IEEE International Conference on Bioinformatics and Biomedicine (BIBM09) Workshop on NLP Approaches for Unmet Information Needs in Health Care. Washington DC, USA, Nov. 1-4, 2009.

Speech Recognition in a Dialog System for Patient Health Monitoring

Lucian Galescu
Institute for Human and Machine Cognition
Pensacola, FL
e-mail: lgalescu@ihmc.us

James Allen, George Ferguson, Jill Quinn, Mary Swift
University of Rochester
Rochester, NY

***Abstract*—We describe CARDIAC, a prototype for an intelligent conversational assistant that provides health monitoring for chronic heart failure patients. CARDIAC supports user initiative through its ability to understand natural language and connect it to intention recognition. The spoken language interface allows patients to interact with CARDIAC without special training. We present speech recognition results obtained during an evaluation with fourteen chronic heart failure patients.**

Keywords - dialog systems; speech recognition; natural language processing; data acquisition; patient self-management; chronic heart failure.

I. INTRODUCTION

Research suggests that the quality of health care is greatly affected by the support patients receive in the home, whether by family caregivers or from nurse practitioners. For example, readmission rates for patients who have experienced congestive heart failure can be significantly reduced with close home monitoring of patients by a nurse practitioner [1]. Unfortunately there is not enough medical personnel available to provide such close monitoring, and even if there was the cost would be prohibitive. New technologies have been developed to permit health monitoring and improve self-management. However, numerous studies have shown that, although the use of telemonitoring technology can lead to better outcomes, and is accepted by a significant number of patients going through clinical studies, significant technical, behavioral, managerial, and financial barriers exist ([2], [3]). Some of the monitoring devices used are invasive; patients may view the technology as intrusive or ill-fitting their lifestyle, most of it is expensive and, finally, access and technical literacy are not yet solved problems. Even specially designed devices touted as easy-to-use are considered by patients as "bothersome, complex, and too lengthy an intervention" [4]. As a consequence, the current assessment is that, besides the intrinsic benefit to the patient, convenience is the most important factor for technology adoption; in particular, usage is more successful if the intervention could be delivered on technology consumers use every day for other purposes [3]. Finally, monitoring devices allow limited input for the patient's subjective health status report (e.g., chest pain, shortness of breath).

Intelligent dialogue systems promise to be an increasingly important tool for collecting information from patients under self-managed care [5]. Not only are dialogue systems more intuitive to use, but they are potentially usable by anyone with a telephone. So far, however, few such systems have been deployed, and most use very restrictive dialogue models based on scripts and finite state machines, and have very shallow knowledge representations. While such systems can be relatively easily built for narrow applications, they lack scalability, tailorability and adaptability (see [5] for a review).

Our research is aimed at developing a plan-based conversational assistant to help chronic care patients look after themselves and provide comprehensive health care monitoring. Our current focus is on chronic heart failure (HF) patients; because their signs and symptoms can be assessed remotely and deterioration can be quickly detected and addressed, HF appears to be an ideal case for testing whether an automated dialogue system would be an effective intervention.

In this paper we first give an overview of CARDIAC (Computer Assistant for Robust Dialogue Interaction and Care), a prototype of an intelligent conversational assistant that provides health monitoring for HF patients. CARDIAC's objective is to conduct regular "checkup" interviews with patients to collect information relevant to their condition. The target population for CARDIAC is patients who are at home following specific self-care guidelines. The CARDIAC checkup is designed to obtain the information required by the self-care guidelines including both objective (e.g., weight) and subjective (e.g., pain) aspects of their condition. The system can also take advantage of other sources of information (such as a network-connected scale or a Personal Health Record) and use them effectively to tailor the checkup interview without additional programming. The system's conversational interface is intuitive and easy to use, a benefit that may encourage patients to report their information more often.

In the second part of the paper we give some speech recognition results obtained during a recently completed feasibility evaluation of CARDIAC. Speech recognition accuracy is likely to be a significant factor in the system's usability, as well as in its ability to accurately comprehend the information reported. Our results suggest that the system would be usable by a significant proportion of the target population.

