Comparing two different bidirectional versions of the limited-domain medical spoken language translator MedSLT

Marianne Starlander[1], Pierrette Bouillon[1], Glenn Flores[2], Manny Rayner[1], Nikos Tsourakis[1]

(1) University of Geneva, TIM/ISSCO/ETI

{Pierrette.Bouillon, Emmanuel.Rayner, Nikolaos.Tsourakis}@issco.unige.ch, Marianne.Starlander@eti.unige.ch

(2) UT Southwestern Medical Center, Children's Medical Center of Dallas

Glenn.Flores@UTsouthwestern.edu

Abstract. This paper reports preliminary results of an evaluation during which two different bidirectional versions of the limited-domain medical spoken language translator MedSLT were compared in a hospital setting. The more restricted version (V.1) only allows Yes-No answers and short elliptical sentences, while the less restricted version (V.2) allows Yes-No answers, short elliptical sentences and full sentences. Although WER is marginally better for V.1, task performance is marginally worse. There appear to be two main reasons for this disparity; short sentences are often badly recognised and patients tend to find it difficult to limit themselves to ellipsis, even if they receive clear instructions about not using full sentences.

1. Introduction

Today, an ideal speech-to-speech translation system enabling a healthcare provider to communicate naturally with a patient across two different languages during a medical encounter is still well beyond what is possible given the state of the art. It is thus necessary to find good tradeoffs, and develop methodologies to build and objectively compare different currently realisable architectures.

In this paper, we compare two different versions of the bidirectional MedSLT system, a medical speech to speech translator, which differ in terms of grammatical coverage. In the first version (V.1), the patient (Pat) can only answer the physician's (Phy) question using elliptical answers (Example 1). The second (V.2), less restricted version, enables the patient to answer with either elliptical answers or full sentences (Example 2).

(1) **Phy:** Where is the pain?
Tran: ¿dónde le duele?
Pat: en la garganta
Trans: I have a sore throat
Phy: For how long have you had a sore throat?
Trans: ¿desde cuándo le duele la garganta?
Pat: Desde hace más de una semana
Trans: I have had a sore throat for more than one week

Example 1. Dialogue sample with Version 1

(2) **Phy:** Where is the pain?
Tran: ¿dónde le duele?
Pat: Me duele la garganta
Trans: I have a sore throat
Phy: How long have you had a sore throat?
Trans: ¿desde cuándo le duele la garganta?
Pat: Me duele desde hace dos días
Trans: I have experienced the pain for two days

Example 2. Dialogue sample with Version 2

The aim of this evaluation is to determine the impact of the restriction in coverage, both in terms of objective system metrics (WER) and also user-oriented performance metrics (task completion time).

In the rest of the paper, we first give a quick overview of the MedSLT system (Section 2). We then explain how the two versions of the

system can be built using the Regulus platform (Section 3). Finally, we provide some preliminary results from an evaluation comparing the two versions of the system, which was undertaken at Children's Medical Center Dallas (Section 4).

2. MedSLT

MedSLT is a medium-vocabulary speech translation system intended to support medical diagnosis dialogues between a physician and a patient who do not share a common language (Bouillon et al., 2005). The topic of conversation is assumed to be limited to a specific medical sub-domain, defined by a related set of symptoms. Typical examples are headaches or chest pains. The architecture has been designed with the following key goals in mind:

- Given the safety-critical nature of the task, precision is more important than recall.
- It should be easy to adapt the core system to new languages and domains.
- The user should be able to adapt to the limitations of the system's coverage with a minimum of training.

The first goal has oriented us towards an architecture that is primarily rule-based, and thus more readily predictable in terms of function, though we also use statistical tuning methods to increase efficiency. The speech recognition component uses the Nuance 8.5 platform (Nuance, 2003), equipped with grammar-based language models. Translation is interlingua-based (Bouillon et al., 2008).

