Evaluating Task Performance for a Unidirectional Controlled Language Medical Speech Translation System

Nikos Chatzichrisafis, Pierrette Bouillon, Manny Rayner, Marianne Santaholma, Marianne Starlander
University of Geneva, TIM/ISSCO
40 bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland

Nikos.Chatzichrisafis@vozZup.com, Pierrette.Bouillon@issco.unige.ch, Emmanuel.Rayner@issco.unige.ch, Marianne.Santaholma@eti.unige.ch, Marianne.Starlander@eti.unige.ch

Beth Ann Hockey
UCSC
NASA Ames Research Center
Moffett Field, CA 94035
bahockey@email.arc.nasa.gov

Abstract

We present a task-level evaluation of the French to English version of MedSLT, a medium-vocabulary unidirectional controlled language medical speech translation system designed for doctor-patient diagnosis interviews. Our main goal was to establish task performance levels of novice users and compare them to expert users. Tests were carried out on eight medical students with no previous exposure to the system, with each student using the system for a total of three sessions. By the end of the third session, all the students were able to use the system confidently, with an average task completion time of about 4 minutes.

1 Introduction

Medical applications have emerged as one of the most promising application areas for spoken language translation, but there is still little agreement about the question of architectures. There are in particular two architectural dimensions which we will address: general processing strategy (statistical or grammar-based), and top-level translation functionality (unidirectional or bidirectional translation). Given the current state of the art in recognition and machine translation technology, what is the most appropriate combination of choices along these two dimensions?

Reflecting current trends, a common approach for speech translation systems is the statistical one. Statistical translation systems rely on parallel corpora of source and target language texts, from which a translation model is trained. However, this is not necessarily the best alternative in safety-critical medical applications. Anecdotally, many doctors express reluctance to trust a translation device whose output is not readily predictable, and most of the speech translation systems which have reached the stage of field testing rely on various types of grammar-based recognition and rule-based translation (Phraselator, 2006; S-MINDS, 2006; MedBridge, 2006). Even though statistical systems exhibit many desirable properties (purely data-driven, domain independence), grammar-based systems utilizing probabilistic context-free grammar tuning appear to deliver better results when training data is sparse (Rayner et al., 2005a).

One drawback of grammar-based systems is that out-of-coverage utterances will be neither recognized nor translated, an objection that critics have sometimes painted as decisive. It is by no means obvious, however, that restricted coverage is such a serious problem. In text processing, work on several generations of controlled language systems has developed a range of techniques for keeping users within the bounds of system coverage (Kittredge, 2003; Mitamura, 1999). If these techniques work for text processing, it is surely not inconceivable that variants of them will be equally successful for spoken language applications. Users are usually able to adapt to a controlled language system given enough time. The critical questions are how to provide efficient support to guide them towards the system's coverage, and how much time they will then need before they have acclimatized.

With regard to top-level translation functionality, the choice is between unidirectional and bidirectional systems. Bidirectional systems are certainly possible today[1], but the arguments in favor of them are not as clear-cut as might first appear. Ceteris paribus, doctors would certainly prefer bidirectional systems; in particular, medical students are trained to conduct examination dialogues using "open questions" (WH-questions), and to avoid leading the patient by asking YN-questions.

The problem with a bidirectional system is, however, that open questions only really work well if the system can reliably handle a broad spectrum of replies from the patients, which is over-optimistic given the current state of the art. In practice, the system's coverage is always more or less restricted, and some experimentation is required before the user can understand what language it is capable of handling. A doctor, who uses the system regularly, will acquire the necessary familiarity. The same might be true for a few patients, if special circumstances mean that they encounter speech translation applications reasonably frequently. Most patients, however, will have had no previous exposure to the system, and may be unwilling to use a type of technology which they have trouble understanding.

A unidirectional system, in which the doctor mostly asks YN-questions, will never be ideal. If, however, the doctor can become proficient in using it, it may still be very much better than the alternative of no translation assistance at all.