II. OVERVIEW OF THE SYSTEM

CARDIAC is an agent-based spoken dialogue system that conducts health monitoring interviews with chronic heart failure patients using natural language. CARDIAC interprets the patient information and uses it to update its user models. The following dialogue excerpt illustrates the sort of interaction that the system supports:

SYS: Do you know your weight
USR: YES TWO SIXTY
SYS: Did you say 260
USR: YES
SYS: Do you have shortness of breath today
USR: YES
SYS: How severe is the shortness of breath
USR: A LITTLE MORE THAN NORMAL

©2009 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

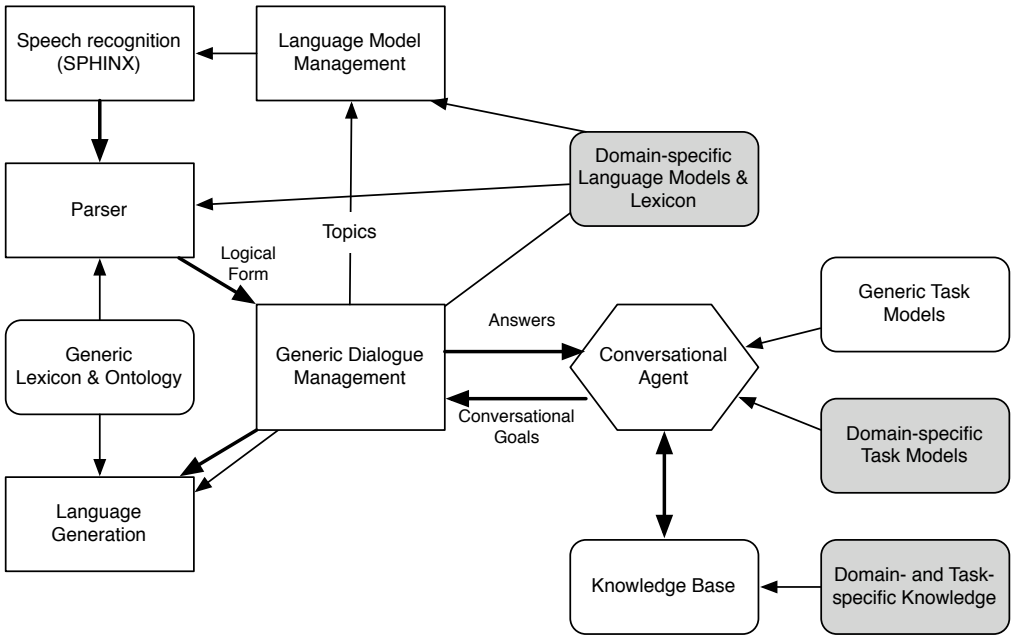

Figure 1. CARDIAC System Architecture

The CARDIAC system is built as an instantiation of the TRIPS generic dialogue architecture [6], which provides broad coverage domain-general parsing, generation, and discourse interpretation facilities. To adapt this generic system to a specific domain, we need to define a task model that the conversational agent uses to drive the interview process without needing to manage the details of natural conversation. Besides the task model, other domain specific components needed are the domain specific lexicon (in this case, medication names) and statistical language models used in speech recognition and generation. In addition, we need ontology mapping rules that translate meanings expressed in the TRIPS generic ontology to the specific ontology used for reasoning in this domain.

The system architecture is given in Fig. 1, with the generic components in white and the domain specific components in grey. To obtain the most accurate speech recognition (SR) we can get, we use dynamic language models based on the current topic. The topic is determined when the system asks a specific question, at which point the SR engine switches the language model accordingly. The best SR hypothesis is used as input to the parser. The generic dialogue manager (DM) receives the parser output and performs contextual interpretation including reference resolution and interpreting elliptical answers. The DM and the Conversational Agent (CA) interact to perform goal-driven dialogue management; i.e., the system is driven by reasoning about goals to acquire knowledge. Each of these components is described below in more detail; for further particulars, the reader is referred to [7].

A. Speech Recognition

The speech recognition component is based on CMU's Sphinx 3 engine [8], enhanced with the capability to load and/or switch language models dynamically, at run-time. The acoustic models we use are generic, gender-independent, trained on broadcast news (HUB-4) data; no adaptation for conversational style or for the application domain is performed.

