LET'S DISCOH: COLLECTING AN ANNOTATED OPEN CORPUS WITH DIALOGUE ACTS AND REWARD SIGNALS FOR NATURAL LANGUAGE HELPDESKS

G. Andreani[*#], *G. Di Fabbrizio*[+], *M. Gilbert*[+], *D. Gillick*[*], *D. Hakkani-Tür*[*] *and O. Lemon*[◇]

[*] ICSI, Berkeley, CA, USA
[#] Speech Village, Ascoli Piceno, Italy
[+] AT&T Labs, Inc. - Research, Florham Park, NJ, USA
[◇] HCRC, School of Informatics, Edinburgh University, UK

ABSTRACT

We motivate and explain the DISCOH project[1], which uses a publicly deployed spoken dialogue system for conference services to collect a richly annotated corpus of mixed-initiative human- machine spoken dialogues. System users are able to call a phone number and learn about a conference, including paper submission, program, venue, accommodation options and costs, etc. The collected corpus is (1) usable for training, evaluating and comparing statistical models, (2) naturally spoken and task oriented, (3) extendible / generalizable, (4) collected using state-of-the-art research and commercial technology, (5) freely available to researchers.

We explain the principles behind the dialogue context representations and reward signals collected by the system, as well as the overall system design, Call Types, and Call Flow. We also present results regarding the initial ASR models and spoken language understanding models. We expect the resulting corpora to be used in advanced dialogue research over the coming years.

Index Terms— Natural language interfaces, Speech communication, User Interfaces, Learning Systems

1. INTRODUCTION

Richly annotated data sets are a necessity for data-driven speech and language processing research. For example, recent work in stochastic approaches to dialogue management [1, 2, 3, 4], user simulation [5, 22, 6], context-sensitive automatic speech recognition (ASR), and stochastic parsing, and so on, all require large amounts of richly annotated dialogue data for training, testing, and evaluation of different models. The DISCOH project is an unprecedented initiative to enable the research community to collect natural language human/machine dialogues from automated conference helpdesk services ("DISCOH ": a Dialogue Service for Conference Help) across different organizations such as IEEE, ACL, etc. The DISCOH system is a general purpose goal-oriented, mixed-initiative, human- machine spoken dialogue system. It is designed to be highly portable and flexible across different conferences and workshops. System users are able to call a phone number and learn about a conference, including paper submission, program, venue, accommodation options and costs, etc. We have deployed the initial system for the IEEE/ACL SLT Workshop, which will take place in December 2006, and other deployments will follow. The first collected corpus will publicly and freely release the annotated spoken dialogues collected from this system for research purposes.

Future releases will extend the initial corpus, validate the annotations, improve system performance, and generalize the dialogue strategy for the help-desk task. The original design also includes real-time capabilities for reinforcement learning policy optimization.

Given that data-driven approaches are getting more popular for many speech and language processing applications, we believe that such a corpus annotated with system prompts, user utterance transcriptions, user intentions, overall task success, etc., will be a useful resource for researchers in dialogue management, spoken language understanding, automatic speech recognition and other related tasks. These annotations can also be extended with user emotion tags, disfluencies, syntactic and semantic parses, etc. in the future.

While no international standards for dialogue context representation yet exist, there has been a recent international workshop (involving the TALK and AMI European projects[2], and the W3C) on this issue [7], and the TALK project has developed a project-wide standard for dialogue context representations and reward annotations [8, 9], a variant of which has been adopted for DISCOH (see section 3.1).

1.1. Related work

Even though there are multiple human-human conversational corpora available such as the Switchboard, Monroe, and ICSI Meeting corpora, the nature of human-machine conversations in the framework of goal-oriented spoken dialogue systems is significantly different. For example, in the latter case, the user's intentions are usually uttered in a more direct and concise way. Furthermore the dialogue structure is different, and these types of human-human dialogues have not proven particularly useful for developing sophisticated goal-oriented dialogue management systems.