To summarize, today's technology definitely lets us build unidirectional grammar-based medical speech translation systems which work for regular users who have had time to adapt to their limitations. While bidirectional systems are possible, the case for them is less obvious, since users on the patient side may not in practice be able to use them effectively.

In this paper, we will empirically investigate the ability of medical students to adapt to the coverage of unidirectional spoken language translation system. We report a series of experiments, carried out using a French to English speech translation system, in which medical students with no previous experience to the system were asked to use it to carry out a series of verbal examinations on subjects who were simulating the symptoms of various types of medical conditions. Evaluation will be focused on usability. We primarily want to know how quickly subjects learn to use the system, and how their performance compares to that of expert users.

2 The MedSLT system

MedSLT (MedSLT, 2005; Bouillon et al., 2005) is a unidirectional, grammar-based medical speech translation system intended for use in doctor-patient diagnosis dialogues. The system is built on top of Regulus (Regulus, 2006), an Open Source platform for developing grammar-based speech applications. Regulus supports rapid construction of complex grammar-based language models using an example-based method (Rayner et al., 2003; Rayner et al., 2006), which extracts most of the structure of the model from a general linguistically motivated resource grammar. Regulus-based recognizers are reasonably easy to maintain, and grammar structure is shared automatically across different subdomains. Resource grammars are now available for several languages, including English, Japanese (Rayner et al., 2005b), French (Bouillon et al., 2006) and Spanish.

MedSLT includes a help module, whose purpose is to add robustness to the system and guide the user towards the supported coverage. The help module uses a second backup recognizer, equipped with a statistical language model; it matches the

[1] For example, the S-MINDS system (S-MINDS, 2006) offers bidirectional translation.

results from this second recognizer against a corpus of utterances, which are within system coverage and have already been judged to give correct translations. In previous studies (Rayner et al., 2005a; Starlander et al., 2005), we showed that the grammar-based recognizer performs much better than the statistical one on in-coverage utterances, and rather worse on out-of-coverage ones. We also found that having the help module available approximately doubled the speed at which subjects learned to use the system, measured as the average difference in semantic error rate between the results for their first quarter-session and their last quarter-session. It is also possible to recover from recognition errors by selecting one of the displayed help sentences; in the cited studies, we found that this increased the number of acceptably processed utterances by about 10%.

The version of MedSLT used for the experiments described in the present paper was configured to translate from spoken French into spoken English in the headache subdomain. Coverage is based on standard headache-related examination questions obtained from a doctor, and consists mostly of yes/no questions. WH-questions and elliptical constructions are also supported. A typical short session with MedSLT might be as follows:

- is the pain in the side of the head?
- does the pain radiate to the neck?
- to the jaw?
- do you usually have headaches in the morning ?

The recognizer's vocabulary is about 1000 surface words; on in-grammar material, Word Error Rate is about 8% and semantic error rate (per utterance) about 10% (Bouillon et al., 2006). Both the main grammar-based recognizer and the statistical recognizer used by the help system were trained from the same corpus of about 975 utterances. Help sentences were also taken from this corpus.

3 Experimental Setup

In previous work, we have shown how to build a robust and extendable speech translation system. We have focused on performance metrics defined in terms of recognition and translation quality, and tested the system on naïve users without any medical background (Bouillon et al., 2005; Rayner et al., 2005a; Starlander et al., 2005).

In this paper, our primary goal was rather to focus on task performance evaluation using plausible potential users. The basic methodology used is common in evaluating usability in software systems in general, and spoken language systems in particular (Cohen et. al 2000). We defined a simulated situation, where a French-speaking doctor was required to carry out a verbal examination of an English-speaking patient who claimed to be suffering from a headache, using the MedSLT system to translate all their questions. The patients were played by members of the development team, who had been trained to answer questions consistently with the symptoms of different medical conditions which could cause headaches. We recruited eight native French-speaking medical students to play the part of the doctor. All of the students had completed at least four years of medical school; five of them were already familiar with the symptoms of different types of headaches, and were experienced in real diagnosis situations.

