word list of the 1000 most frequent terms in German, any word with a chi-square value greater than 3.84 is likely ($p < 0.05$) to be significant for the domain. Words which occurred less than 5 times in the corpus were discarded since the test is likely to be inaccurate. This method yielded very good results even when evaluated on a very small subcorpus. Table 4 shows the top 15 positively significant words for the banking domain, as computed on only 58 dialogs (3 systems) from the domain, and a similar amount of out-of-domain dialogs. The only false hits are words that are very suggestive of customer service SDSs in general ("möchten" / "would like"). These can be excluded by a second stop word list.

Table 4. Significant words in the banking domain.

term	**English**	χ^2	**term**	**English**	χ^2
Kontostand	account balance	56.6	Ziffer	digit	27.6
Kontonummer	account number	54.5	Geburtsdatum	birth date	26.0
möchten	would like	44.1	Hauptmenü	main menu	23.9
Umsätze	transactions	40.7	Bankleitzahl	routing number	22.9
Konto	account	40.2	Servicewunsch	service request	21.8
Überweisung	wire transfer	32.9	beträgt	amounts to	21.3
Cent	Cent	29.1	Gutschrift	credit	20.8
minus	negative	28.1			

We are extracting SDS-specific terminology (such as "customer id", "main menu", etc.) using the same methodology. All dialogs in the VA corpus are used as the positive subcorpus. For the negative examples, we plan to use text extracted from web pages representing a similar range of topics and domains as the VA corpus. This will ensure that only terminology specific to the medium of spoken dialog systems is marked significant by the chi-square test, and not other frequent content words such as domain-specific terms.

User Characteristics In order to perform realistic testing of dialog systems, the user simulation's behavior must be relatively varied. We aim to identify suitable user types from the VA corpus to model them in our user simulation. Broad distinctions such as expert vs. novice users are known from the literature, but aren't easily observable in the corpus, since by far most dialogs are by lay users. Thus, we instead try to distinguish objectively observable characteristics such as the user reaction time, number of barge-ins, etc. We will perform a clustering on each of these variables in order to obtain a "user properties vector" for each caller in the corpus. The obtained user characteristics then become part of the dialog state vector which determines the following user actions. This will account for the differences in behavior of different user types.

Dialog Act Segmentation and Classification Machine learning approaches are the standard approaches to the tasks of dialog act segmentation and classification. Good results can be obtained when the number of classes is not too high, although the quality of the ASR output has a large impact on the accuracy, as well. We distinguish 18 dialog act types (see table 2). Further, the types can be grouped into a flat hierarchy of broad categories such as "request" and "answer". Thus, even in cases where an incoming dialog act has been wrongly classified, SpeechEval's reply may still be appropriate if the misclassified type is of the same super-category.

Our segmentation and classification follows closely the method developed in the AMIDA project [9]. We use the WEKA toolkit to implement separate segmentation and dialog act classification learners. As opposed to this previous work, we use the learned classification modules within an online system. This

means that we cannot make use of dynamic features that require the knowledge of future assignments (as is done in the dialog act classifier). Each determined dialog act type is passed on immediately down the pipeline architecture and is acted upon in further modules.

As a first experiment we have trained a dialog act classifier using Sequential Minimal Optimization (SMO). The subcorpus for this experiment consists of 23 different spoken dialog systems with a total of 355 dialogs. For the dialog act classification, we divided the corpus into a specific training (20 systems, 298 dialogs, 9399 dialog acts) and test set (3 systems, 57 dialogs, 1680 dialog acts). The trained classifier showed very promising results, shown in table 5.

Table 5. SMO dialog act classification results.

total # DAs	**correct**	**incorrect**
1680	1396	284
	.83	.17
Cohen's kappa	0.78	

The kappa statistic of 0.78 shows relatively good agreement of the predicted dialog act with the hand-annotated one, especially considering the fact that human annotators only agree with κ=0.89 as shown above (table 3). The classification accuracy is promising since the test dialogs came from dialog systems and domains that were unseen in the training stage. This suggests that the final classifier trained on the full VA corpus will be very portable across systems and domains. In addition, we are currently using less than 25% of the available data, while the full amount of training data will increase the performance significantly. Furthermore there will be optimizations on the training data itself (for example, the treatment of overlapping segments) and the classification algorithm.

Concerning misclassifications, our evaluation so far has been very strict. The dialog act types can be grouped into a few broad super-categories (request, answer, etc.). Three super-categories are crucial for the interaction with a SDS: requests (*open_question*, *request_info*, *alternative_question*, *yes_no_question*), confirmations (*explicit_confirm*, *implicit_confirm*) and metacommunication (*instruction*, *repeat_please*). The confusion matrix shows that many misclassified instances were assigned to a dialog act class in the correct super-category. This means that the information is at least partially recoverable.

Compared to other recent work, the data reported here is very good. [10] report an accuracy of only 58.78% on dialog act classification of multi-party meeting data, even though they use a very similar feature set and a dialog act scheme of 15 types. This shows that system prompts in spoken dialog systems tend to be very schematic, and generalize well even across systems and domains. This validates our approach of extracting general system-independent knowledge bases for our user simulation.

User Utterance Templates Our corpus shows that by far most user utterances in our corpus consist of just one word. In an initial study, only 12% of the lay user's turns contained more than one word (number sequences such as ID or telephone numbers were excluded). For genuine more-word utterances, we are exploring a grammar induction technique in order to extract possible user utterance templates from our corpus.

5 Conclusion

In this paper we presented an approach to user simulation and spoken dialog system development that allows for very rapid prototyping. We introduced our new corpus of German human-machine dialogs, and discussed the ongoing annotation effort. This corpus constitutes the basis of our statistical methods for extracting both general and domain-dependent knowledge bases for SDSs. We discuss how many resources in a user simulation for SDSs, from the ASR grammar to dialog strategy, can be derived semi-automatically from the general dialog corpus or other supplementary corpora. This ensures easy portability of the user simulation across SDSs and domains and alleviates the need for large specialized corpora or expensive human evaluators.

References

1. Scheffler, T., Roller, R., Reithinger, N.: Speecheval – evaluating spoken dialog systems by user simulation. In: Proceedings of the 6th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Pasadena, CA (2009) 93–98
2. Alexandersson, J., Heisterkamp, P.: Some notes on the complexity of dialogues. In: Proceedings of the 1st Sigdial Workshop on Discourse and Dialogue. Volume 10., Hong Kong (2000) 160–169
3. Ai, H., Litman, T., Litman, D.: Comparing user simulation models for dialog strategy learning. In: Proceedings of NAACL/HLT 2007, Rochester, NY (2007) 1–4
4. López-Cózar, R., de la Torre, A., Segura, J., Rubio, A.: Assessment of dialog systems by means of a new simulation technique. Speech Communication **40** (2003) 387–407
5. Schatzmann, J., Weilhammer, K., Stuttle, M., Young, S.: A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. The Knowledge Engineering Review (2006)
6. Georgila, K., Henderson, J., Lemon, O.: Learning user simulations for information state update dialogue systems. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech), Lisbon, Portugal (2005)
7. Lemon, O., Pietquin, O.: Machine learning for spoken dialogue systems. In: Proceedings of Interspeech. (2007)
8. Rieser, V., Lemon, O.: Learning dialogue strategies for interactive database search. In: Proceedings of Interspeech. (2007)
9. AMIDA: Deliverable D5.2: Report on multimodal content abstraction. Technical report, DFKI GmbH (2007) chapter 4.
10. Germesin, S., Becker, T., Poller, P.: Determining latency for on-line dialog act classification. In: MLMI'08. (September 2008)