TURKOISE: a Mechanical Turk-based Tailor-made Metric for Spoken Language Translation Systems in the Medical Domain

STARLANDER, Marianne

Abstract

In this paper, we will focus on the evaluation of MedSLT, a medium-vocabulary hybrid 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). How can the developers be sure of delivering good translation quality to their users, in a domain where reliability is of the highest importance? With MedSLT sentences are usually translated freely and, as a consequence of spoken input, they are often short. These characteristics entail low BLEU scores (Starlander and Estrella, 2011) as well as poor correlation when using human judgments. In the present paper we will describe the path that led us to using Amazon Mechanical Turk (AMT) as an alternative to more classical automatic or human evaluation, and introduce task-specific human metric, TURKOISE, designed to be used by unskilled AMT evaluators while guaranteeing reasonable level of coherence between the evaluators.

Reference

TURKOISE: a Mechanical Turk-based Tailor-made Metric for Spoken Language Translation Systems in the Medical Domain

Marianne Starlander

University of Geneva, FTI-TIM, 40 Bd du Pont d’Arve, CH-1211 Genève 4
E-mail: Marianne.Starlander@unige.ch

Abstract

In this paper, we will focus on the evaluation of MedSLT, a medium-vocabulary hybrid 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). How can the developers be sure of delivering good translation quality to their users, in a domain where reliability is of the highest importance?

With MedSLT sentences are usually translated freely and, as a consequence of spoken input, they are often short. These characteristics entail low BLEU scores (Starlander and Estrella, 2011) as well as poor correlation when using human judgments.

In the present paper we will describe the path that led us to using Amazon Mechanical Turk (AMT) as an alternative to more classical automatic or human evaluation, and introduce task-specific human metric, TURKOISE, designed to be used by unskilled AMT evaluators while guaranteeing reasonable level of coherence between the evaluators.

Keywords: spoken language translation, evaluation, crowdsourcing

1. Introduction

Speech recognition and machine translation are now widely available on laptops and mobile devices: typical examples are the speech-enabled Google Translate mobile application and Jibbigo (recently acquired by Facebook). A more specialized system is the US military application, now also running on smart-phones (Weiss et al, 2011). These technologies can be of use in many situations, especially when fast, low-cost translations are required. In this paper, we will investigate the application of the above technologies for medical communication purposes. We will focus on the evaluation of MedSLT, a medium-vocabulary hybrid 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 central question faced by the developers is to deliver good translation quality to their users, in a domain where reliability is of the highest importance. Here, we will consider the question of determining if the translations provided by the system are suitable for the task: enabling communication between the physician and his patient without generating ambiguity or errors that could potentially endanger the patient.

In previous research we have evaluated the usability, translation quality and recognition quality of MedSLT (Starlander and Estrella, 2009). This research has revealed discrepancies between usability measures, automatic measures and human evaluation of translation and recognition quality. This made us further investigate the question of translation quality in (Starlander and Estrella, 2011) in order to develop an evaluation method equally suitable both for rule-based spoken language translation (SLT) systems such as MedSLT, and for SMT based systems. One of the findings of the study was that a classic metric like BLEU is not well-suited for the evaluation of MedSLT output, due to the architecture of the system. The problem is that MedSLT was designed with a strong focus on reliability in correct transmission of the message. One of the characteristics of MedSLT is its rule-based architecture, which uses an interlingua approach to produce highly reliable output. This approach discards surface form to keep only the meaning of the sentences. Consequently, sentences are translated freely; for example, “Do you have a sore throat?” is translated as “Le duele la garganta” (closer to “Does your throat hurt?”) instead of the more literal “Tiene dolor de garganta”. Due to these characteristics, and also to the fact that our sentences are short (10% of the corpus consists of sentences counting less than 4 words), the BLEU scores obtained in (Starlander and Estrella, 2011) were low, and did not correlate well with human judgment. This concurs with the common opinion in the MT literature (Callison-Burch et al, 2006, Popescu-Belis et al, 2004) that automatic metrics like BLEU are often not well suited for rule-based machine translation (RBMT) systems, given that they tend to reward translations that are literal and close to a given reference. The results of our experiments are presented in the following section. In section 4, we describe the results obtained when using AMT. Indeed, despite the increasing use of AMT – in a great variety of fields –, the question of the reliability of this type of evaluation compared with a small amount of expert evaluators still remains open, especially with the kind of complex task (requiring bilingual evaluators) we are submitting. On the other hand, AMT workers might arguably be more appropriate judges of translation quality, as they are closer to the real users of the system (on patient side) and are less likely to focus on the linguistic
form than translators.

