Creation of a Doctor-Patient Dialogue Corpus Using Standardized Patients

Robert S. Belvin

HRL Laboratories
3011 Malibu Canyon Road
Malibu, CA 90265
rsbelvin@hrl.com

Win May

USC Keck School of Medicine
1975 Zonal Avenue
Los Angeles, CA 90033
winmay@usc.edu

Shrikanth Narayanan[§], Panayiotis Georgiou[§], Shadi Ganjavi*

[§]Department of Electrical Engineering, *Department of Linguistics
University of Southern California
University Park Campus, Los Angeles, CA 90089
{shri/georgiou}@sipi.usc.edu, ganjavi@usc.edu

Abstract

In this paper we describe the development of a doctor-patient dialogue corpus to support a speech-to-speech machine translation effort for English-Persian medical dialogues. The corpus was developed by recording and transcribing English-to-English dialogues between medical students and standardized patients (actors who have been trained to portray illness or injury victims), and then translated into Persian. We discuss some of the benefits and drawbacks to creating a corpus in this way. Benefits include the ability to customize the corpus in a way that would be infeasible for actual doctor-patient data and avoidance of privacy and legal issues, while drawbacks include the fact that the Persian does not originate as speech, but as text translation of English speech. We address concerns such as the authenticity of the dialogues and the value of such data for system development.

1. Introduction

The DARPA CASTE Program has as its goal the creation of two-way, speech-to-speech language translation systems for narrow domains, including "first encounter" medical care in field environments for several language pairs. HRL Laboratories is part of a collaborative effort with several organizations within USC to develop the English-Persian[1] translation system. In pursuing the project goal, various sources of medical language data have been exploited, including data from the DARPA 1-way translator program, medical phrase-books, and the British National Corpus.

2. Data Requirements

The materials mentioned above provided some basic vocabulary and phrases for carrying out a medical interview, however, we determined after reviewing available resources that there was still a need for material that was focused in such a way as to be suitable for building language and translation models for the specific application we are targeting. The target scenario for us is a situation in which an English-speaking doctor is interviewing/examining a monolingual Persian-speaking patient for a chief complaint of a certain type, and there is an electronic bilingual interpreter present, mediating the interaction. A scenario which would most closely *simulate* this target, but without actually entailing the existence of a system prototype, is exactly the same except there is a human bilingual interpreter present, rather than an electronic one. Thus, the characteristics of a *perfect* development corpus for us (again without entailing the prototype system) would include:

i. A large amount of spontaneous spoken dialogue between a medical professional and a real patient
ii. Coverage of the illnesses and injuries we were most interested in covering (and would not include a large amount of dialogue about medical situations we had no interest in)
iii. An English monolingual doctor, a Persian monolingual patient, and a balanced bilingual interpreter translating for them
iv. Relatively short and to-the-point questions and responses

In practice, such a corpus, especially of an appropriate size (minimum 250,000 words), would be prohibitively expensive to produce, and impossible to find as an existing product. Moreover, obtaining recordings of actual doctor-patient interactions is beset with severe privacy and other legal problems. What we have done as a fallback strategy, therefore, is to create a monolingual English corpus simulating doctor-patient interactions in collaboration with faculty from the Standardized Patient program at the USC Keck School of Medicine, and then translated that corpus into Persian. Although this method still does not capture all of the characteristics we would like to have in our imaginary perfect corpus, it does include most of them and can provide us with sufficient data to build an initial system with which further interaction data can be gathered. Specifically, the method allowed us to achieve (i, ii, and iv) from the list above; we can only partially simulate a corpus with characteristic (iii) using this method, by post-hoc translation of the dialogues into Persian. We are aware of the pitfalls of

[1] Persian is a member of the Iranian branch of the Indo-European languages and is spoken in Iran, Afghanistan and Tajikestan. The dialect spoken in Iran is also known as Farsi, while the one spoken in Afghanistan is known as Dari (Windfuhr, 1987).

trying to simulate a spontaneous Persian dialogue by translating an originally English dialogue. We discuss some of the shortcomings in a later section of the paper.

3. Standardized Patients

We now turn our attention to the process of creating this corpus, which forged a new and very productive alliance between MT system developers and a medical education organization, which to our knowledge is unprecedented.[2] The practice of using Standardized Patients began at the University of Southern California School of Medicine in the 1960's as a way of allowing medical students to gain experience interacting with and diagnosing patients, but without the problems associated with real patients, and with a greater degree of consistency in terms of symptoms displayed and reported.