The experiment was designed to study how well users were able to perform the task using the MedSLT system. In particular, we wished to determine how quickly they could adapt to the restricted language and limited coverage of the system. As a comparison point, representing near-perfect performance, we also carried out the same test on two developers who had been active in implementing the system, and were familiar with its coverage.

Since it seemed reasonable to assume that most users would not interact with the system on a daily basis, we conducted testing in three sessions, with an interval of two days between each session. At the beginning of the first session, subjects were given a standardized 10-minute introduction to the system. This consisted of instruction on how to set up the microphone, a detailed description of the MedSLT push-to-talk interface, and a video clip showing the system in action. At the end of the presentation, the subject was given four sample sentences to get familiar with the system.

After the training was completed, subjects were asked to play the part of a doctor, and conduct an examination through the system. Their task was to identify the headache-related condition simulated by the "patient", out of nine possible conditions. Subjects were given definitions of the simulated headache types, which included conceptual information about location, duration, frequency, onset

and possible other symptoms the particular type of headache might exhibit.

Subjects were instructed to signal the conclusion of their examination when they were sure about the type of simulated headache. The time required to reach a conclusion was noted in the experiment protocols by the experiment supervisor.

The subjects repeated the same diagnosis task on different predetermined sets of simulated conditions during the second and third sessions. The sessions were concluded either when a time limit of 30 minutes was reached, or when the subject completed three headache diagnoses. At the end of the third session, the subject was asked to fill out a questionnaire.

4 Results

Performance of a speech translation system is best evaluated by looking at system performance as a whole, and not separately for each subcomponent in the systems processing pipeline (Rayner et. al. 2000, pp. 297-pp. 312). In this paper, we consequently focus our analysis on objective and subjective usability-oriented measures.

In Section 4.1, we present objective usability measures obtained by analyzing user-system interactions and measuring task performance. In Section 4.2, we present subjective usability figures and a preliminary analysis of translation quality.

4.1 Objective Usability Figures

4.1.1 Analysis of User Interactions

Most of our analysis is based on data from the MedSLT system log, which records all interactions between the user and the system. An interaction is initiated when the user presses the "Start Recognition" button. The system then attempts to recognize what the user says. If it can do so, it next attempts to show the user how it has interpreted the recognition result, by first translating it into the Interlingua, and then translating it back into the source language (in this case, French). If the user decides that the back-translation is correct, they press the "Translate" button. This results in the system attempting to translate the Interlingua representation into the target language (in this case, English), and speak it using a Text-To-Speech engine. The system also displays a list of "help sentences", consisting of examples that are known to be within coverage, and which approximately match the result of performing recognition with the statistical language model. The user has the option of choosing a help sentence from the list, using the mouse, and submitting this to translation instead.

We classify each interaction as either "successful" or "unsuccessful". An interaction is defined to be unsuccessful if either

i) the user re-initiates recognition without asking the system for a translation, or
ii) the system fails to produce a correct translation or back translation.

Our definition of "unsuccessful interaction" includes instances where users accidentally press the wrong button (i.e. "Start Recognition" instead of "Translate"), press the button and then say nothing, or press the button and change their minds about what they want to ask half way through. We observed all of these behaviors during the tests.

Interactions where the system produced a translation were counted as successful, irrespective of whether the translation came directly from the user's spoken input or from the help list. In at least some examples, we found that when the translation came from a help sentence it did not correspond directly to the sentence the user had spoken; to our surprise, it could even be the case that the help sentence expressed the directly opposite question to the one the user had actually asked. This type of interaction was usually caused by some deficiency in the system, normally bad recognition or missing coverage. Our informal observation, however, was that, when this kind of thing happened, the user perceived the help module positively: it enabled them to elicit at least some information from the patient, and was less frustrating than being forced to ask the question again.