Our study focuses on inter-rater agreement, comparing this statistic for our small in-house group of translator-evaluators against a wider group of AMT workers. We also quantify the effort involved in running the AMT evaluation in order to compare the resources needed. Developers and researchers tend to minimize the effort related to the creation of the reference translations needed in order to use BLEU or other reference-based metrics. If AMT workers are found to be reliable, we argue that this type of evaluation is at least as cost- and time-effective as classical automatic metrics, while also providing the advantage of reflecting the end-user’s quality level request.

2. Experiment

This experiment retraces the road from the classic human metrics, to a tailor made human metric and finishes by testing our metric using non-expert evaluators in order to compare our proposed 4-point scale with the traditional 5-point scale fluency/adegacy implemented both by a small group of selected in-house-evaluators and by AMT workers.

We had envisaged experimenting with quality estimation (QE) methods such as described in Specia et al (2010) but our corpus being small, we decided to carry out a first trial with AMT before utterly pursuing in the QE direction.

2.1 Data

Our data set is composed of utterances collected during a test-phase in 2008, where we simulated medical diagnosis dialogues with English-speaking physicians and Spanish-speaking standardised patients at the Dallas Children’s Hospital using our RBMT system MedSLT. The total corpus is made of approximately 1200 utterances, but the test corpus for the actual study consists of an excerpt of approximately 220 English to Spanish translations. We removed sentences that were too short to be of real interest for the task, hence ruling out all the one-word utterances (mainly yes, no, un poco, mucho...) and we only kept the types (removing all multiple occurrence of a sentence and translation).

2.2 Evaluation tasks

The current experiment was divided into the following three tasks:

Monolingual evaluation task:
- fluency of translation results from English to Spanish (Flawless, Good, Non-native, Disfluent, Incomprehensible Spanish)

Bilingual evaluation tasks:
- adequacy of English-Spanish translations using a 5-point scale (All information is present, Almost all information is present, A lot of information is present, Only a little information is present, No information is present)

- rating of English-Spanish translations using our custom 4-point scale (Starlander 2009) as described in section 4.

For these evaluation tasks we used the classical fluency and adequacy human metrics, as well as our own task-centred human metric for the medical domain (Starlander, 2009). This metric relies on human judgements and uses a 4-point scale:

CCOR (4): The translation is completely correct. All the meaning from the source is present in the target sentence.

MEAN (3): The translation is not completely correct. The meaning is slightly different but it represents no danger of miscommunication between doctor and patient.

NONS (2): This translation doesn't make any sense, it is gibberish. This translation is not correct in the target language.

DANG (1): This translation is incorrect and the meaning in the target and source are very different. It is a false sense, dangerous for communication between doctor and patient.

The two categories NONS and DANG might seem similar, but the un-negligible difference between these two categories is that a sentence from the category NONS is much easier identified by the end users (being gibberish or at least incorrect target language); while as a sentence from the category DANG would be a perfectly well built sentence but where the original meaning has been altered in such a way that the translation could represent a danger for the patient using the SLT system. In the table below we provide some evaluation examples using the above-defined scale.

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are you having a fever?</td>
<td>¿El dolor está aliviado cuando tiene fiebre?</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Did you see a doctor this week?</td>
<td>¿Has consultado un médico esta semana?</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Do you have a headache?</td>
<td>¿Tiene tos ayer?</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Examples of the 4-point scale application.

Our tailor-made 4-point scale for SLT in safety-critical domain had so far been used only by very limited numbers of evaluators.

In section 3, we will describe the results obtained by both evaluation groups: our group of in-house translators and the AMT workers. We will analyse the resulting AMT evaluations under the following aspects. First, we aim to determine which of the different scales leads to the best inter-annotator agreement. Then, we will observe how evaluations by AMT workers relate to evaluations by in-house translators for this domain, and finally in section
3.4 we will study how the Kappa evolves according to the number of total evaluation assignments.

Finally, we hope to provide a better insight on how to use AMT on this type of evaluation task. We will conclude on the potential of TURKIOUSE as an alternative to automatic metric or classical human judgement limited to a few expert evaluators.

