The Transonics Spoken Dialogue Translator: An aid for English-Persian Doctor-Patient interviews

**S. Narayanan,*[*] S. Ananthakrishnan,*[*] R. Belvin,[†] E. Ettelaie,[*] S. Gandhe,[*]
S. Ganjavi,[*] P. G. Georgiou,[*] C. M. Hein,[†] S. Kadambe,[†] K. Knight,[*]
D. Marcu,[*] H. E. Neely,[†] N. Srinivasamurthy,[*] D. Traum,[*] and D. Wang[*]**

Abstract

In this paper we describe our spoken english-persian medical dialogue translation system. We describe the data collection effort and give an overview of the component technologies, including speech recognition, translation, dialogue management, and user interface design. The individual modules and system are designed for flexibility, and to be able to leverage different amounts of available resources to maximize the ability for communication between medical care-giver and patient.

Introduction

A spoken dialogue translator shares much in common with a dialogue system. Both require speech recognition, bi-directional language translation, dialogue tracking, and speech synthesis. One difference is that the language translation is from one natural language to another natural language rather than to an internal meaning representation. A dialogue system requires understanding, at least to the level necessary to decide what to say next and perform the desired task. A language translator, on the other hand, does not require understanding of the content, except as necessary to properly translate. There are also differences in the dialogue management. For a dialogue system, the dialogue manager will need to act as one side of the conversation, planning or choosing the utterances for that side. A translation system, on the other hand, mainly conveys the utterances from one human participant to another. While there are similarities in dialogue management, since in both devices a dialogue manager can modulate turn-taking, initiative, and context tracking, there are also big differences in the type of dialogue itself. When the system is one of the dialogue participants, depending on the domain, it can simplify many of the language processing tasks, by regulating the style of initiative and set of expected responses. On the other hand, when the language participants are both human, it is harder for the system to limit the style of interaction without putting a burden on the users. Spoken translation systems are thus both "easier" and "harder" than dialogue systems — less depth of understanding is needed to carry on the conversation, but the conversations themselves may be more complex. This puts burdens on all of the components for breadth of coverage.

As with dialogue systems, other challenges come from the specific domain and task of the conversation, and the particular language(s) spoken. For both language and domain, another limiting factor is the amount of language data available for speech processing, translation, and dialogue processing. Our transonics system is a translator between English and Persian for medical interviews. A doctor, or other medical care-giver speaks English, and is able to interview a persian-speaking patient about symptoms, background information, and diagnosis and treatment. This system was built as part of the Darpa Babylon program[1], which also had other sites building medical translation systems for Chinese, Pashto, and Thai. Some of the medical resources were shared, with individual translations to the individual languages.

System Overview

Our system comprises several spoken language components, as shown in Fig. 1. Modules communicate using a centralized message-passing system. In our architecture, all messages are broadcast with a tag that includes among other information the source and destination modules, but can be seen by all subsystems. The individual subsystems are the *Automatic Speech Recognition* (ASR) subsystem, which works using n-gram *Language Models* (LM) and produces n-best lists/lattices along with the decoding confidence scores. The output of the ASR is sent to the *Dialog Manager* (DM), which displays and passes on to the translation modules, according to a user-configurable state. The DM sends translation requests to the *Machine Translation* (MT) unit. The MT unit works in two modes: Classifier-based MT and a fully Stochastic MT. Depending on the dialogue manager mode, translations can be sent to the unit selection based *Text To Speech* synthesizer (TTS), to provide the spoken output. The same basic pipeline works in both directions: English ASR, English-Persian MT, Persian TTS, or Persian ASR, Persian-English MT, English TTS. There is, however, an asymmetry in the dialogue management and control, given a desire for the English-speaking doctor to be in control of the device. Figure 2 shows some examples of

[*] University of Southern California, Los Angeles, California
[†] HRL Laboratories, Malibu, California