Table I to Table III show the number of total interactions per session, the proportion of successful interactions, and the proportion of interactions completed by selecting a sentence from the help list. The total number of interactions required to complete a session decreased over the three sessions, declining from an average of 98.6 interactions in the first session to 63.4 in the second (36% relative) and 53.9 in the third (45% relative). It is interesting to note that interactions involving the help system did not decrease in frequency, but remained almost constant over the first two sessions

(15.5% and 14.0%), and were in fact most common during the third session (21.7%).

Session 1			
Subject	**Interactions**	**% Successful**	**% Help**
User 1	57	56.1%	0.0%
User 2	98	52.0%	25.5%
User 3	91	63.7%	15.4%
User 4	156	69.9%	10.3%
User 5	86	64.0%	22.1%
User 6	134	47.0%	19.4%
User 7	56	53.6%	5.4%
User 8	111	63.1%	26.1%
AVG	98.6	58.7%	15.5%

Table I Total interaction rounds, percentage of successful interactions, and interactions involving the help system by subject for the 1st session

Session 2			
Subject	**Interactions**	**% Successful**	**% Help**
User 1	50	74.0%	2.0%
User 2	63	55.6%	27.0%
User 3	34	88.2%	23.5%
User 4	96	57.3%	17.7%
User 5	64	65.6%	21.9%
User 6	93	68.8%	10.8%
User 7	48	60.4%	4.2%
User 8	59	79.7%	5.1%
AVG	*63.4*	*68.7%*	*14.0%*

Table II Total interaction rounds, percentage of successful interactions, and interactions involving the help system by subject for the 2nd session

Session 3			
Subject	**Interactions**	**% Successful**	**% Help**
User 1	33	90.9%	33.3%
User 2	57	56.1%	22.8%
User 3	48	72.9%	29.2%
User 4	67	70.2%	16.4%
User 5	68	73.5%	27.9%
User 6	60	70.0%	6.7%
User 7	41	65.9%	14.6%
User 8	57	56.1%	22.8%
AVG	*53.9*	*69.5%*	*21.7%*

Table III Total interaction rounds, percentage of successful interactions, and interactions involving the help system by subject for the 3rd session

In order to establish a performance baseline, we also analyzed interaction data for two expert users, who performed the same experiment. The expert users were two native French-speaking system developers, which were both familiar with the diagnosis domain. Table IV summarizes the results of those users. One of our expert users, listed as Expert 2, is the French grammar developer, and had no failed interactions. This confirms that recognition is very accurate for users who know the coverage.

Session 1 / Expert Users			
Subject	**Interactions**	**% Successful**	**% Help**
Expert 1	36	77.8%	13.9%
Expert 2	30	100.0%	3.3%
AVG	*33*	*88.9%*	*8.6%*

Table IV Number of interactions, and percentages of successful interactions, and interactions involving the help component

The expert users were able to complete the experiment using an average of 33 interaction rounds. Similar performance levels were achieved by some subjects during the second and third session, which suggests that it is possible for at least some new users to achieve performance close to expert level within a few sessions.

4.1.2 Task Level Performance

One of the important performance indicators for end users is how long it takes to perform a given task. During the experiments, the instructors noted completion times required to reach a definite diagnosis in the experiment log. Table VI shows task completion times, categorized by session (columns) and task within the session (rows).

	Session 1	**Session 2**	**Session 3**
Diagnosis 1	17:00 min	11:00 min	7:54 min
Diagnosis 2	11:00 min	6:18 min	5:34 min
Diagnosis 3	7:54 min	4:10 min	4:00 min

Table V Average time required by subjects to complete diagnoses

In the last two sessions, after subjects had acclimatized to the system, a diagnosis takes an average of about four minutes to complete. This compares to a three-minute average required to complete a diagnosis by our expert users.