2.3 Participants

Classically, when launching a human evaluation, the problem is to find suitable evaluators. In our particular context, at the Faculty of Translation and Interpretation of the University of Geneva, it is not a too difficult task to find freshly graduated translators, final year students or fellow translators willing to participate for free at evaluation tasks like ours. This however is only true in our particular context, in “real life” finding enough evaluators can soon turn out to be a very time and money consuming task if not a total nightmare. We have always been able to recruit, indeed, relatively small groups of translators and non-translators, but answering fast (most of the time within 24h).

To investigate another evaluation approach not specific to our context, we have chosen to submit the same tasks to workers recruited on Amazon Mechanical Turk (AMT). The idea behind extending this human evaluation to AMT workers is of course to offer human evaluation but in a faster and cheaper way (Callison-Burch et al., 2009) as has been done at the IWSLT campaigns since 2010 (Paul et al, 2010 and Federico et al; 2012), but also to extend the already wide usage of AMT to an even larger variety of tasks. AMT workers have been involved in tasks ranging from labelling to transcription and have recently moved to spoken dialogue systems evaluation (Jurcicek et al, 2011) with success. Another incentive to experiment with AMT is our interest in comparing the results obtained with evaluators of different backgrounds. Indeed, we have observed that translators tend to focus on the form rather than on the meaning of the provided sentences, which is not the most relevant aspect in our context. Arguably, the AMT are a closer population to the real end-users of such systems as MedSLT. As a consequence of the wider usage of AMT workers for an increasing amount of tasks and resource creation, criticism has risen concerning these practices (Fort et al, 2011). The question of low payment and discrepancies in the quality of the tasks fulfilled are surely topics that deserve discussion. On the first topic, I would like to point out that most of our in-house evaluations have been done in a benevolent manner, in the spirit of helping out fellow translators or researchers. In future work, once a minimum of reliability from the AMT workers could be established for the proposed task in this paper, we would certainly like to investigate more about the ethical and sociological impact of AMT on the research. But at present, we will present the characteristics of the participants to our current experiment.

2.3.1 In-house translators

For the fluency and adequacy evaluation task, we recruited five in-house trained English to Spanish translators. They completed the evaluation in Microsoft Excel spreadsheets. They first graded the fluency by reading only the Spanish target and second, they rated the adequacy of the translations using the classical 5-point scale.

The evaluation using the above described 4-point scale (CCOR, MEAN, NONS, DANG) had been done in previous research. In the first stage, we asked the in-house interpreters of the Dallas Children’s Hospital that participated to the data collection in 2008 to evaluate by email a series of sentences (190 per evaluator). This could somehow be compared to an AMT setting, since we did not know these translators, and also because they did not evaluate all the sentences but only a subset. We had divided the data in such a way that in the end we should have obtained five human evaluations for each sentence. At the same time, we asked these participants to provide a reference translation on a subset of the evaluated translations. We succeeded in gathering three reference translations for each sentence but managing the translators for the evaluation and the production of reference translation was a very time consuming task.

In a second stage, we asked known in-house translators and finally we asked three non-translators to evaluate the entire set of sentences (222 sentences). The time cost for the evaluation in the second phase is comparable to the AMT response time: ranging from 1 hour to 24 hours, according to the respective work-load of our evaluators. Collecting reference translations in order to calculate BLEU and other classical automatic metrics was a much more time consuming task that is however difficult to evaluate.

2.3.2 Crowdsource evaluation

On AMT, translations to evaluate are presented to AMT workers grouped in human intelligence tasks HITs (HITs) of 20 sentences each. These HITs are set up by combining our data with html templates adapted to the different evaluation scales. AMT workers are paid for the evaluation task. We allowed a maximum of 10 assignments per HIT, meaning that 10 different workers would evaluate each sentence. We proceeded in two steps: in the first phase of the experiment we restricted the access to 5 assignments, i.e. a total of five evaluations. The idea is to observe the effect on inter-rater agreement when multiplying the amount of evaluators substantially. We therefore opened the task in a second phase to yet another five assignments in order to observe the behaviour of inter-rater agreement.

In the first phase we studied the reliability of the AMT workers. As we are mainly interested in inter-annotator agreement for the different scales, we have to be particularly careful on how to design the AMT experiment in order to enable detection and exclusion of suspicious responses. Several methods already exist, such
as filtering out extremely short task duration (Kittur, 2008), but this alone is not sufficient in our case. Our main difficulty is making sure that the workers who accept our HITs do in fact have the language skills required, namely good knowledge of both languages involved. While workers are generally careful not to work on HITs that they cannot complete satisfactorily, as this might lead to rejection of their work, we cannot exclude that some might try despite insufficient language skills.