One of the system's distinguishing characteristics, compared to related work (Dillinger and Seligman, 2006; Ehsani et al., 2006; Mana et al., 2003; Schultz et al., 2004; Zhou et al., 2007), is that all grammars used, for recognition, analysis and generation, are compiled from a small number of general linguistically-motivated unification grammars, using the Open Source Regulus platform (Rayner et al., 2006). Early versions of the system used a single core grammar per language; more recent ones have gone further, and merged together grammars for closely related languages (Bouillon et al., 2006b). These core grammars are automatically specialised, using corpus-driven methods based on small corpora, to derive simpler grammars. Specialisation will typically be along all of the following dimensions: task (recognition, analysis, generation), sub-domain (headache, chest pain, etc), and context (physician question, patient response). Each of these specialised unification grammars is then subjected to a second compilation step, which converts it into its executable form. For analysis and generation, this form is a standard parser or generator. For recognition, it is a semantically annotated CFG grammar in the form required by the Nuance engine, which is then subjected to further Nuance-specific compilation steps to derive a speech recognition package. These final compilation steps include a second use of the training corpus to perform statistical tuning of the language model. The overall goal of the Regulus architecture is to simplify the normally very onerous task of writing and maintaining a large number of closely related grammars, retaining internal coherence between them. In particular, coherence between the recognition and analysis grammars guarantees that any spoken expression which is accepted by the recogniser can also be parsed.

Although performance of rule-based recognition systems is typically good on in-grammar coverage, a well-known problem is brittleness: users need to know what language the grammar supports. Our approach to this problem is to equip the system with an intelligent help module (Starlander et al., 2005) which after each utterance provides the user with in-coverage examples, chosen to be as close to the user's actual utterance as possible. The help module's output is based on a library of utterances which have already been evaluated as being within grammar coverage and producing correct translations. At runtime, the system carries out a second round of recognition with a backup statistical recogniser, and uses the result to select examples from the library which are similar to the statistical recogniser's result in terms of a backed-off N-gram metric. On the patient side, additional information is provided: the help module first searches the help library to find the stored question most similar to the current one, and

then shows a predefined list of possible answers associated with it.

The user interacts with the system through a push to talk interface (Figure 1). In order to address the issue of reliability, the recognised sentence is always back-translated into the source language.

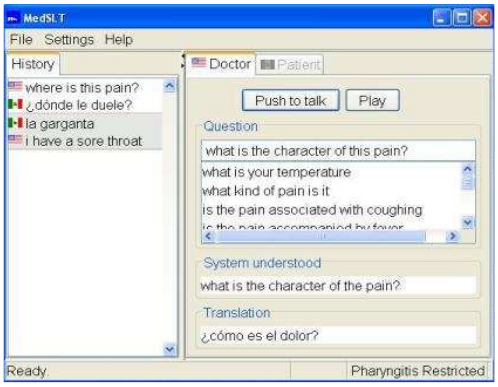

Fig. 1. MedSLT bidirectional screenshot – V1

The intention is that this back-translation ("System understood" in Figure 1) should be checked by the user before the translation is sent. Since we are using an Interlingua approach to translation, it is reasonable to assume that, if the sentence is correctly translated back to the source language, it will be equally well translated into the target language. Evaluations show that this is indeed the case more than 98% of the time.

3. Two bidirectional versions

The two bidirectional versions we evaluated (more restricted and less restricted) were developed for the English to Spanish language pair, in the sore throat domain. Since both recognition and translation are rule-based, each version uses six specialised grammars:

- An English recognition grammar for the physician's questions
- A Spanish recognition grammar for the patient's answers
- An English generation grammar for the translation of the Spanish patient's answers
- A Spanish generation grammar for the translation of the English physician's questions
- An English generation grammar for the back-translation of the physician's questions
- A Spanish generation grammar for the back-translation of the patient's answers

Writing these 12 grammars by hand would have been extremely tedious and error-prone. Deriving them from general unification grammars with the help of the Regulus platform was, however, quite straightforward. To produce the different versions, we only need to vary three criteria. For each specialised grammar we (1) specify the general grammar used, either English for the physician's side or Romance for the patient's side (Bouillon et al., 2006a), (2) define a short training corpus from which the specific coverage is learned, and (3) define operationality criteria. The corpora give at least one example of each type of structure or word that should be included in the different specialised grammars. The operationality criteria specify which constituents should be included in the resulting grammar and how they should be organised. Together, the corpus and the operationality criteria make it possible to take into account essential differences between recognition and generation grammars, in order to produce an appropriate grammar for each task and domain. For example, the recognition grammars need to be general enough to cover different combinations of nouns and specifiers ("when do you have a headache?", "when do you have your headaches?", "when do you have headaches?" etc). The operationality criteria are thus defined so as to generalise all the specifiers and nouns from the examples in the corpus (spec → {a, your, etc.}; noun → {headaches, headache}) and to learn complex NP rules (np → spec, noun). The generation grammars, on the other hand, need to be more constrained, and generate only a preferred specifier-noun pair for each particular domain ("when do you have headaches?"). The criteria here are defined to learn fixed noun phrases from the generation corpus (in our example, np → headaches).