4.1.3 System coverage

Table VI shows the percentage of in-coverage sentences uttered by the users on interactions that did not involve invocation of the help component.

	IN-COVERAGE SENTENCES
Session 1	54.9%
Session 2	60.7%
Session 3	64.6%

Table VI Percentage of in-coverage sentences

This indicates that subjects learn and adapt to the system coverage as they use the system more. The average proportion of in-coverage utterances is 10 percent higher during the third session than during the first session.

4.2 Subjective Usability Measures

4.2.1 Results of Questionnaire

After finishing the third session, subjects were asked to fill in a short questionnaire, where responses were on a five-point scale ranging from 1 ("strongly disagree") to 5 ("strongly agree"). The results are presented in Table VIII.

STATEMENT	SCORE
I quickly learned how to use the system.	4.4
System response times were generally satisfactory.	4.5
When the system did not understand me, the help system usually showed me another way to ask the question.	4.6
When I knew what I could say, the system usually recognized me correctly.	4.3
I was often unable to ask the questions I wanted.	3.8
I could ask enough questions that I was sure of my diagnosis.	4.3
This system is more effective than non-verbal communication using gestures.	4.3
I would use this system again in a similar situation.	4.1

Table VIII Subject responses to questionnaire. Scores are on a 5-point scale, averaged over all answers.

Answers were in general positive, and most of the subjects were clearly very comfortable with the system after just an hour and a half of use. Interestingly, even though most of the subjects answered "yes" to the question "I was often unable to ask the questions I wanted", the good performance of the help system appeared to compensate adequately for missing coverage.

4.2.2 Translation Performance

In order to evaluate the translation quality of the newly developed French-to-English system, we conducted a preliminary performance evaluation, similar to the evaluation method described in (Bouillon 2005).

We performed translation judgment in two rounds. In the first round, an English-speaking judge was asked to categorize target utterances as comprehensible or not without looking at corresponding source sentences. 91.1% of the sentences were judged as comprehensible. The remaining 8.9% consisted of sentences where the terminology used was not familiar to the judge and of sentences where the translation component failed to produce a sufficiently good translation. An example sentence is

- Are the headaches better when you experience dark room?

which stems from the French source sentence

- Vos maux de tête sont ils soulagés par obscurité?

In the second round, English-speaking judges, sufficiently fluent in French to understand source language utterances, were shown the French source utterance, and asked to decide whether the target language utterance correctly reflected the meaning of the source language utterance. They were also asked to judge the style of the target language utterance. Specifically, judges were asked to classify sentences as "BAD" if the meaning of the English sentence did not reflect the meaning of the French sentence. Sentences were categorized as "OK" if the meaning was transferred correctly and the sentence was comprehensible, but the style of the resulting English sentence was not perfect. Sentences were judged as "GOOD" when they were comprehensible, and both meaning and style were considered to be completely correct. Table VIII summarizes results of two judges.

	Good	OK	Bad
Judge 1	15.8%	73.80%	10.3%
Judge 2	46.6%	47.1%	6.3%

Table VIII Judgments of the quality of the translations of 546 utterances

It is apparent that translation judging is a highly subjective process. When translations were marked as "bad", the problem most often seemed to be related to lexical items where it was challenging to find an exact correspondence between French and English. Two common examples were "troubles de la vision", which was translated as "blurred vision", and "faiblesse musculaire", which was translated as "weakness". It is likely that a more careful choice of lexical translation rules would deal with at least some of these cases.

5 Summary

We have presented a first end-to-end evaluation of the MedSLT spoken language translation system. The medical students who tested it were all able to use the system well, with performance in some cases comparable to that of that of system developers after only two sessions. At least for the fairly simple type of diagnoses covered by our scenario, the system's performance appeared clearly adequate for the task.

This is particularly encouraging, since the French to English version of the system is quite new, and has not yet received the level of attention required for a clinical system. The robustness added by the help system was sufficient to compensate for that, and in most cases, subjects were able to find ways to maneuver around coverage holes and other problems. It is entirely reasonable to hope that performance, which is already fairly good, would be substantially better with another couple of months of development work.

In summary, we feel that this study shows that the conservative architecture we have chosen shows genuine potential for use in medical diagnosis situations. Before the end of 2006, we hope to have advanced to the stage where we can start initial trials with real doctors and patients.

Acknowledgments

We would like to thank Agnes Lisowska, Alia Rahal, and Nancy Underwood for being impartial judges over our system's results.

This work was funded by the Swiss National Science Foundation.

References

P. Bouillon, M. Rayner, N. Chatzichrisafis, B.A. Hockey, M. Santaholma, M. Starlander, Y. Nakao, K. Kanzaki, and H. Isahara. 2005. *A generic multilingual open source platform for limited-domain medical speech translation*. In Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT), Budapest, Hungary.

P. Bouillon, M. Rayner, B. Novellas, Y. Nakao, M. Santaholma, M. Starlander, and N. Chatzichrisafis. 2006. *Une grammaire multilingue partagée pour la reconnaissance et la génération*. In Proceedings of TALN 2006, Leuwen, Belgium.

M. Cohen, J. Giangola, and J. Balogh. 2004, Voice User Interface Design. Addison Wesley Publishing.

R. I. Kittredge. 2003. *Sublanguages and comtrolled languages*. In R. Mitkov, editor, The Oxford Handbook of Computational Linguistics, pages 430–447. Oxford University Press.

MedBridge, 2006. *http://www.medtablet.com/*. As of 15th March 2006.

MedSLT, 2005. http://sourceforge.net/projects/medslt/. As of 15th March 2006.

T. Mitamura. 1999. *Controlled language for multilingual machine translation*. In Proceedings of Machine Translation Summit VII, Singapore.

Phraselator, 2006. *http://www.phraselator.com*. As of 15 February 2006.

M. Rayner, B.A. Hockey, and J. Dowding. 2003. *An open source environment for compiling typed unification grammars into speech recognisers*. In Proceedings of the 10th EACL (demo track), Budapest, Hungary.

M. Rayner, N. Chatzichrisafis, P. Bouillon, Y. Nakao, H. Isahara, K. Kanzaki, and B.A. Hockey. 2005b. *Japanese speech understanding using grammar specialization*. In HLT-NAACL 2005: Demo Session, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

M. Rayner, P. Bouillon, N. Chatzichrisafis, B.A. Hockey, M. Santaholma,M. Starlander, H. Isahara,

K. Kankazi, and Y. Nakao. 2005a. *A methodology for comparing grammar-based and robust approaches to speech understanding*. In Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP), Lisboa, Portugal.

M. Rayner, D. Carter, P. Bouillon, V. Digalakis, and M. Wirén. 2000. The Spoken Language Translator, Cambridge University Press.

M. Rayner, N. Chatzichrisafis, P. Bouillon, Y. Nakao, H. Isahara, K. Kanzaki, and B.A. Hockey. 2005b. *Japanese speech understanding using grammar specialization*. In HLT-NAACL 2005: Demo Session, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

M. Rayner, B.A. Hockey, and P. Bouillon. 2006. *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*. CSLI Press, Chicago.

Regulus, 2006. *http://sourceforge.net/projects/regulus/*. As of 15 March 2006.

S-MINDS, 2006. *http://www.sehda.com/*. As of 15 March 2006.

M. Starlander, P. Bouillon, N. Chatzichrisafis, M. Santaholma, M. Rayner, B.A. Hockey, H. Isahara, K. Kanzaki, and Y. Nakao. 2005. *Practicing controlled language through a help system integrated into the medical speech translation system (MedSLT)*. In Proceedings of the MT Summit X, Phuket, Thailand