Standardized Patients (SPs) have been carefully trained to portray all of the characteristics of a real patient, in order to provide the opportunity for a student to learn, or be evaluated, on clinical skills first hand. The term "Standardized Patient" was coined by Norman on the basis of the fact that the student-patient challenge to each student remains the same. We note that, since our purpose is not concerned with evaluating the medical students, the SP training was appropriately simplified, and will be discussed in more detail shortly.

The cases which standardized patients portray are based on actual patients encountered by physicians. The SP is trained to simulate not only the signs and symptoms, but also the emotional and personality characteristics of the patient, thus presenting the "gestalt" of the patient being simulated (Barrows, 1987). Unlike real patients who may be too ill for repeated interactions with medical students, SPs can reproduce the history, physical findings and the behaviors of the patients over and over again.

Although there is understandable concern regarding the authenticity of the dialogues which occur in an SP interaction, we will discuss some characteristics of the SP dialogue corpus which we gathered shortly, which indicate its usefulness in the system development effort. We believe it is also a significant fact, and one that appears as a vote of confidence for the authenticity of the patients, that the use of SPs is becoming an essential part of not only the training but also the licensing of MDs, both in this country and abroad. A survey of the 142 curriculum deans of US medical schools conducted in 1993 showed that 111 (80%) of the 138 responding schools indicated that SPs were being used in teaching and assessment at their schools (Anderson et al, 1994).

The Educational Commission for Foreign Medical Graduates (ECFMG) already uses a Clinical Skills Assessment, where SPs are used to test the clinical skills of international medical graduates seeking certification to enter residency programs in the United States (Educational Commission for Foreign Medical Graduates, 2004). The National Board of Medical Examiners (NBME) will be administering a Step 2 Clinical Skills (CS) Exam utilizing standardized patients, as part of the examination procedures for licensure for the medical school class of 2005 (United States Medical Licensing Examination, 2004). There are three primary components for the Step 2 CS Exam: (1) the Integrated Clinical Encounter, which comprises history taking and physical examination as well as completion of the patient note, (2) Communication/Interpersonal Skills, and (3) Spoken English Proficiency. National certification and /or licensure examinations using standardized patients are used in Canada (Medical Council of Canada, 2004) and the United Kingdom (General Medical Council, 2004).

4. Standardized Patient Case Materials

Standardized Patient (SP) cases are created by health professionals such as MDs and RNs who have had first-hand experience with such cases (as noted earlier), in collaboration with standardized patient educators. The cases consist of a detailed description of the symptoms the standardized patient is to report, the physical signs they need to exhibit, as well as a one-page synopsis of some of the patient's vital signs (which may differ from their actual vital signs) but which will serve as important indicators to the students in coming towards the correct diagnosis (shown below).

> *The cough started **about 3 months ago**. It is **constan**t and produces sputum that is usually **thick** and **yellow** and **occasionall**y has some **flecks of blood** in it. The sputum **does not have any bad smel**l. 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.*

1. Example of Instructions to the Patient[3]

Notice that the instructions are very specific, but do not tell the patient exactly *how* they should report the symptoms. This is important for the dialogue data collection, as we are interested in collecting variations on the way that particular symptoms are reported. The training that the patients received for our project was somewhat different than for the true "standardizing" process, in that they were *encouraged* to vary their wording and playing of the role for each interaction, rather than trying to ensure that it was the *same* for each interaction.

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

2. Examples of the vital signs, which will be seen by both the patient and the medical student

5. Balancing the Concerns of the Two Organizations

The protocol we developed includes a mix of considerations:

[2] Though apparently the NESPOLE! project has employed a similar method for collecting medical dialogue data also (thanks to an anonymous reviewer for pointing this out to us).

[3] O*nly* the patient sees these; the medical student does not see these.

a. Because the Standardized Patient programis attracting the Medical Students in part by advertising this event as an opportunity to gain additional practice in interacting with patients, they are interested in keeping the dialogue as unconstrained and natural as possible.
b. Because as system developers we are interested in data that we can transcribe reasonably easily and use in acoustic and language models, we don't want it to be *too* unconstrained, e.g. we want to minimize speaker overlap and encourage short turns, and minimize digressions by the patient.

