Connecting Medical Informatics and Bio-Informatics
R. Engelbrecht et al. (Eds.)
ENMI, 2005

Breaking the Language Barrier: Machine Assisted Diagnosis Using the Medical Speech Translator

Marianne Starlander[a], Pierrette Bouillon[a], Manny Rayner[a,b], Nikos Chatzichrisafis[a], Beth Ann Hockey[b], Hitoshi Isahara[c], Kyoko Kanzaki[c], Yukie Nakao[c], Marianne Santaholma[a]

[a] *University of Geneva, TIM/ISSCO, Geneva, Switzerland*
[b] *UCSC/NASA Ames Research Center, Moffett Field,California, USA*
[c] *NICT, Kyoto, Japan*

Abstract

In this paper, we describe and evaluate an Open Source medical speech translation system (MedSLT) intended for safety-critical applications. The aim of this system is to eliminate the language barriers in emergency situation. It translates spoken questions from English into French, Japanese and Finnish in three medical sub-domains (headache, chest pain and abdominal pain), using a vocabulary of about 250-400 words per sub-domain. The architecture is a compromise between fixed-phrase translation on one hand and complex linguistically-based systems on the other. Recognition is guided by a Context Free Grammar Language Model compiled from a general unification grammar, automatically specialised for the domain. We present an evaluation of this initial prototype that shows the advantages of this grammar-based approach for this particular translation task in term of both reliability and use.

Keywords
Medical informatics; Diagnosis, computer assisted; Natural language processing; speech translation.

1. Introduction

Language is crucial to medical diagnosis. During the initial evaluation of a patient in an emergency department, obtaining an accurate history of the chief complaint is of equal importance to the physical examination. However, this physician-patient communication is often made substantially difficult because of language barriers: in many parts of the world there are large recent immigrant populations that require medical care but are unable to communicate fluently in the local language.

One solution to this problem would be to use human translators. Unfortunately, trained interpreters are only too rarely available in emergency cases, as it is quite expensive to provide every hospital with medical interpretation resources (see [1] for a description of the situation in the USA). Most of the time doctors have to rely on improvised translators such

as relatives, acquaintances or hospital employees with no medical training that happen to speak the language in question. This situation can be dramatic for the patient. In particular, the study by Glenn Flores, Professor of Paediatrics at the Boston University Schools of Medicine, shows that errors of interpretation are often responsible for errors in diagnosis [2]. It is therefore crucial to find a reliable and cost-effective alternative to the more expensive solution of providing a pool of trained emergency interpreters for each hospital. Our system is designed to address this problem using *spoken machine translation.*

Designing a spoken translation system to obtain a detailed medical history would be well beyond state of the art if we had to build a general system capable of translating anything the doctor or patient might wish to say. The reason that the use of spoken translation technology is feasible is because what is actually needed in the emergency setting is more limited. Since medical histories traditionally are obtained through two-way physician-patient conversations that are mostly physician initiative, there is a pre-established limiting structure that we can follow in designing the translation system. Our starting point is that this limited structure allows a physician to successfully use one way translation to elicit and restrict the range of patient responses while still obtaining the necessary information.

Another helpful constraint is that examinations can be divided into smaller sub-domains based on symptom types, for example headaches, chest pains, abdominal pains, and so on. This gives the possibility of further constraining the range of utterances that needs to be recognized at any point in the dialogue. For example, standard examination questions about chest pain include intensity, location, duration, quality of pain, and factors that increase or decrease the pain. The answers to these questions can be successfully communicated by a limited number of one or two word responses (e.g. yes/no, left/right, numbers) or even gestures (e.g. nodding or shaking the head, pointing to an area of the body). The above observations suggest that this is a domain in which the constraints of the task are sufficient for a limited domain, one way spoken translation system to be a useful tool.

The aim of this paper is to describe an Open Source toolkit (MedSLT, [3]) supporting quick development of this kind of speech translation systems for limited emergency diagnosis sub-domains. Although most spoken translation systems use statistical speech recognition [4], we show that a grammar-based approach is more suitable for this kind of task. This approach produces greater speech recognition accuracy, which is more important in the medical setting than robustness.

2. The MedSLT system

MedSLT [3] is an Open Source project which is developing a generic platform for building medical speech translation systems; early versions are described in [5], [6]. The basic philosophy behind the MedSLT system architecture is to attempt an intelligent compromise between fixed-phrase translation on one hand (e.g. (Phraselator [7]) and linguistically motivated grammar-based processing on the other (e.g. Verbmobil [8]) and Spoken Language Translator [9].

