A Small-Vocabulary Shared Task for Medical Speech Translation

Manny Rayner[1], **Pierrette Bouillon**[1], **Glenn Flores**[2], **Farzad Ehsani**[3]
Marianne Starlander[1], **Beth Ann Hockey**[4], **Jane Brotanek**[2], **Lukas Biewald**[5]

[1] University of Geneva, TIM/ISSCO, 40 bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
{Emmanuel.Rayner,Pierrette.Bouillon}@issco.unige.ch
Marianne.Starlander@eti.unige.ch

[2] UT Southwestern Medical Center, Children's Medical Center of Dallas
{Glenn.Flores,Jane.Brotanek}@utsouthwestern.edu

[3] Fluential, Inc, 1153 Bordeaux Drive, Suite 211, Sunnyvale, CA 94089, USA
farzad@fluentialinc.com

[4] Mail Stop 19-26, UCSC UARC, NASA Ames Research Center, Moffett Field, CA 94035–1000
bahockey@ucsc.edu

[5] Dolores Labs
lukeab@gmail.com

Abstract

We outline a possible small-vocabulary shared task for the emerging medical speech translation community. Data would consist of about 2000 recorded and transcribed utterances collected during an evaluation of an English ↔ Spanish version of the Open Source MedSLT system; the vocabulary covered consisted of about 450 words in English, and 250 in Spanish. The key problem in defining the task is to agree on a scoring system which is acceptable both to medical professionals and to the speech and language community. We suggest a framework for defining and administering a scoring system of this kind.

1 Introduction

In computer science research, a "shared task" is a competition between interested teams, where the goal is to achieve as good performance as possible on a well-defined problem that everyone agrees to work on. The shared task has three main components: training data, test data, and an evaluation metric. Both test and training data are divided up into sets of items, which are to be processed. The evaluation metric defines a score for each processed item. Competitors are first given the training data, which they use to construct and/or train their systems. They are then evaluated on the test data, which they have not previously seen.

In many areas of speech and language processing, agreement on a shared task has been a major step forward. Often, it has in effect created a new subfield, since it allows objective comparison of results between different groups. For example, it is very common at speech conference to have special sessions devoted to recognition within a particular shared task database. In fact, a conference without at least a couple of such sessions would be an anomaly. A recent success story in language processing is the Recognizing Textual Entailment (RTE) task[1]. Since its inception in 2004, this has become extremely popular; the yearly RTE workshop now attracts around 40 submissions, and error rates on the task have more than halved.

Automatic medical speech translation would clearly benefit from a shared task. As was made apparent at the initial 2006 workshop in New York[2], nearly every group has both a unique architecture and a unique set of data, essentially making comparisons impossible. In this note, we will suggest an initial small-vocabulary medical

[1] http://www.pascal-network.org/Challenges/RTE/

[2] http://www.issco.unige.ch/pub/SLT_workshop_proceedings_book.pdf

Coling 2008: Proceedings of the workshop on Speech Processing for Safety Critical Translation and Pervasive Applications, pages 60–63
Manchester, August 2008

shared task. The aspect of the task that is hardest to define is the evaluation metric, since there unfortunately appears to be considerable tension between the preferences of medical professionals and speech system implementers. Medical professionals would prefer to carry out a "deep" evaluation, in terms of possible clinical consequences following from a mistranslation. System evaluators will on the other hand prefer an evaluation method that can be carried out quickly, enabling frequent evaluations of evolving systems. The plan we will sketch out is intended to be a compromise between these two opposing positions.

The rest of the note is organised as follows. Section 2 describes the data we propose to use, and Section 3 discusses our approach to evaluation metrics. Section 4 concludes.

2 Data

The data we would use in the task is for the English ↔ Spanish language pair, and was collected using two different versions of the MedSLT system[3]. In each case, the scenario imagines an English-speaking doctor conducting a verbal examination of a Spanish-speaking patient, who was assumed to be have visited the doctor because they were displaying symptoms which included a sore throat. The doctor's task was to use the translation system to determine the likely reason for the patient's symptoms.

The two versions of the system differed in terms of the linguistic coverage offered. The more restricted version supported a minimal range of English questions (vocabulary size, about 200 words), and only allowed the patient to respond using short phrases (vocabulary size, 100 words). Thus for example the doctor could ask "How long have you had a sore throat?", and the patient would respond *Hace dos días* ("for two days"). The less restricted version supported a broader range of doctor questions (vocabulary size, about 450 words), and allowed the patient to respond using both short phrases and complete sentences (vocabulary size, about 225 words). Thus in response to "How long have you had a sore throat?", the patient could say either *Hace dos días* ("for two days") or *Tengo dolor en la garganta hace dos días* ("I have had a sore throat for two days").

Data was collected in 64 sessions, carried out over two days in February 2008 at the University of Texas Medical Center, Dallas. In each session, the part of the "doctor" was played by a real physician, and the part of the "patient" by a Spanish-speaking interpreter. This resulted in 1005 English utterances, and 967 Spanish utterances. All speech data is available in SPHERE-headed form, and totals about 90 MB. A master file, organised in spreadsheet form, lists metadata for each recorded file. This includes a transcription, a possible valid translation (verified by a bilingual translator), IDs for the "doctor", the "patient", the session and the system version, and the preceding context. Context is primarily required for short answers, and consists of the most recent preceding doctor question.

3 Evaluation metrics

The job of the evaluation component in the shared task is to assign a score to each translated utterance. Our basic model will be the usual one for shared tasks in speech and language. Each processed utterance will be assigned to a category; each category will be associated with a specified score; the score for a complete testset will the sum of the scores for all of its utterances. We thus have three sub-problems: deciding what the categories are, deciding how to assign a category to a processing utterance, and deciding what scores to associate with each category.

3.1 Defining categories

If the system attempts to translate an utterance, there are *a priori* three things that can happen: it can produce a correct translation, an incorrect translation, or no translation. Medical speech translation is a safety-critical problem; a mistranslation may have serious consequences, up to and including the death of the patient. This implies that the negative score for an incorrect translation should be high in comparison to the positive score for a correct translation. So a naive scoring function might be "1 point for a correct translation, 0 points for no translation, –1000 points for an incorrect translation."

However, since the high negative score for a mistranslation is justified by the possible serious consequences, not all mistranslations are equal; some are much more likely than others to result in clinical consequences. For example, consider the possible consequences of two different mistrans-

[3] http://www.issco.unige.ch/projects/medslt/

lations of the Spanish sentence *La penicilina me da alergias*. Ideally, we would like the system to translate this as "I am allergic to penicillin". If it instead says "I am allergic to *the* penicillin", the translation is slightly imperfect, but it is hard to see any important misunderstanding arising as a result. In contrast, the translation "I am *not* allergic to penicillin", which might be produced as the result of a mistake in speech recognition, could have very serious consequences indeed. (Note in passing that both errors are single-word insertions). Another type of result is a nonsensical translation, perhaps due to an internal system error. For instance, suppose the translation of our sample sentence were "The allergy penicillin does me". In this case, it is not clear what will happen. Most users will probably dismiss the output as meaningless; a few might be tempted to try and decipher it, with unpredictable results.

Examples like these show that it is important for the scoring metric to differentiate between different classes of mistranslations, with the differentiation based on possible clinical consequences of the error. For similar reasons, it is important to think about the clinical consequences when the system produces correct translations, or fails to produce a translation. For example, when the system correctly translates "Hello" as *Buenas días*, there are not likely to be any clinical consequences, so it is reasonable to reward it with a lower score than the one assigned to a clinically contentful utterance. When no translation is produced, it also seems correct to distinguish the case where the user was able recover by a suitably rephrasing the utterance from the one where they simply gave up. For example, if the system failed to translate "How long has this cough been troubling you?", but correctly handled the simpler formulation "How long have you had a cough?", we would give this a small positive score, rather than a simple zero.

Summarising, we propose to classify translations into the following seven categories:

1. Perfect translation, useful clinical consequences.
2. Perfect translation, no useful clinical consequences.
3. Imperfect translation, but not dangerous in terms of clinical consequences.
4. Imperfect translation, potentially dangerous.
5. Nonsense.
6. No translation produced, but later rephrased in a way the system handled adequately.
7. No translation produced, but not rephrased in a way the system handled adequately.