In order to recruit AMT workers that seem to present the required skills (fluency in English and Spanish), we first posted a selection HIT of 20 gold-standard sentences. This gold standard set was composed of sentences where all five of our in-house evaluators had reached unanimous judgement. The AMT workers obtaining a minimum of 75% of agreement on the selection HIT were attributed a qualification, and given access to the three real tasks divided into a total of 29 HITS. Very rapidly (within 24 hours) we had assigned the qualification to 20 “suitable” workers, and within three days, our five assignments for all the proposed hits where completed. As mentioned above, we had limited the number of assignments to five for the first stage of the experiment with AMT in order to first verify the feasibility of the task, since it involved English-Spanish bilingual AMT workers. We were surprised to find sufficient English-Spanish participants on the AMT to reach our assignment of five within such a limited time. We were able to give the aforementioned qualification to a total of 11 workers, within 24 hours. However, on the second phase, only 3 out of 13 invited participants participated within an acceptable responding time ranging (within a week). We could have opened the task to more AMT workers, but this meant reopening a qualification round, which we decided not to do and provide the results for a maximum of eight assignments instead.

For the second phase of the experiment, we first identified which qualified AMT workers had not yet achieved the task in the first stage, and quite simply, by changing the level of qualification, we were able to invite them to participate to the new stage of the task by opened 5 assignments for each task (fluency, adequacy and TURKoise). The aim of this new phase was to study the impact of multiplying the number of assignments, through doubling the number of evaluations for each translation, and finally trying to identify what the critical threshold would be: when does the inter-rater agreement reach a peak or degrade. We intended to observe how the inter-rater agreement evolves when multiplying the number of evaluations. The idea behind this second experiment was to find out the impact of having more evaluators working on a task. We hypothesized that if we obtain the same type of inter-rater agreement (i.e. evaluation quality) with 3 assignments, 5 assignments and 8 assignments, it would mean that a small crowd is sufficient to achieve our goal of proposing a reliable Turk-based evaluation.

The next section describes the results obtained for the three tasks and two groups of evaluators. In the section Kappa Evolution we will present the results obtained in the second phase of the experiment.

3. Results

Handling human metrics always implies checking their coherence and inter-rater agreement. We start with the human metrics: fluency and adequacy, using the classic 5-point-scale, then we briefly present the results obtained with our tailor-made human metric. In the first column, we present the results achieved by our in-house evaluators and in the second column contains the results for the AMT workers. We calculated the percentage of unanimous and majority agreement, but we also present Fleiss’ Kappa. In our previous study (Starlander & Estrella 2011) we discussed the difficulty of interpretation of Kappa, and decided to follow (Hamon et al, 2009)’s example and calculate the percentage of total agreement between judges, that is the number of times all judges agreed on a category. We apply this on all three evaluation task (fluency, adequacy and TURKoise).

3.1 Fluency

In table below, the difficulty and subjectivity of the fluency evaluation tasks appears clearly. The percentage of the majority agreement only amounts to 84% for the in-house evaluators and 79% for the AMT workers. For this task, the Fleiss Kappa is slightly better for the in-house translators.

<table>
<thead>
<tr>
<th>Fluency</th>
<th>In-house</th>
<th>AMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>unanimous</td>
<td>21%</td>
<td>22%</td>
</tr>
<tr>
<td>4 agree</td>
<td>32%</td>
<td>29%</td>
</tr>
<tr>
<td>3 agree</td>
<td>31%</td>
<td>28%</td>
</tr>
<tr>
<td>majority</td>
<td>84%</td>
<td>79%</td>
</tr>
<tr>
<td>Fleiss Kappa</td>
<td>0.174</td>
<td>0.164</td>
</tr>
</tbody>
</table>

Table 2: Fluency

As expected, the best agreement is obtained for the highest categories, when the sentences are fluent (obtaining 5 out of 5). This observation can be done for both categories of participants. In this table we present the result for five assignments. As we will further comment in section 3.4, the fluency task clearly comes out as the mostly subjective one, were the variety of answers is rich and the agreement poor. Using AMT for this type of task could be an advantage as the crowd would compensate for the agreement.

3.2 Adequacy

Regarding the adequacy of the translations obtained by MedSLT, a divergence between the percentage of majority agreement and the Fleiss Kappa appears. For the first, the in-house evaluators reach 93% compared to 86% for the AMT workers, while as for the latter, the AMT
receives a clearly higher Fleiss Kappa (0.236) than the in-house translators (0.121).