[1] http://darpa-babylon.mitre.org

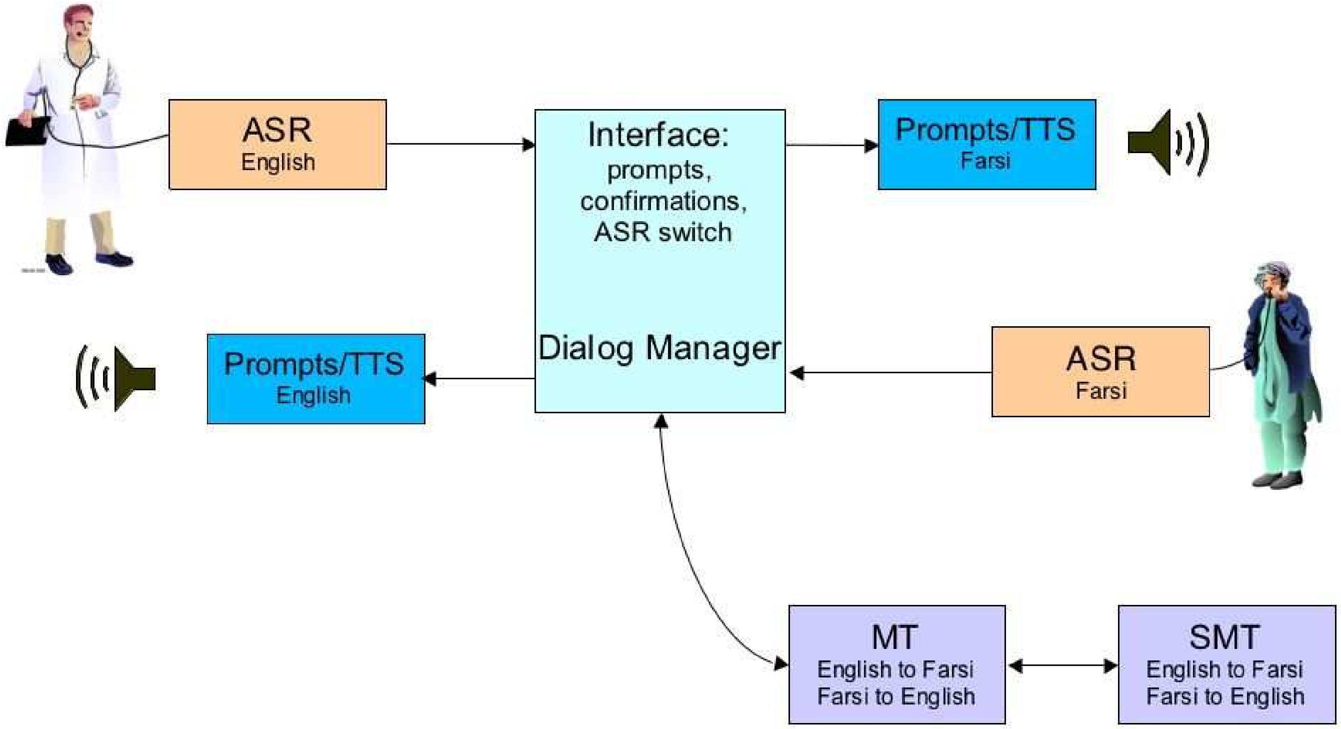


Figure 1: Block diagram of system. Note that the communication server allows interaction between all subsystems, and the broadcast of messages. Our vision is that only the doctor will have access to the GUI, and the patient will only be given a microphone headset.