3.2 Assigning utterances to categories

At the moment, medical professionals will only accept the validity of category assignments made by trained physicians. In the worst case, it is clearly true that a layman, even one who has received some training, will not be able to determine whether or not a mistranslation has clinical significance.

Physician time is, however, a scarce and valuable resource, and, as usual, typical case and worst case may be very different. Particularly for routine testing during system development, it is clearly not possible to rely on expert physician assessments. We consequently suggest a compromise strategy. We will first carry out an evaluation using medical experts, in order to establish a gold standard. We will then repeat this evaluation using non-experts, and determine how large the differential is in practice.

We initially intend to experiment with two different groups of non-experts. At Geneva University, we will use students from the School of Translation. These students will be selected for competence in English and Spanish, and will receive a few hours of training on determination of clinical significance in translation, using guidelines developed in collaboration with Glenn Flores and his colleagues at the UT Southwestern Medical Center, Texas. Given that the corpus material is simple and sterotypical, we think that this approach should yield a useful approximation to expert judgements.

Although translation students are far cheaper than doctors, they are still quite expensive, and evaluation turn-around will be slow. For these reasons, we also propose to investigate the idea of performing evaluations using Amazon's Mechanical Turk[4]. This will be done by Dolores Labs, a new startup specialising in Turk-based crowdsourcing.

3.3 Scores for categories

We have not yet agreed on exact scores for the different categories, and this is something that is

[4] `http://www.mturk.com/mturk/welcome`

probably best decided after mutual discussion at the workshop. Some basic principles will be evident from the preceding discussion. The scale will be normalised so that failure to produce a translation is counted as zero; potentially dangerous mistranslations will be associated with a negative score large in comparison to the positive score for a useful correct translation. Inability to communicate can certainly be dangerous (this is the point of having a translation system in the first place), but mistakenly believing that one has communicated is usually much worse. As Mark Twain put it: "It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so".

3.4 Discarding uncertain responses

Given that both speech recognition and machine translation are uncertain technologies, a high penalty for mistranslations means that systems which attempt to translate everything may easily end up with an average negative score - in other words, they would score worse than a system which did nothing! For the shared task to be interesting, we must address this problem, and in the doctor to patient direction there is a natural way to do so. Since the doctor can reasonably be assumed to be a trained professional who has had time to learn to operate the system, we can say that he has the option of aborting any translation where the machine does not appear to have understood correctly.

We thus relativise the task with respect to a "filter": for each utterance, we produce both a translation in the target language, and a "reference translation" in the source language, which in some way gives information about what the machine has understood. The simplest way to produce this "reference translation" is to show the words produced by speech recognition. When scoring, we evaluate both translations, and ignore all examples where the reference translation is evaluated as incorrect. To go back to the "penicillin" example, suppose that Spanish source-language speech recognition has incorrectly recognised *La penicilina me da alergias* as *La penicilina* no *me da alergias*. Even if this produces the seriously incorrect translation "I am not allergic to penicillin", we can score it as a zero rather than a negative, on the grounds that the speech recognition result already shows the Spanish-speaking doctor that something has gone wrong before any translation has happened. The reference translation may also be produced in a more elaborate way; a common approach is to translate back from the target language result into the source language.

Although the "filtered" version of the medical speech translation task makes good sense in the doctor to patient direction, it is less clear how meaningful it is in the patient to doctor direction. Most patients will not have used the system before, and may be distressed or in pain. It is consequently less reasonable to expect them to be able to pay attention to the reference translation when using the system.

4 Summary and conclusions

The preceding notes are intended to form a framework which will serve as a basis for discussion at the workshop. As already indicated, the key challenge here is to arrive at metrics which are acceptable to both the medical and the speech and language community. This will certainly require more negotiation. We are however encouraged by the fact that the proposal, as presented here, has been developed jointly by representatives of both communities, and that we appear to be fairly near agreement. Another important parameter which we have intentionally left blank is the duration of the task; we think it will be more productive to determine this based on the schedules of interested parties.

Realistically, the initial definition of the metric can hardly be more than a rough guess. Experimentation during the course of the shared task will probably show that some adjustment will be desirable, in order to make it conform more closely to the requirements of the medical community. If we do this, we will, in the interests of fairness, score competing systems using all versions of the metric.