The main advantage of the Regulus platform for this study is that it is possible to get easily comparable grammars for the two bidirectional versions. Since the two bidirectional versions under study here mainly differed with respect to

the patient recognition grammar, it was sufficient to construct two different versions of the patient recognition training corpus (141 corpus entries for V.1, and 453 entries for V.2). The more restricted version is specialised to recognise only elliptical sentences (incomplete sentences), while the less restricted version can recognise both full and incomplete sentences. For Spanish, the specialised version for the less restricted version (allowing patients to answer with full sentences) does not contain any rule for interrogative sentences. An utterance like "Tengo fiebre" will thus only be analysed as "I have a fever" and never as "Do I have a fever?"

	UG rules	**Words**
V.1	38	100
V.2	132	228

Table 1. Size of the Spanish specialised recognition grammars for V.1 and V.2

Table 1 shows the number of unification grammar rules and vocabulary items for Spanish after respectively specialising for V.1 and V.2.

4. Evaluation and conclusion

In order to compare the two versions of the system, we organised a data collection with physicians and standardised patients at Children's Medical Center Dallas. The standardised patients were professional medical Spanish interpreters from Children's Medical Center Dallas trained for a specific task. The aim of the task was to determine whether the patient suffered from a bacterial infection (streptococcal pharyngitis) or not. Eight physicians and 16 patients participated. We asked the standardised patients to simulate a history and physical findings consistent with viral pharyngitis or streptococcal pharyngitis, using eight different carefully scripted, fixed scenarios. None of the participants had used the system before. Each standardised patient used both versions of the system with two different physicians. Half of the standardised patients started with the more constrained version (V.1) and half with the less restricted version (V.2).

Our expectation when setting up the study was that the more restricted bidirectional version (V.1) would give better results, for at least two reasons. The sentences that need to be recognised in V.1 are shorter, which we thought would help recognition, and the vocabulary and syntax are more limited, which we assumed would help the patients remain within the system's coverage. However, the evaluation in fact suggested the contrary, which could also explain why in the satisfaction questionnaire both physicians and patients were clearly in favour of the less restricted version (V.2). Although the word error rate (WER) is, as expected, slightly better for the more restricted version (V.1), as shown in Table 2, communication seems more successful with the less restricted version (V.2).

	English	**Spanish**
V.1	29.12%	29.37%
V.2	31.99%	32.44%

Table 2. Word error rate

The findings in Table 3 indicate that on one hand the number of interactions and the mean time taken to achieve a diagnosis are slightly higher for the more restricted version (V.1).

	V.1	**V.2**
Total number of utterances	967	943
Average time in minutes	8	7

Table 3. Number of utterances and time to diagnosis by system

On the other hand, the number of recognised sentences accepted by the participants on the basis of the back-translation, and thus sent to translation is slightly higher with the less restricted version (V.2), both on the patient's and the physician's sides (Table 4).

	V.1	**Confidence interval**	**V.2**
Pat	70.5	(67.14, 74.02)	72.3
Phys	72.5	(69.19, 75.82)	75.67
All	71.56	(69.17, 73.95)	74.55

Table 4. Recognition performance in terms of % of accepted sentences by the users

Regardless of the statistically significance of the figures in the table above, a thorough examination of badly recognised sentences which were not accepted by the users indicates that each version has its own set of limitations. The more restricted version (V.1) seems to encounter two major challenges that look

difficult to address. First, very short sentences are often badly recognised (26 out of 140 unaccepted sentences are misrecognitions for the elliptical answers “mucho” and “un poco”). This is consistent with the results of (Bouillon et al., 2007) which studied the impact of ellipsis on recognition and showed that they tend to increase the WER. Second, patients tend to find it difficult to limit themselves to ellipsis, even with the help system, and despite the fact that they received clear instructions for doing so – 44/140 misrecognised sentences were in fact complete sentences that were outside the more restricted version’s (V.1) coverage and thus could not be recognised.