At run-time, the system behaves essentially like a phrasal translator which allows some variation in the input language. This is close in spirit to the approach used in most normal phrase-books, which typically allow "slots" in at least some phrases ("How do I get to---?"). However, in order to minimize the overhead associated with defining and maintaining large sets of phrasal patterns, these patterns are derived from a single large linguistically motivated unification grammar, using the Open Source Regulus platform [6], [10], which implements an example-based specialisation method driven by small corpora of examples.

The linguistically motivated compile-time architecture makes the system easy to extend and modify. In particular, it makes it easy to port the grammar between different medical

sub-domains, which seem to be quite convergent. For example, the first version of the system covered only the headache sub-domain; when we ported the English grammar to the new chest pain sub-domain, over 80% of the training sentences could be analysed correctly as soon as we had added the relevant new vocabulary.

The translation module is implemented in SICStus Prolog, and is interlingua-based. Translation consists of four stages illustrated in Figure 1: (1) mapping from the source representation to interlingua; (2) ellipsis processing; (3) mapping from interlingua to the target representation and (4) generation, using a suitably compiled Regulus grammar for the target language. In accordance with the generally minimalist design philosophy of the project, semantic representations have been kept as simple as possible, namely a flat list of attribute-value pairs.

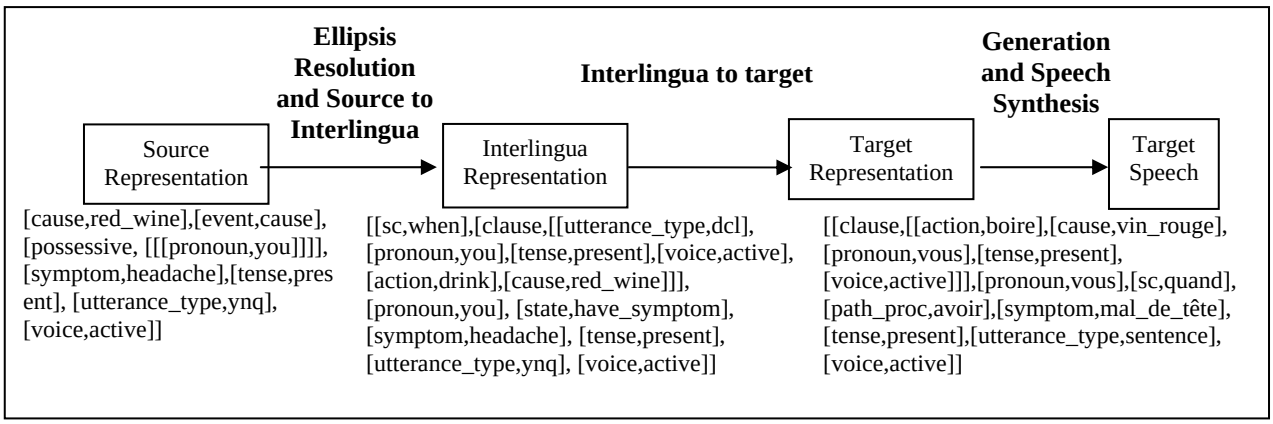

Figure 1: Translation flow in MedSLT for the source sentence: "Does red wine cause your headaches"
Target: "Avez-vous vos maux de tête quand vous buvez du vin rouge"

The run time system provides a GUI based interface, which allows the user to select the input and output languages and the sub-domain. The user initiates speech recognition through a push-to-talk interface.

Both the string of words produced by the speech recognizer (what the system heard) and a back-translation from the interlingua (what the system understood), are displayed on the screen, after which the user can choose either to proceed with the translation or to abort in the case of apparently unsuccessful speech understanding. Output speech is produced using either pre-recorded waveforms or the Nuance VocalizerTM TTS engine, depending on the output language. It is important to realise that the recognised words and the back-translation can be quite different, particularly when translating incomplete utterances. For example, if the previous utterance was “Do the headaches typically last more than an hour?”, the follow-on question “More than two hours?” would be back-translated as “Do the headaches usually last more than two hours?”.

If the system is unable to produce any translation, the system invokes a simple context-sensitive help module. This uses the result of traditional recognition using a statistical language model (SLM) to display a list of in-coverage example sentences. Examples are selected from a predefined list, using a heuristic which prioritises sentences maximizing the number of words and phrases shared with those extracted from the SLM recognition result.

The current prototype covers the three subdomains of headaches, chest pains, and abdominal pains, and can translate from English into French, Japanese and Finnish. The Finnish version is still under development, and is not as mature as the other two. Initial versions supporting translation from Japanese, French and Spanish and to Spanish are also available. Some examples of English-French translations are given in Table 1.

Table 1: Examples of collected data and translation by MedSLT

Source language	**Target language**
'Do sudden head movements make the pain worse?'	'La douleur empire-t-elle quand vous bougez soudainement la tête ?'
'Is your headache relieved by sleep?'	'Vos maux de tête s'améliorent-ils quand vous dormez ?'

3. Evaluation

In the long-term, the real question we would like to answer when evaluating the prototype is whether this system is practically useful for doctors. As a first step, we compare the performance of the grammar-based architecture on the medical examination task with that of a second version of the system. This version was built using a statistical language model (SLM) created with the help of the Nuance SayAnything tool [11], and trained on the same data as the GLM version. In the following sections, we present a comparison of these two versions of the MedSLT system. It is striking to see that the two systems give nearly the same results on the training data, but that when judged on a real task the GLM version clearly outperforms the SLM. We used the headache version of the Open Source MedSLT system [3] to perform our experiments. Both versions of the recogniser were trained from the same corpus of 575 standard examination questions put together by Dr. Vol Van Dalsem III[1].

We collected data from 12 native speakers of English. Each subject was first given a short acclimatization session, where they used a prepared list of ten in-coverage sentences to learn how to use the microphone and the push-to-talk interface. They were then encouraged to play the part of a doctor, and conduct an examination interview, through the system, on a team member who simulated a patient suffering from a specific type of headache. The subject's task was to identify the type correctly out of a list of eight possibilities. Half of the subjects used the grammar-based version of the system, and half used the SLM based version. We collected a total of 870 recorded utterances.

The recorded data was first transcribed, and then processed through offline versions of both the grammar-based (GLM) and statistical SLM processing paths in the system. This was done as follows. We first set the system to translate from English into English (via the interlingua), and then had an English-speaking judge evaluate each back-translation. Utterances for which the back-translation was judged acceptable were regarded as correctly recognised, and were then translated further into the target languages French and Japanese.

Translations to French and Japanese were judged for acceptability by native speaker judges for each language: there were six judges for French, and three for Japanese. Judges were asked to categorise translations as "good", "ok" or "bad". For each target language, and each processing method (GLM or SLM), we consolidated the results using a majority voting scheme. If two-thirds of the judges (i.e. four for French, or two for Japanese) agreed that the translation was clearly "good" or "bad", we counted the translation as belonging to the appropriate category. Otherwise, we counted it as "ok". The results of this judging are shown below.

[1] El Camino Hospital, Mountain View, California.

Table 2: Recognition and Translation quality

		GLM	SLM		GLM	SLM	
French	**Bad recognition**	54.6%	59.8%	**Japanese**	54.6%	59.8%	**Bad recognition**
	Good translation	34.4%	30.8%		36.4%	32.8%	**Good translation**
	Ok translation	8.7%	7.7%		3.6%	3.3%	**Ok translation**
	Bad translation	0.3%	0.2%		0.5%	0.5%	**Bad translation**
	No translation	2.0%	1.5%		4.9%	3.7%	**No translation**

4. Discussion

We will now attempt to draw some general conclusions about the relative performance of GLM and SLM processing in these experiments. As shown in Table 2, the GLM produces fewer recognition errors (54.6%) than the SLM (59.8%). Although these figures are quite high for both systems, these results should be interpreted in the light of further results (cf. [12]) clearly showing the difference in performance between the two recognisers on in-coverage data. The GLM (5.7%) clearly out-performs the SLM (12.7%) on in-coverage sentences measured in terms of WER. This pattern is inverted for the out of coverage sentences, where the SLM outscores the GLM by 47.8% to 57.5%.

This confirms our intuition that the SLM version is more robust than the GLM but that the GLM is more precise. If a sentence is in the coverage of the GLM, global constraints usually insure that the sentence is well recognised and translated. The extra robustness offered by the SLM does indeed result in a lower word error rate on the out-of-coverage data, but what counts in this type of medical speech translation task, where partial translations are worse than useless, is to achieve a correct output on in-coverage data.

The ratio of in-coverage to out-of-coverage in the dataset is however mainly a function of how familiar the subjects are with the system's coverage. An experienced user will produce mostly in-coverage data; a novice user will produce mostly out-of-coverage data.

Table 3: Learning effect: improvement of recognition quality.

		GLM	SLM			GLM	SLM
Help system **OFF**	**All data**	54.0%	61.1%	Help system **ON**	**All data**	55.3%	58.4%
	First quarter	58.6%	65.8%		**First quarter**	63.1%	64.1%
	Last quarter	52.1%	58.1%		**Last quarter**	45.9%	56.0%
	Improvement	6.5%	7.7%		**Improvement**	17.2%	8.1%

A critical point is thus the capacity of the subjects to improve their performance with increased familiarity. This improvement is especially noticeable for the subjects using the GLM system: as people become more expert, they gravitate towards the intended coverage, and robustness becomes less important. We can get some idea of what's happening here by contrasting performance for the two architectures averaged over the first and last quarters of each session. Table 3 presents the recognition scores for the GLM and SLM comparing the start of a session to the end of the session. A lot of the improvement seems to be due to the help system. Subjects with access to the help system improved much more between the first quarter session and the last. The difference in improvement only occurs if the GLM system is being used, not surprisingly since the help system is steering users towards the grammar's coverage.

5. Conclusion

We have described an approach to automatic limited medical speech translation, and

compared two different architectures. We conclude that a grammar-based architecture is more effective for the task, particularly when combined with the inclusion of an intelligent help component. With the grammar-based architecture and the help component, Table 3 shows a dramatic improvement in subjects' performance even over short sessions averaging 60 utterances in length. These results, and other informal studies we have conducted, suggest that a new user would be able to achieve a useful level of performance after only a few hours of practice with the system. Within the next few months, we hope to be able to gain more data on the system's utility as we begin to test it in a simulated emergency room setting.

6. Acknowledments

The MedSLT project is funded by the Fonds National de la Recherche Suisse (FNRS) and the Japanese National Institute of Information and Communications Technology. Finally we would like to thank Vol Van Dalsem for his expert advice.

7. References

[1] Loviglio, J. Interpreters Lower Risks in Hospitals, Article by Associated Press writer published on various internet news sites on 21 November 2004

[2] Flores G, Laws B, Mayo S, Zuckerman B, Abreu M, Medina L, and Hardt EJ. Errors in medical interpretation and their potential clinical consequences in pediatric encounters. *Pediatrics* 2003:111 pp: 6-14.

[3] MedSLT, http://sourceforge.net/projects/medslt/. As of 12 January 2005.

[4] Akiba Y, Federico M, Kando N, Nakaiwa H, Paul M and Tsujii J. Overview of the IWSLT04 Evaluation Campaign. In: *Proceedings of the International Workshop on Spoken Language Translation IWSLT04*, Kyoto, Japan, pp. 1-12.

[5] Rayner M, Bouillon P. A flexible Speech to Speech Phrasebook Translator. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics Workshop on Speech-to-Speech Translation: Algorithms and Systems*, Philadelphia, PA, 2002, pp. 69-76.

[6] Rayner M, Hockey BA and Dowding J. An Open Source Environment for Compiling Typed Unification Grammars into Speech Recognisers. In: *Proceedings of the 10th European Association for Computational Linguistics (demo track)*, Budapest, Hungary, 2003, pp. 223-226.

[7] Phraselator, http://www.phraselator.com. As of 12 January 2005.

[8] Wahlster W. *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, 2000.

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

[10] Regulus, http://sourceforge.net/projects/regulus/. As of 12 January 2005.

[11] Nuance, http://www.nuance.com. As of 12 January 2005.

[12] Bouillon P, Rayner M, Chatzichrisafis N, Hockey BA, Santaholma M, Starlander M, Nakao Y, Kanzaki K, Isahara H. A Generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation. In: Proceedings of the 10th International Conference of the European Association for Machine Translation, Budapest, Hungary, 2005.

Address for correspondence

Marianne Starlander, University of Geneva, ETI/TIM/ISSCO, 40, bd. du Pont d'Arve, 1211 Genève 4,
Tel. ++41.22.379.86.78 Email. Marianne.starlander@eti.unige.ch http://www.issco.unige.ch/projects/medslt/