<table>
<thead>
<tr>
<th>Adequacy</th>
<th>In-house</th>
<th>AMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>unanimous</td>
<td>26%</td>
<td>36%</td>
</tr>
<tr>
<td>4 agree</td>
<td>40%</td>
<td>27%</td>
</tr>
<tr>
<td>3 agree</td>
<td>27%</td>
<td>24%</td>
</tr>
<tr>
<td>majority</td>
<td>93%</td>
<td>86%</td>
</tr>
<tr>
<td>Fleiss Kappa</td>
<td>0.121</td>
<td>0.236</td>
</tr>
</tbody>
</table>

Table 3: Adequacy

This result is quite encouraging, and tends to indicate that the path of using AMT for evaluation of SLT output could be followed with success. We are keen on comparing this relatively good result with those coming from the tailor-made 4-point scale metric designed for AMT: TURKoise.

3.3 TURKoise

As expected from the results obtained for the adequacy evaluation, the results for TURKoise by the AMT are absolutely comparable to the results obtained by our in-house evaluators. Again, the Fleiss Kappa for the AMT workers is around 0.232, which, although this figure is in the lower range of the interpretation grid for Kappa (Landis & Koch, 1977), the trend is out-classing the equivalent in-house evaluation.

<table>
<thead>
<tr>
<th>TURKoise</th>
<th>In-house</th>
<th>AMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>unanimous</td>
<td>15%</td>
<td>32%</td>
</tr>
<tr>
<td>4 agree</td>
<td>35%</td>
<td>26%</td>
</tr>
<tr>
<td>3 agree</td>
<td>42%</td>
<td>37%</td>
</tr>
<tr>
<td>majority</td>
<td>92%</td>
<td>95%</td>
</tr>
<tr>
<td>Fleiss Kappa</td>
<td>0.199</td>
<td>0.232</td>
</tr>
</tbody>
</table>

Table 4: TURKoise

In most cases, the graders disagree on only one point of the scale, but agreeing on the general quality of the rated sentence.

The agreement is far higher for the CCOR category than for the NON or DANG categories, this could be observed in general for all evaluations (fluency, adequacy and TURKoise).

Following these observations, we would like to make some experiments by collapsing similar categories: regrouping on one hand CCOR and MEAN and on the other hand NONS and DANG together, as it is well known that the more point on a scale the less the agreement. In previous research we had observed much higher Kappa on binary evaluation task (ranking task). It would be interesting to test this with the AMT in further research.

In a preliminary study about the participating population, our evaluators recruited so far were always translators and we decided to add non-translators to our evaluators’ population, in order to verify the impression we had about our evaluators being particularly severe. As you can see from the table below, it is almost impossible to reach full agreement of 6 evaluators, it only happens in 18.5% of the tested sentences.

<table>
<thead>
<tr>
<th>In-house evaluators</th>
<th>% of full agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>All 6 evaluators</td>
<td>18.5%</td>
</tr>
<tr>
<td>Translators (3)</td>
<td>33.8%</td>
</tr>
<tr>
<td>Non-translators (3)</td>
<td>41.9%</td>
</tr>
</tbody>
</table>

Table 6: Inter-rater agreement in % of full agreement

However, in the table above, we still observe low inter-annotator agreement when using our 4-point scale, but a more “positive attitude” from the non-translators and a higher inter-annotator agreement, which suggests that translators might not be the best judges as they find it difficult not to focus on the form. Of course, this type of categorisation for our evaluators is impossible to apply to crowdsourcing, since we don’t know the profiles of our AMT workers. The only information we gathered is that they are fluent enough to achieve our task. But this is also where the idea of using crowdsourcing occurred to us, since the amount of evaluators would have two effects: smoothing the inter-rater differences and provide a significant amount of bilingual evaluators without being translators.

3.4 Kappa evolution

When using crowdsourcing, the temptation of multiplying the number of evaluators to compensate for the inherent incoherence of human evaluation is great, but the question we wanted to investigate is: Does an increasing number of evaluators really make a difference? The fact that AMT workers are quite easily available and cost-effective does not automatically mean that “more is better”.

In the figure below we present the results obtained in our second AMT phase, when we added three supplementary evaluations for each sentence. We compare Fleiss’ Kappa
obtained for respectively three, five and eight assignments compared to the Kappa achieved by our 5-inhouse evaluators, who each evaluated the entire corpus (222 sentences).