In the 1990s, the DARPA-funded Airline Travel Information System (ATIS) [10] and Communicator [11, 12, 13] projects resulted in collection and annotation of human-machine spoken dialogues from the travel planning and information domain. These corpora were used (and are still being used since there are no other alternatives) by many researchers and have led to the next generation of technologies for speech and language processing, where machine learning approaches are more widely used even for dialogue management and natural language generation. However, there are some problems with the COMMUNICATOR corpora when we try to use them for statistical approaches to dialogue systems (see [14]):

- the data sets were not large enough (less than 3,000 dialogues),

Authors are in alphabetical order. This work was partially funded by NSF, IIS 0624389, and the TALK project, EC IST 507802. We thank AT&T Labs, Inc. - Research for providing equipment, phone lines, research and development resources. The views herein are those of the authors and do not reflect the views of the funding agencies.

[1] www.discoh.org

[2] www.talk-project.org and www.ami-project.org

- some important data types (e.g., ASR confidence scores) were omitted.
- no representation of dialogue contexts or speech-act history was used
- user inputs were not labelled with speech acts (the DATE scheme [15] was only used for system outputs).

Recent work extending the original COMMUNICATOR corpora [14] has tried to remedy some of these problems. However, vital information is still missing and cannot be obtained (e.g. ASR confidence scores).

The COMMUNICATOR corpora are very useful in that task success was recorded, but this is sadly missing from later dialogue data collections, for example the AMITIES [16] data collection did not record task success / failure, so cannot be used for many machine learning approaches for dialogue management, such as reinforcement learning.

The W99 spoken dialogue system [23] developed in AT&T for the ASRU99 workshop and the VoiceIF [24] created for the 2000 edition of the AT&T Innovation Forum Workshop are more similar in terms of task and structure to the DISCOH system. Unfortunately, these corpora are all proprietary and are thus unavailable for general use such as benchmarking, and have only been beneficial to limited communities.

2. SYSTEM OVERVIEW

In developing the system we used the AT&T VoiceTone® Spoken Dialogue System tools. Below, we briefly describe this system, ASR and SLU models, corpus components, and annotation types.

2.1. The AT&T VoiceTone® Spoken Dialogue System

The AT&T VoiceTone® Spoken Dialogue System [17] is a part of the AT&T VoiceTone® retail branded services offered to many of AT&T 's business customers. It has been designed to increase the level of customer care automation and provides a better user experience while reducing call center operation costs. It provides network grade scalable services with large vocabulary speech recognition [18], spoken language understanding (SLU) [19] and mixed-inititative dialogue management [20]. The users are greeted with the open-ended *"I am an automated assistant. You can speak naturally to me. How may I help you?"* prompt encouraging them to utter their requests in natural language. The system then tries to identify the customer's intention (call-type) and associated named entities (NE) using a natural language understanding component. For call-type classification and NE extraction, statistical classifiers are used. As a call classification example, consider the utterance *"yes um I want to know if my paper was accepted"*, in our conference help-desk domain. Assuming that the utterance is recognized correctly, the corresponding intent or the call-type would be Request(paper_status). The action would be to prompt for the paper identification number and provide the current status of the paper submission (e.g. incomplete submission, review in progress, accepted, not accepted). In the event the system is unable to understand the caller with high enough confidence, the conversation usually proceeds with either a re-prompt or a confirmation prompt. Overall the dialogues are not rigidly structured, since users are free to ask for any information at any time. Thus we do not expect the corpus to be constrained by the initial system design. Here is an example dialogue with DISCOH :

```
Sys: I'm an automated assistant. You can speak
     naturally to me. How may I help you?
Usr: I wanna know about the the paper submission
     deadlines
Sys: Do you want to know about the current deadlines?
Usr: Yes I do.
Sys: The paper submission deadline is Friday, July 21, ...
     Is there anything else I can help you with today?
Usr: How do I register for the workshop?
Sys: Registration can only be done online at the
     workshop website...
```

Training Data	No. of utt./sent	No. of words	Vocabulary Size
W99	11,275	56,436	1,541
W99+GD	20,786	122,953	1,543
W99+GD+WEB	21,012	128,184	2,552
W99+GD+WEB+MD	21,479	131,722	2,561
test	213	1,157	187

Table 1. Characteristics of the data sets used in the experiments.

2.2. Speech recognition and understanding models

One of the largest obstacles when building a system for a new domain is the lack of annotated data for training the statistical models. The speech recognizer acoustic model for DISCOH is trained using telephone speech collected from previous AT&T VoiceTone® applications. The initial speech recognizer language model is built using the data set (W99) collected from a similar previous spoken dialogue system [23, 24], and was improved using data collected from the web pages of the conference (WEB), artificially generated utterances using the semantic parses and predicates and arguments of W99 and WEB data using conversational templates learned from previous applications [25] (GD), and a small set of utterances that are estimated to be seen in the domain (MD). We have tested these models using a test set of *all* utterances collected from the first deployment of the system. Table 1 lists the properties of these training and test data, and Figure 1 shows the run-time in real time versus ASR word accuracy on the test set using models trained on these data sets. As we added more data, the test set performance improved at all operating points.

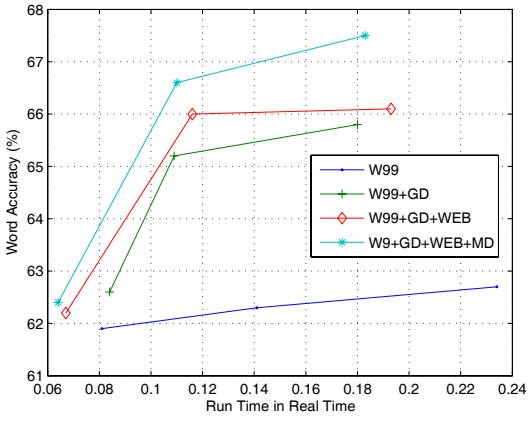

Fig. 1. Run time in real time curves using a combination of various training data sets. As we added more data, the test set performance improved at all operating points.

The set of user intentions (Call Types) for spoken language un-

Training Data	No. examples	Ref		ASR	
		Error	F	Error	F
sW99	4,877	38.9%	62.7%	50.4%	52.8%
sW99+MLD	5,213	39.3%	63.5%	49.5%	53.5%

Table 2. The SLU error rate and F-measure (F) on reference transcriptions and ASR output.

derstanding were designed using semantic parses of the previously collected data sets [26], and were improved manually, resulting in 54 categories. Spoken language understanding models were trained using a manually labeled subset of W99 utterances (sW99), and an additional subset of manually created and labeled utterances (MLD) to incorporate new user intentions for the IEEE/ACL SLT workshop. The SLU error rates on the manual transcriptions and ASR output of the test set are shown in Table 2. The test set included user intentions that were not seen in the sW99 and MLD data sets, and we expect both the ASR and SLU performance to improve as we collect more data from the deployment.

2.3. User Interface Design

Spoken natural language user interface design faces the challenge of many contradicting requirements when offered to a large population of heterogeneous users. A general principle is to increase the likelihood that users can successfully complete their task with the minimum amount of effort and confusion [12]. However, this has to be balanced with the overwhelming amount of information available on the workshop web site, which must be rendered over the limited telephony channel [27] and has to be effective with the different behaviors of naïve and expert users. To achieve these goals, we first divided the workshop information published on the web site into 30 dialogue categories, and hierarchically organized them in three levels of increasing information detail. For example, there are basic-level contents such as general workshop and hotel information, with the corresponding call types:

- Request_info(workshop(general))
- Request_info(hotel(general))

and more detailed content has the following call types:

- Request_info(workshop(schedule))
- Request_info(workshop(invited_talks))
- Request_info(workshop(social_program))
- Request_info(workshop(technical_program))

This was done keeping in mind a simple navigation schema that allows users to consistently require more details on a specific topic when needed. Secondly, the web verbiage was rewritten in a more natural and crisp formulation, closer to a colloquial style. Thirdly, the prompt transcripts were refined adding greetings, contextual help, and confirmation requests to give a chance to validate low confidence results from the SLU component. Finally, completing the prompt design phase, we selected a female voice talent to record the prompts in a professional audio studio. Prompts are the most visible part of the application and contribute substantially to the overall user experience. We instructed the voice talent to provide a cheerful and trustworthy personality, slightly enthusiastic about the whole workshop event and the venue. The reading pace was somewhat faster than normal to give a lively, energetic involvement.

The Call Flow proceeds from an open top-level *"How can I help you?"* prompt and then allows the user to freely request information at any level of detail. After information at a general level is given to the user, they are then offered more detailed options for that topic (e.g. *"I can tell you about the workshop location"*), but they may also switch to a different topic, again at any level of detail.

3. COMPONENTS OF THE CORPUS

The resulting (anonymized) corpus will at least contain the data listed below which will be generated from system outputs and internal logs. We invite the research community to extend the annotations of the collected data, for example in terms of prosody, turn taking, alignment, and so on.

- Audio files of user utterances
- Best hypothesis of the ASR
- ASR confidence scores (whole utterance)
- n-best hypotheses of the ASR (with confidence scores)
- ASR word lattices
- System prompts
- Call-type hypotheses and/or dialogue acts from the SLU with their confidence scores
- Named entities and/or filled/confirmed slots from the SLU
- Dialogue context (e.g. speech act history), (see section 3.1)
- Task context (e.g. named entities and/or filled/confirmed slots)
- Reward signals (see section 3.2)
- System agenda (e.g., what the system plans to say next)
- Dialogue length and number of errors
- Dialogue design/policy.

The corpus represents the dialogues in a hierarchical XML structure. Each dialogue consists of a sequence of turns, which includes a system prompt and a user utterance, and the dialogue context and reward after each utterance. The context and reward annotations are crucial for training and testing new approaches in stochastic dialogue management, parsing, and context-sensitive speech recognition.

3.1. Representing dialogue context

In [8] a method for representing dialogue contexts is proposed, based on an extension of the DATE annotation scheme [15], and has been used when training the learned dialogue policies of [4, 21] and user simulations of [22]. The basic idea is to log and/or annotate features of the dialogue context after each system and user move. The dialogue context contains features such as turn, speech-act, user intentions, speech-act history, filled slots (named entities), confirmed slots, etc., see [4] for examples.

3.2. Collecting reward signals

We have implemented a new element of the Call Flow to automate the collection of reward signals from the user. This is a sequence of questions that the user is asked upon closing the system, which will help to determine:

- Perceived task completion ("Did you get all the information that you wanted?")
- Future use ("Would you use the system in the future?")
- Ease of use ("Did you find the system easy to use?")

Such reward signals are critical for training and testing stochastic dialogue system components using various types of reinforcement learning [1, 2, 3, 4]. The regression analysis of [12] has established that they are correlated with overall user satisfaction. The current system collects final reward, and additional annotations could be developed for various types of interim reward.

3.3. Potential Manual Annotation Types

To be useful for future spoken dialogue systems research the corpus should include the manual transcription of user utterances, manually annotated user utterance call-types and objective task success / failure information.

We aim to collect 600-800 dialogues with the initial deployment, and more dialogues with the system deployed / improved for future conferences / workshops. The annotations can also be extended by adding user gender, age, emotion, accent of the speaker, disfluencies, syntactic and semantic parse of the user utterances to be useful for multiple research purposes.

4. CONCLUSION

The contribution of this project is a mixed-initiative human- machine spoken dialogue corpus, which is: 1) useful for training, evaluating and comparing statistical models, 2) naturally spoken, 3) extendible / generalizable, 4) collected using state-of-the-art commerical technology, 5) freely available to researchers.

We explained the motivations behind the DISCOH project, the representations of dialogue context and reward, and presented a system overview. We explained the principles behind the dialogue context representations and reward signals collected by the system, as well as the overall system design, Call Types, and Call Flow. We also presented results regarding the system's initial ASR models and spoken language understanding models. All collected data will be freely released to the research community.

5. REFERENCES

[1] E. Levin and R. Pieraccini, "A stochastic model of computer-human interaction for learning dialogue strategies," in *Proc. Eurospeech*, 1997, pp. 1883–1886.

[2] D. Litman, M. Kearns, S. Singh, and M. Walker, "Automatic optimization of dialogue management," in *Proc. COLING*, 2000.

[3] K. Scheffler and S. Young, "Corpus-based dialogue simulation for automatic strategy learning and evaluation," in *Proc. NAACL Workshop on Adaptation in Dialogue Systems*, 2001.

[4] J. Henderson, O. Lemon, and K. Georgila, "Hybrid Reinforcement/Supervised Learning for Dialogue Policies from COMMUNICATOR data," in *IJCAI workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 2005.

[5] J. Schatzmann, M. Stuttle, K. Weilhammer, and S. Young, "Effects of the user model on simulation-based learning of dialogue strategies," in *ASRU* 2005.

[6] V. Rieser and O. Lemon, "Cluster-based user simulations for learning dialogue strategies," in *Interspeech/ICSLP*, 2006, p. (to appear).

[7] O. Lemon, Ed., *TALK/AMI/W3C workshop on Standards for Multimodal Dialogue Context*. HCRC, Edinburgh University, 2005.

[8] O. Lemon, K. Georgila, J. Henderson, M. Gabsdil, I. Meza-Ruiz, and S. Young, "D4.1: Integration of Learning and Adaptivity with the ISU approach," Tech. Rep., TALK Project, 2005.

[9] V. Rieser, I. Kruijff-Korbayová, and O. Lemon, "A corpus collection and annotation framework for learning multimodal clarification strategies," in *6th SIGdial Workshop*, 2005.

[10] S. Seneff, L. Hirschman, and V. W. Zue, "Interactive problem solving and dialogue in the ATIS domain," in *Proc. Fourth DARPA Speech and Natural Language Workshop*, 1991.

[11] M. Walker, R. Passonneau, and J. Boland, "Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems," in Proc. *ACL*, 2001

[12] M. Walker, C. Kamm, and D. Litman., "Towards Developing General Models of Usability with PARADISE," *Natural Language Engineering*, vol. 6, no. 3, 2000.

[13] M. Walker, A. Rudnicky, Aberdeen J., E. Bratt, Garofolo J., H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, R. Prasad, Roukos S., G. Sanders, S. Seneff, D. Stallard, and S. Whittaker, " DARPA Communicator Evaluation: Progress from 2000 to 2001," in *Proc. ICSLP*, 2002.

[14] K. Georgila, O. Lemon, and J. Henderson, "Automatic annotation of COMMUNICATOR dialogue data for learning dialogue strategies and user simulations," in Proc9th SEMDIAL, *DIALOR)*, 2005.

[15] M. Walker and R. Passonneau, "DATE: A dialogue act tagging scheme for evaluation of spoken dialogue systems," in *Proc. Human Language Technology,* 2001.

[16] H. Hardy, K. Baker, L. Devillers, L. Lamel, S. Rosset, T. Strzalkowski, C. Ursu, and N. Webb, "Multi-layer dialogue annotation for automated multilingual customer service," in *Proc. ISLE Workshop on Dialogue Tagging for Multi-Modal Human Computer Interaction*, 2002.

[17] M. Gilbert, J. Wilpon, B. Stern, and G. Di Fabbrizio, "Intelligent virtual agents for contact center automation.," in *IEEE Signal Processing Magazine*. 2005.

[18] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tur, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, and M. Saraclar, "The AT&T Watson speech recognizer," in *ICASSP 2005*

[19] N. Gupta, G. Tur, D. Hakkani-Tur, S. Bangalore, G. Riccardi, and M Rahim, "The AT&T spoken language understanding system," *IEEE Trans. on Speech and Audio Processing.*, 2006.

[20] G. Di Fabbrizio and C. Lewis, "Florence: a dialogue manager framework for spoken dialogue systems," in Proc. *ICSLP*, 2004

[21] M. Frampton and O. Lemon, "Learning more effective dialogue strategies using limited dialogue move features," in Proc. *ACL*, 2006.

[22] K. Georgila, J. Henderson, and O. Lemon, "Learning User Simulations for Information State Update Dialogue Systems," in *Interspeech*, 2005.

[23] M. Rahim, R. Pieraccini, W. Eckert, E. Levin, G. Di Fabbrizio, G. Riccardi, C. Kamm and S. Narayanan, "A Spoken Dialog System for Conference/Workshop Services," in *ICSLP*, 2000.

[24] M. Rahim, G. Di Fabbrizio, C. Kamm, M. Walker, A. Pokrovsky, P. Ruscitti, E. Levin, S. Lee, A. Syrdal and K. Schlosser, "VoiceIF: A Mixed-Initiative Spoken Dialogue System for AT&T Conference Services," in *Eurospeech*, 2001.

[25] D. Hakkani-Tür and M. Gilbert, "Bootstrapping Language Models for Spoken Dialog Systems from the World Wide Web," in *ICASSP 2006*

[26] G. Tur, D. Hakkani-Tür, and A. Chotimongkol, "Semi-Supervised Learning for Spoken Language Understanding using Semantic Role Labeling," in ASRU, 2005.

[27] Cowan, N., Morey, C.C., and Chen, Z., "The legend of the magical number seven" in S. Della Sala (Ed.), *Tall tales about the brain: Things we think we know about the mind, but ain't so.*, Oxford University Press, (in press)