the messages sent by the modules. The first tag shows who the message is to and from, and the type of message. The next field is an utterance ID number – shared by all messages pertaining to this utterance. After this is the content field, which depends on the message type. The first message is from the speech recognizer to the dialogue manager, showing the five best hypotheses with confidence scores for an english utterance. The dialogue manager sends messages to the GUI to update the display, and then sends the hypotheses one by one to the machine translation system, which returns the translation (in USCPers) with confidence, paired with the class (and original input, if different).

```
FADT 0260|YOU HAVE ANY QUESTIONS *0.079513|
          DO YOU HAVE ANY QUESTIONS *0.079700|
          HE YOU HAVE ANY QUESTIONS *0.080245|
          YOU HAD ANY QUESTIONS *0.080305|
          SEE YOU HAVE ANY QUESTIONS *0.080591
FDGC 0260|StatusBar*-1.0| Working on Translating
          english to persian*1.0
FDMT 0260|YOU HAVE ANY QUESTIONS *0.079513
FMDT 0260|467>sWAl dygry ndAryd*0.00763885|
          5467>DO YOU HAVE ANY QUESTIONS*0.00763885|
FDGT 0260|YOU HAVE ANY QUESTIONS *0.079513|
         5467>DO YOU HAVE ANY QUESTIONS*0.00763885|
          467>sWAl dygry ndAryd*0.00763885
FDMT 0260|DO YOU HAVE ANY QUESTIONS *0.0797
FMDT 0260|467>sWAl dygry ndAryd*0.00471005|
          5467>DO YOU HAVE ANY QUESTIONS*0.00471005|
```

Figure 2: Examples of System messages

Data Collection & Transcription

All of the major system modules rely on data-driven approaches. Since there was no ready-made source of large-scale data, this required a major data-collection effort within the project. Our approach to this problem was multi-pronged: leverage and adapt existing resources, and develop new resources. The ideal data set would be mediated/translated dialogues, of precisely the sort that the system would engage in. Next best would be dialogues mediated by human translators. The initiative and size, speed, and complexity of the dialogues would be different, but there would still be ready-made examples of speech and translations. Finally, mono-lingual Doctor-patient dialogues would at least give domain coverage, and could be translated off-line. Here the mediated nature is also missing. Other modules could also make use of other data sources, such as paraphrases of existing utterances, concept lists, translated material (in and out of domain), and spoken samples in Farsi and English.

Table 1 provides a summary of the currently identified data sources for the Transonics system development. It

Data Description	Size and original form	Used in
English questions and answers (Marine Acoustics)	600 sentences in English and Persian - text	FSG, CBMT
Paraphrasing above	2000 sentences - text and audio	CBMT
WoZ experiments	100+ utterances in audio format	All modules
Medical phrasebooks	600+ Q&A utterances. Translated and tran	All modules
Persian newspaper mining	Virtually unlimited text. Continuously converted to USCPers+	LM
DLI data - spontaneous speech	About 5h of mediated doctor-patient interaction – audio	All modules
CECOM data - scripted	500 English with Persian translations - No transcription	
CECOM data - semi-spontaneous	75 Q&A pairs in both English and Persian - No transcription	
IBM data - semi-spontaneous	Force protection data. About 350 short interactions.	LM, DM
USC Medical school	Videotaped medical interactions – standardized patient examinations	LM, DM
USC/HRL medical data collection	Doctor patient interaction, using trained patients and medical students. 200 Dialogs in audio format.	LM, DM

Table 1: List of data sources and uses. (FSG -finite state grammar, LM - Language Model, CBMT - classifier based machine translation, DM - dialog manager).

includes domain material gathered from existing resources (and translated) or material being specifically collected as a part of the project. The first 4 rows are target-domain material completely transcribed/translated in both English and Persian. For the medical domain, the initial bootstrapping was based on the availability of a large amount of common medical expressions, obtained from Marine Acoustics Inc[2] and a number of medical phrase books. These were not only useful for the creation of fixed state grammars and the Classifier based MT, but are also valuable in enriching our medical domain vocabulary (especially the medical phrase book data). In addition, for supporting development of language models for generic larger vocabulary recognition in Persian, we have been gathering Persian text corpora from mining publicly available newspapers. Due to the tremendous amount of data needs and processing involved, we are continuing the collection and transcription process.

The other major focus of our data needs is spoken language interaction data in the target domains. Although we identified some limited sources of existing spoken dialog interaction data (in English for *e.g.*, doctor-patient dialogs), these are significantly different from the mediated dialogs of the Transonics system. Hence, a significant portion of current efforts is focused on generating actual interaction data (both monolingual and bilingual modes) in the target domain. In addition to feeding the ASR and MT modules, these data are valuable for designing the dialog interface.

USC Standardized Patient Data Collection

The most important data collection effort that has been undertaken by USC/HRL, and in collaboration with the USC Keck School of Medicine, is the *Standardized Patient* data collection. The practice of using *Standardized Patients* began in Los Angeles in the 1960s as a way of allowing medical students to gain experience interacting with and diagnosing patients, and with a greater degree of consistency in terms of symptoms displayed; moreover, the patients are trained to rate the students on their bedside manner, handling of the physical examination, and methods of diagnosis.

Standardized Patient cases are created by MD's and RNs, ideally ones who have had first-hand experience with such medical instances. The cases consist of a detailed description of the symptoms the standardized patient is to report (some brief samples are shown below), as well as a one-page synopsis of some of the patient's vital signs, which will differ from their actual vital signs, but which will serve as important indicators to the students in forming a proper diagnosis (see Fig. 3). The SP goes through extensive training and two practice run examinations by qualified MD's. Notice that the instructions are very specific, but do not tell the patient exactly how to report the symptoms. This is important for the dialog data collection, as we are interested in collecting variations on the way that particular symptoms are reported.

The cases, which include among others, tuberculosis (TB), malaria, flu, heart attack, severe diarrhea, *etc*., were chosen not only to get better balance of illnesses to injuries – the vast majority of the IBM data collection are injuries – but also after research into published material by military insti-

[2] http://www.sarich.com

(A) Brief patient instructions:
The cough started about 3 months ago. It is constant and produces sputum that is usually thick and yellow and occasionally has some flecks of blood in it. The sputum does not have any bad smell. The cough is deep and you have occasional coughing "fits." The cough is fairly constant, happening often during both the day and the night. You have also lost weight during this time without dieting. You have noticed your skirt/pants have become very loose.

(B) Doctors chart of vital signs:

Temperature:	99 degrees F
Pulse:	100
Respiration:	18
Blood Pressure:	112/80 mm Hg

Figure 3: (A) Only seen by patient. (B) Seen by both medical student and patient

tutions (such as the Naval Medical Research Institute, Army sources, Army Research Inst. of Environmental Medicine, *etc*.).

Transcription Methods

Data transcription in Persian is a harder task than in many languages. First, Persian is traditionally represented in Arabic script, with stylizations of some characters depending on the position of the character within the word. One issue is that there is no accepted standard for computer encoding of the characters, with different codings used for different archives of existing text. Thus, first one must choose a character set to represent the Arabic script that allows easy access to non-Persian speakers, avoids the multiple character forms, and reduces the transcription overhead. Our choice has been the creation of the `USCPers` ascii-based transcription scheme. Additionally, we need to provide pronunciations of each word (for ASR & TTS), and we have created the `USCPron` transcription system to address this issue. Furthermore, the Persian written system does not include the vowel sounds in its written form, and thus multiple transcriptions on USCPers or the original Arabic script can result in both different pronunciations and different meanings. This prompts the need for a different transcription scheme that enables a one-to-one mapping between the acoustic representation (excluding user variability) and the transcription method, which we introduce as `USCPers+`. An in-depth analysis of our transcription schemes is given in (Ganjavi, Georgiou, & Narayanan 2003).

System Components

In this section we briefly describe the major components of the system. Each of them is tuned for translated medical dialogues, using a variety of techniques and data sources.

Speech Recognition

To recognize speech utterances we employ separate English and Persian Speech recognizers, built using Sonic (Pellom 2001), with local models and training data. For English, we are able to take advantage of existing large language models, such as the wall street journal corpus. In Persian, however, there is a lack of adequate existing speech data. To overcome this drawback we adopted a data driven language-adaptive technique. We borrowed acoustic data from English to compensate for the lack of data in Persian. The key requirement in enabling the use of English data for Persian ASR is the development of a phoneme mapping between the two. We used a novel Earth Movers Distance based sub-phonetic/phonetic mapping (Srinivasamurthy & Narayanan 2003). Additionally, for adaptation/re-training, the Marine Acoustics data was translated into Persian and read by 18 native Persian speakers (9 females and 7 males). We compared Persian adapted/re-trained ASRs using seed models from (i) sparse Persian speech data (FARSDAT), (ii) knowledge based English phonemes, (iii) data driven phonetic models and (iv) data driven sub-phonetic models as shown on Table 2. The results we obtained are very encouraging, illustrating that it is possible to make use of acoustic data even between diverse languages like English and Persian to improve the performance of ASRs in languages constrained by sparse data. We also observe that our proposed technique while having better performance when the re-training is used does not perform as well when only adaptation is used. A possible reason for this is that the adaptation scheme used, MLLR, is restricted to only linear transformations, which may not be sufficient to model differences in phonemes between different languages, where phoneme contexts play an important role.

Seed Models	Phoneme Error Rate	
	Re-training	Adaptation
FARSDAT	20.35%	38.95%
Knowledge based	20.00%	39.87%
Phonetic mapping	20.13%	57.03%
Sub-phonetic mapping	19.80%	51.48%

Table 2: Phoneme error rates obtained for different approaches. Observe that sub-phonetic mapping ASR achieved the best recognition performance when re-training was used.

To develop a full LM based ASR for Persian, we have successfully mined data from Persian news sources. Conversion to USCPers can be automated to a large degree. The data are subsequently processed by our team of transliterators to create the USCPers+ script of the same text, while at the same time minor modifications may be made to reflect predefined classes. The LM generated from this data as well as our existing English language LM will be interpolated with the ones we expect to create from the limited amount of medical data available, such as the standardized patient examination data and the USC/HRL collection effort.

Machine Translation

The approach we employ for the Machine Translation unit is twofold. A classifier is applied as the main translator unit of the system because of its faster and more accurate performance, while a *statistical machine translator* (SMT) is kept as the backup unit for the cases when the classifier response is not within an acceptable confidence margin. These cases should be relatively infrequent if significantly large number of classes are chosen for the classifier based MT.

As a first step in building a classifier, the proper set of standard questions and answers that covered the context was selected. Every standard question or answer was chosen as a representation of a class.

Following each input utterance, the system is expected to classify it in one of the predefined classes and generate the pre-stored translation. This requires training data to create the classes that are represented by each standard question or answer. To collect this data we created an online tool where a sentence was presented and users were asked to paraphrase it, thus expanding the coverage area of our training corpus. In addition, and in order to model expected errors introduced from the ASR module, we collected acoustic paraphrasing data that were not cleaned (*i.e.,* the recognized transcript may not match the uttered speech).

The resulting dataset was used to train a naive Bayesian classifier with uniform prior probabilities. The test set was gathered from an ASR and consists of both standard questions and paraphrased data. Since the test phase paraphrasing is separate and additional to the training set paraphrasing, there is significant test phrases that are new for the system.

From the collected data we have established a monotonically increasing performance in the MT classification as the paraphrasing increases. With the available paraphrasing the performance is roughly linearly increasing with over 1-2% per paraphrase round. We are continuously collecting more data to improve our classification quality, and additionally we are introducing more original phrases to increase our domain coverage.

A more sophisticated classification scheme that consists of a lattice of finite state transducers has been under development. A group of FST's model each main block in the real system. Thus, the ASR is modeled by a phoneme corrupter FST followed by a phoneme-to-word transducer. A set of unigram based FST's associated with each class followed by a bigram filter and a word-to-phoneme FST forms the speaker model. For every utterance from the (real) ASR a detection procedure, *e.g.,* Viterbi algorithm is performed to get the corresponding class. Early experiments with this system show 3.4% increase over the accuracy of the naive Bayesian classifier. Better ways of using the training data to build these FST's are under investigation.

The second method, to be used in the case of poor classification confidence, is the Stochastic MT method. SMT is based on word-to-word translation and can generate the translation for every input sentence. However, accuracy of the SMT, although a function of the training corpus, is in general expected to be lower than the accuracy of the classifier. The best performance for SMT can be achieved by employing a large amount of bilingual parallel text for training. This training corpus can be used to build a language model for the target language along with a statistical translation table, which relates words in the source language and their counterparts in the target language. However, due to the lack of any significant amount of English-Persian parallel text, we are following the approach of using the initial and target language models, and combining these with a dictionary approach for transition between the two languages. The LMs used are the same as has already been discussed in the ASR section.

Finally, another consideration in favor of the classifier system is the high computational demands of the SMT algorithms. In a speech to speech system where latency is crucial, the SMT system would always be kept as a backup choice after the utilization of the faster classifier system.

Text to Speech

We rely on a hybrid unit selection based speech synthesis. In the default case, when the output is chosen from a classifier-based MT, the generated phrases are known a priori. Hence, our first system release enabled us to use a prompt based system for spoken output. The other end of the unit selection possibility is through diphone concatenation. We have implemented such a synthesizer based on Festival[3] for English and Persian. Note that there are 29 sounds in the Persian language (6 vowels & 23 consonants), which results in the theoretical number of 900 diphones, fewer than needed for English that has a larger vowel inventory.

Dialogue Manager and User Interface

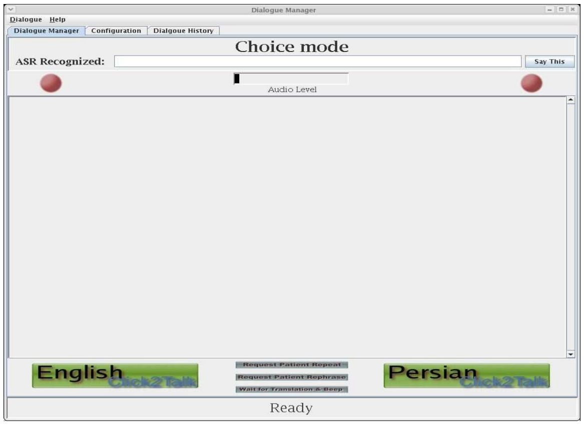
The dialogue manager component is closely bundled with the user interface and has the main task of using the other components to promote effective communication between the participants. The dialogue manager performs the following tasks:

- presents the medical care-giver with options for style of dialogue flow
- presents the care-giver with a visualization of the dialogue, speech recognition, and translation processes
- depending on the user-settable configuration and user choices, calls other modules for processing
- keep track of the discourse history
- provide hypotheses of most likely next utterances
- manage the turn-taking and grounding interaction between speakers

We have chosen an asymmetric dialogue flow, assuming that the medical care-giver will maintain the initiative in the dialogue, will have sole access to the controls and display of the translation device, and will operate the push-to-talk controls for both him or herself and the persian patient. There are several reasons for this choice, ranging from the predominance of Doctor initiative to the practicalities of knowledge of how to operate the device, keeping possession, as well as limitations of current technology such that push-to-talk recognition is currently more reliable than "always listening" systems.

Figure 4 shows the initial screen in choice mode. The medical care-giver can choose to click on the left green button to speak in English, or on the right to have the patient speak in Persian. Also, there are three smaller buttons in between, which will play recorded Persian requests for the

[3] http://www.cstr.ed.ac.uk/projects/festival



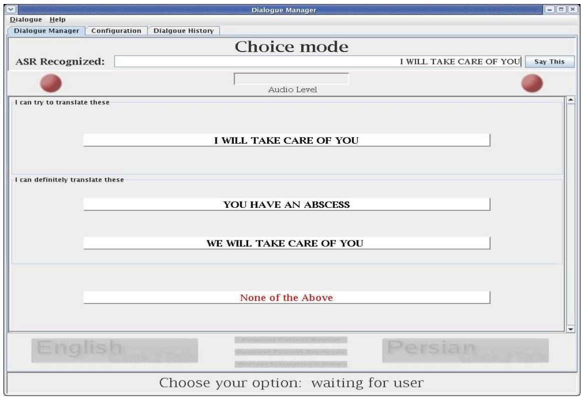
Choice mode

ASR Recognized:

Ready

Figure 4: Transonics Interface Ready for Speech

patient to repeat, rephrase, or wait for the beep before speaking. While the speech recognizers are active, the audio-level meter shows sound signal, and listening buttons turn green. As the speech recognition module is producing recognition hypotheses, the current best hypothesis is shown and updated in the "ASR recognized" box. When speech recognition is finished, the top five results are sent to the machine translation system. In choice mode, for English to Persian these are presented in two sections, as in Figure 5, headed "I can try to translate these" (results from the SMT, which may or may not end up in fluent persian), and "I can definitely translate these" (results from the classifier, which will be fluent, but might not be close enough to what the English speaker said). The care-giver can click any of these buttons to translate the selection to Persian, or can choose "none of the above", and try to speak again (either another attempt at the same thing, or changing the input). For Farsi recognition, the same process is used, except that the first choice is always said, while all translations are available to the care-giver for inspection. The care-giver can choose a response from the set presented (for updating the dialogue history), if there is one that is sufficiently clear and relevant, or can ask the Persian speaker to confirm (or can try to ask a follow-up question).



Choice mode

ASR Recognized: I WILL TAKE CARE OF YOU

I can try to translate these

I WILL TAKE CARE OF YOU

I can definitely translate these

YOU HAVE AN ABSCESS

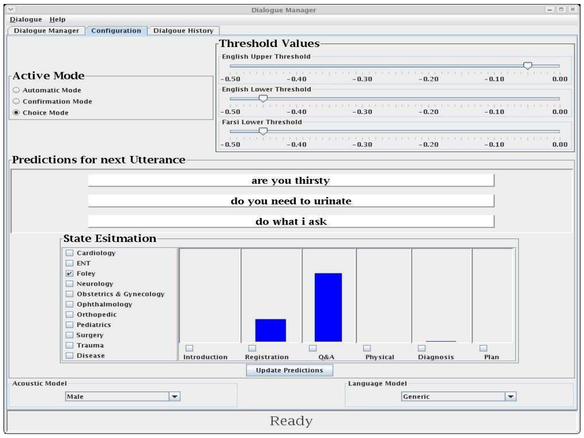
WE WILL TAKE CARE OF YOU

None of the Above

Choose your option: waiting for user

Figure 5: Choice mode

Figure 6 shows the configuration screen, giving an advanced user additional options for system performance modes. In the left corner, there is the choice of basic modes. In addition to "choice mode", described above, there is "automatic mode", in which the first choice translation is played without need for gui intervention. Confirmation mode is between the two, where only the top choice is presented, but the user must still select whether to speak the translation or try again. In the upper right corner are confidence thresholds for whether the translations are good enough to proceed. The top translation with confidence higher than the upper threshold (english only) will be translated without need for confirmation. Any hypotheses below the lower threshold will be pruned and not even presented to the user. The translations in between the thresholds will be presented for user choice and confirmation.



Active Mode

○ Automatic Mode

○ Confirmation Mode

● Choice Mode

Predictions for next Utterance

are you thirsty

do you need to urinate

do what i ask

□ Cardiology
□ ENT
☑ Foley
□ Neurology
□ Obstetrics & Gynecology
□ Ophthalmology
□ Orthopedic
□ Pediatrics
□ Surgery
□ Trauma
□ Disease

Acoustic Model: Male

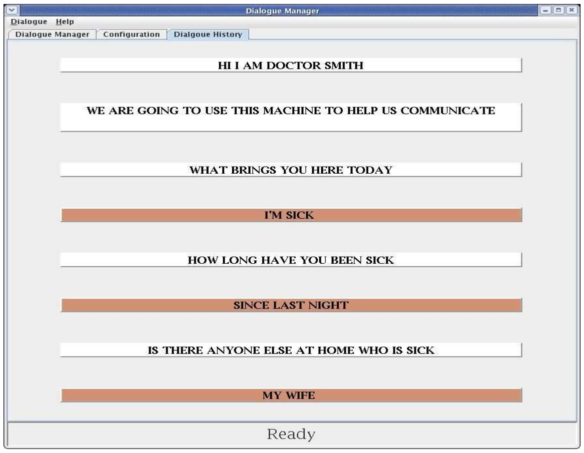
Language Model: Generic

Ready

Figure 6: Configuration Screen

The bottom part of Figure 6 shows the dialogue manager's calculation of the current medical case and phase within the examination. The phases include *introduction, registration, Q&A, Physical examination, Diagnosis*, and *Treatment Plan*. The dialogue manager's current estimation of the probability distribution for the phases is shown here (the classifier was built using the MALLET toolkit (McCallum 2002). In this example, Q&A is the most likely possibility, though registration is still somewhat likely. Probability estimates are based on the correspondence between phases and translation classes, as calculated from hand-labelled collected dialogues. There are also click buttons so that the user can inform the system of the current case and phase when that is known (in this case, the probability becomes close to 100% for the next utterance). Just above the phase estimation is a set of predictions of most likely next utterances for the care-giver, based on the estimation of case and phase. The user can select one of these to translate rather than speaking his own utterance, if desired.

Finally, the dialogue manager keeps a history of the accepted utterances in the dialogue (English utterances that were translated to Persian, and Persian replies accepted by the care-giver), as shown in Figure 7. The care-giver utterances are shown in white, and the patient replies shown in



HI I AM DOCTOR SMITH

WE ARE GOING TO USE THIS MACHINE TO HELP US COMMUNICATE

WHAT BRINGS YOU HERE TODAY

I'M SICK

HOW LONG HAVE YOU BEEN SICK

SINCE LAST NIGHT

IS THERE ANYONE ELSE AT HOME WHO IS SICK

MY WIFE

Figure 7: Dialogue History view

brown. This can be a useful memory aid for the doctor, and can also be used to go back and revisit any vague or troublesome points. Clicking on one of the utterances will cause the system to replay the associated translations.

Evaluation

As part of the Darpa Babylon program, our transonic system has been evaluated for basic usability by MITRE. The evaluation involved MITRE-recruited English speaking government Doctors and Nurses, and Farsi speakers acting as patients (with training similar to that for the standardized patient program). While the formal analysis of the results are not yet available, informal observation of the process shows that the participants can sometimes communicate quite effectively using the device. On the other hand, some interactions were less successful at reaching a correct diagnosis and treatment plan. The translation protocol may be difficult to learn, especially for older patients. More work is still required to determine how much the system-mediated interviews degrade from mono-lingual interaction, and how much they add in a situation where no other translation assistance is available.

Acknowledgments

This work was supported by the DARPA Babylon program, contract N66001-02-C-6023.

References

Ganjavi, S.; Georgiou, P. G.; and Narayanan, S. 2003. Ascii based transcription schemes for languages with the arabic script: The case of persian. In *ASRU*.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Pellom, B. 2001. Sonic: The university of colorado continuous speech recognizer. Technical Report TR-CSLR-2001-01, University of Colorado.

Srinivasamurthy, N., and Narayanan, S. 2003. Language-adaptive persian speech recognition. In *Eurospeech*.