Because no domain-specific textual data was available to train statistical language models for speech recognition, we used the technique described in [9]. In short, this technique allows us to build quickly language models using a process of collecting domain-specific utterances, then generalizing them (via synonyms and rephrasing) into a finite-state grammar from which we generate, by random walking, a large corpus of sentences; statistical language models are trained on this synthetic corpus. For example, the utterance *My back hurts a little* may be transformed in the template *<body-part> hurts [<degree>]*, where *<body-part>* may stand for *my back*, *my head*, *the left leg*, etc., and the optional *<degree>* phrase might stand for *a little*, *a lot*, *some*, *more than yesterday*, and so on. As usual, non-terminals can be defined in terms of other non-terminals. Although these grammars involve somewhat time-intensive manual labor, it is not too difficult to develop extensive such grammars to cover a lot more n-grams than could be obtained at greater expense by conducting Wizard-of-Oz experiments to collect "real" data from potential users of the system. While the synthetic corpus generated can only yield a crude approximation of the true (but unknown) distribution of word/n-grams for the application domain, in our experience the approach works quite well; for medium-sized vocabularies, word error rates in the 20-30% range would be possible [9].

We use statistical language models because they offer a lot more flexibility than fixed grammars, where the user has to phrase their answer in a (small) number of pre-specified ways, and a lot more naturalness than guided systems, where the system tells the user what options are available and the user can only select one of them (cf. [10]).

For the CARDIAC system we trained nine topic-specific grammars and language models; topics were based on broad categories, such as symptoms (e.g., fatigue, edema, etc.), medication use, diet and exercise. Each topic-specific grammar and language model accounts for potential user speech related to the specific topic. Additional topic-independent grammars and language models were built to account for generic

conversational (social) speech, and for patient speech that doesn't fall under any of the nine topics (e.g., descriptions of how they feel). Because we expect that in reality the system would likely be faced with off-topic over-answers (e.g., *No swelling, but I was very short of breath yesterday after my daily exercise*), in the final system all the topic-specific and topic-independent models were linearly interpolated. Linear interpolation is a commonly used method of combining language models whereby the new LM is computed as a convex combination of the component LMs [11]. Thus, each of the nine topic-specific interpolated LMs was a mixture of all the component LMs, with most of the emphasis on the LM for the main topic; this way we traded off some specificity for greater robustness. A tenth, topic-independent model, is used as a fall-back, when no topic-specific LM is available; it is also computed as a linear interpolation of the individual models, with all topics being equally weighted, and more emphasis put on the topic-independent components.

During the development of the system we collected a small amount of data from simulated dialogues with nursing students. We used the data to enhance the grammars' coverage and to tune the interpolation weights.

For the current version of the system, all LMs are trigram models with Witten-Bell discounting, built with the SRILM toolkit [12]. The vocabulary size for all LMs is 1647; each LM includes about 54K bigrams and 174K trigrams.

B. Parsing

The best SR hypothesis is fed to a deep, broad coverage parser [13]. The parser uses a general grammar and a domain-independent semantic lexicon augmented by domain-specific words, which in this system are medication names, to produce a semantic representation of the utterance. In order to deal with speech recognition errors, the parser is designed to find the most semantically coherent sequence of sentence fragments when it cannot construct a complete analysis.

C. Dialogue Management

The DM receives the parser output and performs contextual interpretation including reference resolution, surface speech act recognition, and ellipsis interpretation. It generates hypotheses about the user's intended meaning, which are then passed to the conversational agent (CA) for evaluation as to whether the hypothesis would be a coherent statement. The DM then either chooses one hypothesis as the final interpretation and passes it to the conversational agent, or determines that it didn't understand the user's response. Often, because of recognition errors, some part of the utterance is not interpretable. If the DM can identify a fragment that satisfies the current goal (determined by the CA), then it ignores the fragments that cannot be interpreted. While the conversational agent determines the current goal for the interview (e.g., find out the patient's weight), the DM manages the details of the actual dialogue. Thus, if the user's answer is not understood, it will re-ask the question, possibly also giving hints such as asking the patient to speak more simply. In addition, when dealing with answers that could easily be misrecognized, such as answers involving numbers, the DM may initiate a confirmation subdialogue to verify the answer. If the system fails to understand the patient a number of times, the DM abandons its discourse-level goal and notifies the conversational agent to abandon the current task-level goal. The DM also applies ontology mapping rules to convert the generic semantic representation output by the parser into the domain-specific representation whenever it communicates with the CA.

D. Conversational Agent

The CA is responsible for the system's overall behavior. This includes the following responsibilities:

- It is driven by a declarative model of the task(s) that the system can perform. Based on these tasks, it manages the goals that drive the system's behavior.
- During execution, it maintains the knowledge base that stores what the system knows about the current situation.
- It interacts with the language understanding components to support the interpretation of user utterances, and with the natural language generation components to produce system utterances.
- Finally, it responds reactively to changes in its environment, including utterances from the user and other sources of information.