Thus, we asked our medical colleagues to train the patients (as part of their SP training) not to interrupt the doctor, and to make their answers relevant and relatively short. We also imposed a simple push-to-talk *prop*, as a way of discouraging interruptions and digressions. The prop was handed back and forth between the patient and the medical student, and included a button they were told to push before beginning to speak. This strategy was effective for some of the interactions, but was often abandoned part way through the interaction. In retrospect, it would have been more effective had it been an *actual* push-to-talk device.

Also, it is worth mentioning that the cost of producing this custom corpus was reasonable, due to several factors, including the fact that the medical students were willing to work for a fraction of what most medical professionals would charge for their time, in large part because they regarded the activity as educational, helping to prepare them both for interacting with real patients, as well as for medical school exams which include an SP component. The most expensive part of the endeavor was the post-processing activities of transcription and validation.

6. Resulting Corpus

The audio and text corpus produced by this activity includes 300 dialogues of approximate average duration of 12 minutes, and approximate average length of 1200 words. A small sample of a dialogue is shown below:

Doctor: <UM> how are you feeling today?
Patient: overall okay but I've been having this cough that has been <UH> bugging me
Doctor: when did you first notice the cough?
Patient: I'd say approximately three months ago
Doctor: and can you, do you think the cough has been <UM> the same kind of cough over the three months
Patient: yeah that would be a fair assessment yes
Doctor: can you tell me if you're producing any <UH> phlegm or sputum anything coming out when you cough
Patient: yeah I notice there has been some sputum coming out yes

3. Sample extract from the SP Corpus

The audio corpus was collected using high-quality head-worn (close-talking) microphones and DAT recorders sampling at 48kHz.

A natural question to ask is how authentic this corpus is in terms of the language generated by the student-to-SP interaction. A closely related question is how appropriate the data is for use as the basis of a speech-to-speech translation system. Regarding the first question, we note that, impressionistically, there are readily observable differences between some of the medical students and experienced doctors in terms of the level of explanation given to patients, the extent to which permission is requested of the patient, and generally the efficiency with which the interview and exam are carried out. The students appeared to offer more and longer explanations than experienced medical professionals, to ask permission more often (for example at the beginning of the physical exam), and generally to carry out a longer interview. These contrasts are based on comparing a small amount of data we have from advanced (4^{th} year) medical students, an RN and an MD (the more experienced group) with data from the largest group, who were 2^{nd} year medical students.

7. Discussion

Our overall impression of this method for generating appropriately focused medical interaction data is that it was very successful and provides a new approach to obtaining medical dialogues, and one which is not beset with many of the privacy and legal issues associated with obtaining and/or putting true medical dialogues into the public domain. It seems plausible, in fact, that for certain kinds of data requirements, even if one had the possibility of gathering genuine doctor-patient interactions, one might still opt to use this method as a means of gathering a large amount of the relevant kinds of cases quickly.

In spite of the apparent success of this method, we do not mean to be turning a blind eye to potential deficiencies and pitfalls inherent in it. Perhaps the most serious of these is the fact that the data is not bilingually generated, but only becomes bilingual by virtue of a translation bureau translating the transcribed English-English dialogues. There are several reasons why this method might lead to problematic data:

(i) There is very little room for cultural differences to emerge, including potentially fundamental discrepancies in aspects of worldview pertaining to illness, healing and medicine.
(ii) The foreign language originates as written language, not spontaneous spoken language.
(iii) It is probably very difficult to capture aspects of the speech of a person who is very ill or who has recently experienced a traumatic injury.

Instances of (i) which have emerged in some preliminary data collection we undertook with an Eastern Persian-speaking MD early in the project included problems such as:
a. Rural people may have the attitude that "more medicine is better"; the MD had seen cases in which people reported different symptoms than they actually were experiencing because they knew that was the way to get the most medicine.
b. People often do not finish courses of antibiotics because they feel better after a few days, and want to save some of the pills "for a rainy day."

c. If a person does not feel better shortly after taking a prescribed medication, they may double or triple doses to try to get better faster, in some cases leading to serious or even fatal overdoses.

Although problems such as these pertain to a different phase of patient care than the focused history and physical exam (which was the target interaction type for our effort), they nevertheless represent a kind of mismatch in worldview that the data collected from our method may be deficient in representing. However, it is also obvious that there are many aspects of the culture-clash issue that simply go beyond a corpus development problem. [4]

Regarding the problem noted in (ii), it is well-known that spontaneous spoken language differs from language which has originated as text. The fact that the foreign language is a translation of spontaneously spoken English may or may not contribute to language which more closely approximates spontaneously spoken Persian. It is important to note that we have not had time to adequately study the Persian translations of the English dialogues so as to be able to assess this at this point. We hope to report on this issue in future work.

In spite of the concerns noted above, it seems quite clear that there are many respects in which the data we have collected using this method is good data for system development purposes, especially in comparison to certain alternatives which have been used in various speech-to-speech translation projects (for example chat rooms, more scripted interactions, and so on). For example, we have collected more than 200 distinct instances of medical students asking a question to determine why the patient has come to them (some variants are "Can you tell me why you came in?", "Okay Martin what brings you to the clinic today?", "So what brings you in?", "Can you tell me what brings you here today?", "I wanted to talk to you about what brings you to the hospital today", "Okay so tell me what brings you here today", etc.) Such variations in the phrasing of a standard question are extremely valuable for system development purposes. In addition to the phrases we have collected in our SP data, the collection of paraphrases for what we anticipate to be relevant utterances for this domain has been an important component of our research, but the use of the SPs should provide by far the largest source for what amounts to paraphrases of critical questions and instructions on the part of the doctor or medic. So this characteristic of the data alone is an important result from this effort.

Another feature of the dialogues we collected is that they provide many instantiations of important diagnostic procedural "acronyms" used by all diagnosticians. An example of one of these is the pain series "PQRST" (Place, Quality, (what gives) Relief, Severity, Time). Moreover, despite the more tentative quality that many of the dialogues display, the core natural language diagnostic predicates appear to be very similar across different experience levels. For example, the utterances of the second year students on the one hand, and the fourth year students, the RN and MD on the other, displayed a high degree of lexical overlap in the predicates contained in diagnostic questions concerned with actual symptoms—where there are noticeable differences are in the area of various kinds of discourse markers: more hedges, filled pauses, and various other moderating devices. We will continue to study these differences and report on them in future work.

Finally, the issue noted in (iii) is simply beyond the scope of what one might hope to completely accurately represent in any kind of simulation. However, we note that for all three of the issues raised here, this method is not intended to be the be-all and end-all of the data collection process. Rather, it is intended to be the beginning. Once a prototype system has been developed, it must be augmented with a great deal more data, which is "real" or at least "realer". At this point, such data still appears to be very hard to come by. It is our hope that the medical community will become more aware of the potential of the kind of technology that this data collection enables, and that some of the administrative and legal obstacles will be eased.

[4] Still, certain considerations have been addressed in the Persian translations. For instance there is a formality difference between Persian and English in doctor-patient interactions. While in English-speaking North America it is customary for the doctor to address the patient in an informal manner, that is not the case among Persian speaking cultures. Therefore, in translating the English-English data, we attended to this difference.

8. Acknowledgments

This work was supported by the DARPA Babylon program, contract N66001-02-C-6023. We would like to thank the following individuals for their help, comments and suggestions: Dr. Naveen Srinivasamurthy, Emil Ettelaie, Josephine Cruz, Denise Souder, Cheryl Hein, and Howard Neely.

References

Anderson, M.B., Stillman, P.L., and Wang, Y. (1994). Growing use of standardized patients in teaching and evaluation in medical education. *Teaching and Learning in Medicine,* 6, pp 15-22.

Barrows, H.S. (1987) *Simulated (Standardized) Patients and Other Human Simulations*. Health Sciences Consortium, 201 Silver Cedar Court, Chapel Hill, North Carolina 27514.

EDUCATIONAL COMMISSION FOR FOREIGN MEDICAL GRADUATES (ECFMG) [http://www.ecfmg.org/csa/index.html, accessed 7 February, 2004]

GENERAL MEDICAL COUNCIL (2004) [http://www.gmc-uk.org/download/plab2.pdf, accessed 7 February, 2004]

MEDICAL COUNCIL OF CANADA (2003) Information Pamphlet on the Medical Council of Canada Qualifying Examination Part II (MCCQE Part II) pp 1-24.

UNITED STATES MEDICAL LICENSING EXAMINATION(USMLE) [http://www.usmle.org/step2/step2cs/2004step2cs.htm, accessed 7 February, 2004]

Windfuhr, Gernot L. (1987). Persian. In *the World's Major Languages*, Bernard Comrie (Ed.), pp. 523-546. New York: Oxford University Press.