By contrast, in the less restricted version (V.2) the types of badly recognised sentences are more diverse. Many of these are out of coverage formulations or vocabulary items that clearly need to be added to the system. Our preliminary conclusion is thus that the restrictions on the more limited version do not improve the system, and that it seems more feasible to further develop the less restricted version by expanding it. From experience, we know that several data collections will be needed before reaching a satisfactory level of coverage is achieved. We are therefore planning a second evaluation round that will concentrate on the less restricted version, once the coverage problems have been addressed.

5. References

BOUILLON, Pierrette, HALIMI, Sonia, NAKAO, Yukie, KANZAKI, Kyoko, ISAHARA, Hitoshi, TSOURAKIS, Nikos, STARLANDER, Marianne, HOCKEY, Beth Ann and RAYNER, Manny (2008). 'Developing Non-European Translation Pairs in a Medium-Vocabulary Medical Speech Translation System'. In Proceedings of the 6th International Conference on Language Resources and Evaluation.

BOUILLON, Pierrette, RAYNER, Manny, CHATZICHRISAFIS, Nikos, HOCKEY, Beth Ann, SANTAHOLMA, Marianne, STARLANDER, Marianne, ISAHARA, Hitoshi, KANZAKI, Kyoko and NAKAO, Yukie (2005). 'A generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation'. In Proceedings of the Tenth Conference of the European Association of Machine Translation, pp. 50-58.

BOUILLON, Pierrette, RAYNER, Manny, NOVELLAS VALL, Bruna, NAKAO, Yukie, SANTAHOLMA, Marianne, STARLANDER, Marianne and CHATZICHRISAFIS, Nikos (2006a). 'Une grammaire multilingue partagée pour la traduction automatique de la parole'. In Proceedings of the Traitement Automatique des Langues Naturelles, pp. 93-102.

BOUILLON, Pierrette, RAYNER, Manny, NOVELLAS VALL, Bruna, STARLANDER, Marianne, SANTAHOLMA, Marianne, NAKAO, Yukie and CHATZICHRISAFIS, Nikos (2006b). 'Une grammaire partagée multi-tâche pour le traitement de la parole : application aux langues romanes'. Revue TAL, 47(3), pp. 155-173.

BOUILLON, Pierrette, RAYNER, Manny, STARLANDER, Marianne and SANTAHOLMA, Marianne (2007). 'Les ellipses dans un système de Traduction Automatique de la Parole'. In Proceedings of the Traitement Automatique des Langues Naturelles, pp. 53-62.

DILLINGER, Mike and SELIGMAN, Mark (2006). 'Converser (TM): Highly Interactive Speech-to-Speech Translation for Healthcare'. In Proceedings of the Workshop on Medical Speech Translation at HLT-NACCL, pp. 40-43.

EHSANI, Farzad, KINZEY, Jim, MASTER, Demetrios, LESEA, Karen and PARK, Hunil (2006). 'Speech to speech Translation for Medical Triage in Korean'. In Proceedings of the Workshop on Medical Speech Translation at HLT-NACCL, pp. 17-23.

MANA, Nadia, BURGER, Susanne, CATTONI, Roldano, BESACIER, Laurent, MACLAREN, Victoria, MCDONOUGH, John and METZE, Florian (2003). 'The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domains'. In Proceedings of the Eurospeech, pp. 1589-1593.

Nuance (2003). Nuance.

RAYNER, Manny, HOCKEY, Beth Ann and BOUILLON, Pierrette (2006). 'Putting Linguistics into Speech Recognition'. Stanford, California: Stanford University Center for the Study of Language and Information.

SCHULTZ, Tanja, ALEXANDER, Dorcas, BLACK, Alan W., Peterson, Kay, Suebvisai, Sinaporn and Waibel , Alex (2004). 'A Thai speech translation system for medical dialogs'. In Proceedings of the Human Language Technologies (HLT), pp. 263-264.

STARLANDER, Marianne, BOUILLON, Pierrette, CHATZICHRISAFIS, Nikos, SANTAHOLMA, Marianne, RAYNER, Manny, HOCKEY, Beth Ann, ISAHARA, Hitoshi, KANZAKI, Kyoko and NAKAO, Yukie (2005). 'Practising Controlled Language through a Help System integrated into the Medical Speech Translation System (MedSLT)'. In Proceedings of the MT Summit X, pp. 188-194.

ZHOU, Bowen, BESACIER, Laurent and GAO, Yuqing (2007). 'On efficient coupling of ASR and SMT For Speech Translation'. In Proceedings of the International Conference on Accoustics, Speech and Signal Processing.