This section briefly elaborates on each of these aspects.

The CA is based on a domain-independent engine that executes tasks specified declaratively. Building on a long tradition in AI ([14] - [16]), these tasks generally consist of goals that need to be achieved. Other tasks are invoked to achieve the sub-goals, eventually bottoming out in so-called "primitive" goals that can be achieved by built-in mechanisms. The CA execution engine includes predefined mechanisms for sequential and conditional tasks, and is easily extended.

Significantly, the CA can introspect on its execution. It can inspect the set of active and pending goals, the goal-subgoal hierarchy, the tasks chosen for active goals, and the state of those tasks. This ability is crucial for collaborative systems. First, it enables intention recognition for interpreting language and identifying discourse phenomena such as topic shifts, corrections, and perhaps misunderstandings. It is also necessary to support explicit discussion of the problem-solving process, where the participants explicitly discuss what goals to pursue or whether to abandon them, for example.

The CA defines a number of abstract tasks that form the basis for the specification of the system's behavior. Because the system is designed from the outset to support collaboration, these predefined tasks include tasks for accomplishing knowledge goals. Our representation of the system's knowledge is loosely based on standard models of knowledge and belief and their relationship to language and action (e.g., [17] - [19]). Briefly, what this means is that goals may involve the system's knowing something, or that the system agree with the user about something, or it could involve getting the user to do something. Standard tasks accomplish these knowledge goals by inspecting the system's beliefs regarding the content of the goal (including its beliefs about the user's beliefs) and initiating conversational acts whose effects achieve the goal. For example, to agree whether the patient is experiencing any swelling, the system does not know the answer but believes that the user does (because swelling is a type of symptom that the system knows corresponds to an internal state of the user). This eventually results in the system *asking* the user. On the

other hand, in agreeing about the user's weight, the system might already know the value (perhaps from an automated bathroom scale). This would result in the system *informing* the user as a way of reaching agreement.

The CA supports interpretation of user utterances by a combination of intention recognition and knowledge-based interpretation. Even with topic-specific language modeling for speech recognition and deep, semantic parsing, speech recognition errors can lead to grammatical but incorrect interpretations of user input. The system uses its knowledge of what it is doing (the goal hierarchy that led to the current question, if we're interpreting an answer) and the reasoning capabilities of the knowledge base to validate interpretation hypotheses. Crucially, the interpretation process is applied equally to the user's answers to the system's questions and to statements that the user makes on their own, supporting the user-initiated style of interaction described below.

The agent's reactive behavior and explicit goal state and task model allow it to support over-answering and, more generally, user initiative. In the case of over-answering, the content of a user's answer is taken from its linguistic interpretation, not from the fact that it is an answer to the question. The linguistic interpretation must make sense as an answer (as described above), but once committed, the entire content of the utterance becomes system knowledge. If that knowledge is the subject of subsequent goals, the CA execution engine will automatically use the knowledge and behave appropriately (for example, not asking a question for which it already knows the answer).

The system can also interpret user utterances that are not responses to questions. Again, the interpretation of the utterance becomes system knowledge (and this time the interpretation must make sense relative to the task being performed, a form of intention recognition). This symmetric treatment of questions and statements means that the system automatically supports mixed-initiative interaction with no task- or domain-specific programming. Patients using our prototype for the first time typically let the system ask the questions and answer quite specifically. If they were to become more familiar with the system's goals (i.e., what the system needs from them), they could make the checkup interview even simpler by describing how they're doing from the outset. The system would interpret all this, and follow up only on things that were not mentioned or are not inferable from what the system knows already (from environmental sensors, PHR, etc.).

III. EVALUATION

We have recently conducted a feasibility evaluation of CARDIAC with actual chronic heart failure patients in a cardiology practice. The focus of the evaluation was whether the system could identify with high accuracy the information the patient provides in the interview. The evaluation required comparing CARDIAC's analysis of patient responses with that of nurse practitioners. To this end, we created a web interface where nurse practitioners can listen to the audio of the system interviews and record their interpretation of patient responses. A detailed analysis is in progress. Preliminary observations suggest that the system can perform the CHF self-care checkup with reasonable accuracy, and that most patients believe the system is easy to use and would be helpful to them in managing their care.

Phase	Sentences	Words	Unique words
P1	191	639	168
P2	615	2087	423
C	282	755	164
Total	1088	3481	512

Table 1. Sentence and word counts for the collected data.

In the following we describe the data collected and discuss evaluation results for the speech recognition component.

A. Data

So far, we collected data from fourteen CHF patients, eight male and six female. A session with a patient consisted of two distinct phases:

- a practice phase, designed to familiarize the subject with the capabilities of the speech recognizer. During this phase, the system asked 20 generic questions, and the subject was free to answer as many times as they wished. This phase was further subdivided into two sub-phases:

  (P1) for the first five questions, the SR hypothesis was visible on the screen;

  (P2) for the other questions, there was no feedback.

- a checkup phase (C), during which the system conducted an interview with the patient.

During the practice phase only the topic-independent LM was used; during the checkup phase, we used dynamic, topic-specific LMs.

Table 1 shows summary statistics about the data.

B. Speech Recognition Results

Speech recognition performance is shown in Figures 2 and 3. The first graph shows the percentage of correct sentences (SC) recognized. The second graph displays word error rate (WER) performance. Along with per-speaker results, for the checkup phase we also display per-topic results (labeled "C/t").

Clearly, when the speech recognizer gets very poor performance for a patient, it is unlikely that that patient would benefit from this technology. We estimate that the threshold is around 20-25% WER; more than that would practically guarantee that the system will either fail to collect the information needed, or, worse, the accuracy of the collected information would be severely compromised. Nine of the fourteen participants had word error rates for the checkup phase under 21%. The other five (two male and three female) subjects had error rates above 30%; for these patients, it is likely that incremental improvements on the current technology will not be enough to enable them to use our system.

A significant number of errors occur simply due to vocabulary coverage lapses in our LMs. When the speech recognizer encounters an out-of-vocabulary (OOV) word, it will not only misrecognize the word in question, but most often it will misrecognize adjacent words as well; it is estimated that each OOV occurrence may trigger two to three word errors. In

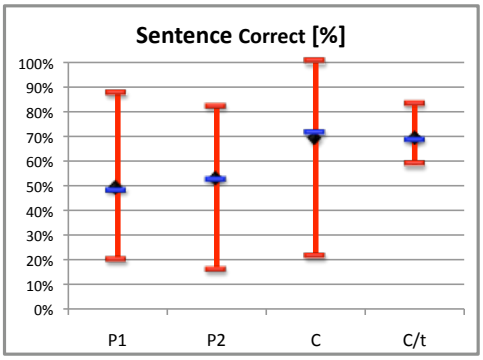

Figure 2. Sentence correct performance.

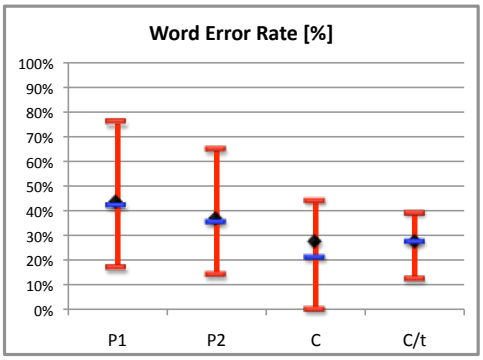

Figure 3. Word error rate performance.

our data we found an overall OOV rate of about 3% for the checkup phase (and as high as 5% overall for the P1 condition). These values would be rather large for a deployed system and point to the need for additional LM training data and/or expansion of the LM grammars. At the same time, given that no development data collected from real patients was available before the evaluation, this is not an unsatisfactory result.

Compared to the performance of the topic-independent LM, dynamic switching between topic-specific models was responsible for a 4.2% relative improvement in the overall WER and about 8% relative improvement in per-speaker average WER. Perplexity decreased about 10%, from just above 35 to under 32. Overall there were 5.4% fewer sentence errors. While these are very positive results, early indications based on the simulated dialogues seemed to suggest even better improvements would be possible. In fact, two of the topic-specific LMs actually had somewhat lower performance compared to the topic-independent LM. This points to weaknesses in the current LMs; one of them appears to suffer from insufficient coverage (it had the highest OOV rate of all models). It is likely that the interpolation weights will need further tuning.

As mentioned above, LMs based on synthetic data tend to have reasonable n-gram coverage, but may exhibit poor fit to the actual word distribution they are estimating. If that were the case, we'd expect that even though the top hypothesis may not be the right one, a better hypothesis should be present somewhere in the recognition lattice. The closer the LMs are to the real distribution, the higher the correct hypothesis would be ranked in the lattice. We obtained the 10 best hypotheses for each utterance and checked how much performance improvement could be obtained had the LMs been better at predicting users' utterances. Results for both the SC and WER measures are shown in Figures 4 and 5. We see that, when the correct hypothesis is covered by the LMs, it is typically among the 7 best hypotheses, and the largest improvement is obtained by going from the 1-best to 2-best hypotheses. Pushing through these apparent performance bounds would require, at the very least, extending the coverage; adding user-specific data, for example based on prior conversations, would also help.

Given that perfect LMs are a practical impossibility, the natural language understanding components in our system are designed to work with multiple SR hypotheses; specifically, they could pick the interpretation resulting from a lower-ranked hypothesis, if it makes more sense in the current context. While we did not use this feature in the experiments reported here, it appears that very significant improvements could be obtained by using it. The cost of handling 7-best hypotheses would not pose an unreasonable burden on the system to affect its responsiveness. Of course, this feature can also be responsible for introducing interpretation errors; in follow-up experiments we plan on testing empirically the advantage of using multiple hypotheses in our system. We are also planning on using confidence measures to decide, for each utterance, how many hypotheses should be output.

C. *Discussion*

All subjects managed to finish their sessions, and felt fairly comfortable with our system. The final verdict on the system's usability will come from the results of the system's understanding performance; however, the SR performance is

Figure 4. N-best performance (sentence correct).

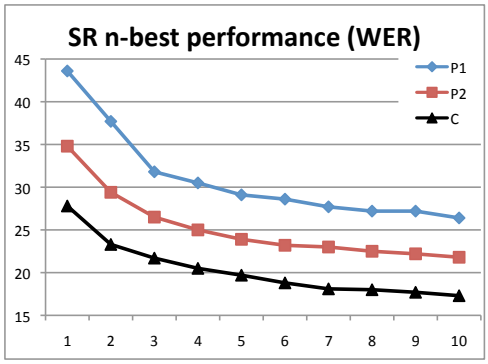

Figure 5. N-best performance (word error rates).

quite encouraging, especially given the difficulties posed by the speech data we collected (acoustic model mismatches, OOV words, disfluencies, etc.). Nonetheless, it is clear that there is plenty of room for further improvement, and above we outlined a number of avenues for improving the language models.

In addition, better acoustic models will be crucial for successful deployment. In our study, the average male participant got only half the WER of the average female participant. Since female subjects spoke in shorter utterances, and uttered fewer OOV words, and both these conditions are correlated with better performance, it seems the poorer performance can only be explained as being due to the acoustic models' being inadequate for female conversational style (this observation corroborates our prior experience with the gender-independent acoustic models we use). Ultimately, the application we envision would use speaker-adapted models; ideally we'd want to only use modest amounts of transcribed data (for example, just the data obtained during the first conversation) and then perform unsupervised adaptation. Fortunately, the literature suggests that it is possible to obtain significant recognition improvements with only a few utterances [20] and we intend to test this empirically in follow-up studies.

IV. CONCLUSION

We have described CARDIAC, an intelligent conversational assistant designed to promote successful health outcomes with patient-centered health monitoring technology. It is a first step towards developing a system that helps patients and/or their caregivers manage their medical care, provide reminders, answer questions, and engage in dialogue to collect information for monitoring a patient's current state. The information that the system collects during its interviews with the patient could be used by an interface which allowed the patient to view trends in their data, or exported to a PHR to share longitudinal data with their healthcare providers.

For the moment we have finished an analysis of the speech recognition performance on fourteen chronic heart failure patients. Although we used generic acoustic models and had practically no data for training domain-specific language models, by using synthetic data and topic-specific LMs we were able to field a usable system. Our initial estimate is that the current system's SR capabilities could handle successfully nine of the fourteen patients. We outlined a number of avenues for obtaining further improvements, to support both more accurate understanding, and thereby widen the potential pool of patients for which our technology could prove useful.

Along with removing the identified weaknesses from the system, longer-term we plan to evaluate the system with telephone speech, so it could be used by patients from their home. Eventually we hope to conduct in-home studies over several months with HF patients.

ACKNOWLEDGMENTS

Support for this research was provided in part by a grant from the Robert Wood Johnson Foundation. Additional funding was provided by NIH/NHLBI grant R21HL085396 "Feasibility of Conversational Systems for Patient Care." The first author also received support from the Office of Naval Research (N000140510314). We are grateful for helpful comments from three anonymous reviewers.

REFERENCES

[1] M.D. Naylor, D.A. Brooten, R.L. Campbell, G. Maislin, K.M. McCauley, and J.S. Schwartz, "Transitional care of older adults hospitalized with heart failure: A randomized, control trial," Journal of American Geriatric Society, vol. 52, pp. 675-684, 2004.

[2] E. Geisler and N. Wickramasinghe. "The role and use of wireless technology in the management and monitoring of chronic diseases," IBM Center for The Business of Government, Tech. Rep., June 2009.

[3] H. Jimison, P. Gorman, S. Woods, P. Nygren, M. Walker, S. Norris, and W. Hersh, "Barriers and drivers of health information technology use for the elderly, chronically ill, and underserved." AHRQ Evidence report/ technology assessment, no. 175, November 2008.

[4] L. M. LaFramboise, J. Woster, A. Yager, and B. C. Yates, "A technological life buoy: patient perceptions of the health buddy." The Journal of Cardiovascular Nursing, vol. 24, no. 3, pp. 216-224, 2009.

[5] T. Bickmore, and T. Giorgino, "Health dialog systems for patients and consumers." Journal of Biomedical Informatics, vol. 39:5, pp. 556-571, 2006.

[6] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent. "An architecture for a generic dialogue shell," Journal of Natural Language Engineering, vol. 6, pp. 1-16, 2000.

[7] G. Ferguson, J. Allen, L. Galescu, J. Quinn, and M. Swift. "CARDIAC: An Intelligent Conversational Assistant for Chronic Heart Failure Patient Health Monitoring," AAAI Fall Symposium Series: Virtual Health Care Interaction (VHI 09), Arlington, VA, 2009.

[8] CMU Sphinx Open Source Speech Recognition Engines, http://cmusphinx.sourceforge.net/html/cmusphinx.php

[9] L. Galescu, E. Ringger, and J. Allen, "Rapid language model development for new task domains," Proc. First International Conference on Language Resources and Evaluation (LREC), Granada, Spain, 1998.

[10] I. Azzini, D. Falavigna, T. Giorgino, R. Gretter, S. Quaglini, C. Rognoni, et al. "Automated spoken dialog system for home care and data acquisition from chronic patients," Studies in health technology and informatics. vol. 95, pp. 146-51, 2003.

[11] A. Kalai, S. Chen, A. Blum, and R. Rosenfeld, "On-line Algorithms for Combining Language Models," Proceedings of the International Conference on Accoustics, Speech, and Signal Processing (ICASSP '99), Phoenix, AZ, 1999.

[12] A. Stolcke, "SRILM - An extensible language modeling toolkit.", in Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado, September 2002. http://www.speech.sri.com/projects/srilm

[13] J. Allen, M. Dzikovska, M. Manshadi, M. Swift. "Deep linguistic processing for spoken dialogue systems," in Proc. ACL 2007 Workshop on Deep Linguistic Processing, pp. 49-56, Prague, Czech Republic, 2007.

[14] M.P. Georgeff and A.L. Lansky, "Reactive reasoning and planning," in Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87), 1987, pp. 667–682.

[15] A. Tate, B. Drabble, and R. Kirby, "O-Plan2: An open architecture for command, planning, and control," in Intelligent Scheduling, M. Zweben and M. S. Fox, Eds., Morgan Kaufman, 1994, pp. 213-240.

[16] D. Morley and K.L. Myers, "The SPARK agent framework," Proc. Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-2004), ACM Press, 2004, pp. 714-721.

[17] J. F. Allen and C. R. Perrault, "Analyzing intention in utterances," Artificial Intelligence, vol. 15, pp. 143–178, 1980.

[18] J. F. Allen and D.J. Litman, "Discourse processing and commonsense plans." in Intentions and Communication, P.R. Cohen, J. Morgan, and M. Pollack, Eds., MIT Press, 1990.

[19] P. Cohen and H. Levesque, "Intention is choice with commitment," Artificial Intelligence, vol. 42, pp. 213–261, 1990.

[20] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," Computer, Speech and Language, vol. 9, pp. 171-185, 1995.