<table>
<thead>
<tr>
<th>Number of eval.</th>
<th>Fluency</th>
<th>Adequacy</th>
<th>TURKoise</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-times AMT</td>
<td>-0.052</td>
<td>0.135</td>
<td>0.181</td>
</tr>
<tr>
<td>5-times AMT</td>
<td>0.164</td>
<td>0.236</td>
<td>0.232</td>
</tr>
<tr>
<td>8-times AMT</td>
<td>0.134</td>
<td>0.226</td>
<td>0.227</td>
</tr>
<tr>
<td>5-inhouse eval.</td>
<td>0.174</td>
<td>0.121</td>
<td>0.199</td>
</tr>
</tbody>
</table>

Table 7: Kappa evolution according to the number of evaluations

It is interesting to observe that the task obtaining the worse Kappa is fluency. It appears that the more "subjective" the task the more evaluations are needed to compensate and reach a slightly better Kappa. Generally, it appears that the AMT does well, and achieves comparable Kappa levels with the human in-house evaluators. Clearly, for the fluency task, three assignments are not enough, since the achieved Kappa doesn’t show any agreement and even show a slight disagreement between the evaluators (Landis & Koch, 1977). Asking AMT workers to achieve a total of five evaluations in a crowd effort seems to be the best alternative regarding the above table.

Another interesting result from this experiment is the fact that TURKoise achieves a higher Kappa with five assignments than with eight, and confirms that the Kappa is higher than for the in-house evaluators as soon as more than three assignments are accomplished by AMT workers.

4. Conclusion and further work

These preliminary results using AMT workers are encouraging. Overall, it seems that crowdsourcing can be considered to be an effective method to replace or enhance small group-evaluations. The results obtained by both groups are quite comparable on the level of inter-rater agreement. The results obtained tend to show that our selected AMT workers are reliable and surely are cost- and time- effective. Our experiment could also determine which of the three evaluation scales lead to the inter-rater agreement. Our test shows that there is no need to multiply the number of assignments in order to gather more than 5 evaluations for each sentence. In our experiment we needed a total of 23 AMT workers to achieve this goal. The total cost can thus even more be restricted than for our experiment, since five assignments gave better results than eight. Hence, we could identify the “ideal” size of the effectively used crowd to be 5 assignments ventilated on as many AMT workers as necessary, knowing that the total cost for phase one was of 55$, we can also conclude that this method is cost-effective (compared to 50$ for non-benevolent each evaluator): expanding the number of assignments while still remaining competitive in terms of cost and quality. Our general conclusion is thus that there is a clear potential of using TURKoise (i.e. with AMT workers) as an alternative to classic fluency and adequacy human evaluation but also to classic reference-based automatic evaluation.

Then, as mentioned in the introduction, a supplementary aim of our study is to provide the research community with an evaluation method that would not be biased in favour of SMT or RBMT, but that would equally suit to compare spoken language translation output being produced by SMT, RBMT or hybrid systems. We would therefore further investigate the suitability of TURKoise using it this time on both MedSLT output and on the translation provided by an SMT system (probably Google translate) to build on the results from (Starlander & Estrella, 2011). For this purpose we will thus also add a human binary ranking task of MedSLT translations vs. translations obtained with Google Translate, and maybe also experiment with AMT workers using a simplified 2-point scale resulting from the above mentioned scale (Meaning, Danger).

In the present paper we focused on a reduced version of our Corpus made of English to Spanish translations, being originally questions asked by the physicians and translated into Spanish by MedSLT to enable the patients to interact with the doctor. The whole corpus So far we have not yet tackled the trickier data of Spanish to English patient responses. These utterances are often even shorter and rely on ellipsis resolution (Bouillon et al., 2005), which represent extra difficulties to be handled by TURKoise. Hence, the next step would be to extend the present study to the remaining real data collection.

Further work, once these experiments achieved, would be to exploit the collected ratings through our “controlled” crowdsourcing to implement quality estimation specifically chosen to suite SLT systems in a safety critical domain, with the “handicap” of having only a very small amount of specific data to add to the general models used (Specia et al. 2010). A viable perspective could be to enhance the corpus through targeted web-search, as there is a massive amount of medical diagnosis type data available out on the net.

5. Acknowledgments

We would like to acknowledge our benevolent in-house evaluators for participating freely to our experiment. Special thanks to my colleague Johanna Gerlach that shared her precious experience and knowledge of the AMT platform.

6. References


Google translate site, http://translate.google.com/

Jibbigo information site:


