

Looking for the best Evaluation Method for Interlingua-based Spoken Language Translation in the medical Domain

Marianne Starlander and Paula Estrella

University of Geneva, ETI-TIM-ISSCO, 40 Bd du Pont d'Arve, 1211 Genève 4

`Marianne.Starlander@unige.ch`

University of Córdoba, FaMAF, Haya de la Torre s/n, Ciudad Universitaria, 5000 Córdoba

`pestrella@famaf.unc.edu.ar`

Abstract. This paper focuses on the quality of rule-based machine translations collected using our open-source limited-domain medical spoken language translator (SLT) tested at the Dallas Children's hospital. Our aim is to find the best suited metrics for our Interlingua rule based machine translation (RBMT) system. We applied both human metrics and a set of well known automatic metrics (BLEU, WER and TER) to a corpus of translations produced by our system during a controlled experiment. We also compared the scores obtained for both type of evaluation with those obtained on translations produced by the well known statistical machine translation (SMT) system GoogleTranslate¹ in order to have a point of comparison. Our aim is to find the best suited metric for our type of Interlingua RBMT SLT system.

Keywords. Key words: Machine translation evaluation, Spoken language translation, Automatic metrics

1 Introduction

MedSLT is a medium-vocabulary open-source speech translation system intended to support medical diagnosis dialogues between a physician and a patient who do not share a common language [1]. The translation module is rule-based in order to provide a more predictable translation, prioritizing precision over recall given the safety-critical nature of the task. This implies that we prefer to produce no translation at all instead of bad translations that would account for recall, but that

¹ <http://translate.google.com/>

could entail communication errors between the physician and the patient, potentially leading to diagnosis errors.

More specifically, the translation is interlingua-based making the system multilingual, translating a dozen of language combinations to/from {ENG, FRE, JPN, SPA, ARA, CAT} [2]. This Interlingua architecture helps us add new languages more easily but at the cost of reaching a common representation for several languages by focusing on the meaning of the sentences to produce the corresponding common semantic representation. Therefore, the resulting translations are less literal than those produced by other models such as example-based or statistical models. This approach avoids problems of divergences and discrepancies that would inevitably arise between the large varieties of language families handled.

In order to assess our system, we have carried out a set of evaluations using different methodologies in the quest for the most appropriate one for our system [3-4]. We have applied a set of state-of-the-art metrics, including human-based and automatic ones. However, as it has often been mentioned in the MT literature [5], automatic metrics based on computing the similarity of an output against one or more references (like BLEU, WER, and most of the commonly used metrics) seem to be less suited for rule-based machine translation (RBMT) systems, given that they tend to reward translations that are more literal and close to a given reference. Thus, our Interlingua architecture seems at first sight incompatible with this type of automatic metrics for the evaluation of its translation quality. In this paper we want to further study the suitability of these automatic metrics compared to tailor-made human metrics and we will therefore apply a set of metrics to a corpus gathered during an experiment, where we tested the medical spoken translation system in a controlled environment, very close to that of real use. The results of running a series of tests on the RBMT data are then compared to those obtained using statistical translations produced by GoogleTranslate (GT)[6]. We chose to use GT as a baseline, to have a point of comparison with a statistical machine translation (SMT) system, but we are aware that the comparison is slightly unfair since GT has not been particularly trained for this task. However, many bilingual resources exist on the web in the domain of medical diagnosis, so this choice is less unfair than if we had chosen a SMT system trained, for example, on the Europarl corpus [7]. Some tests have been conducted using automatically generated data to build a SMT system equivalent to MedSLT for English-French and English-Japanese, but not for English-Spanish since the resulting translation does not outperform the RBMT [8].

The rest of the paper is organised as follows: In section 2 we give more background about MedSLT explaining in more details why our system is Interlingua and RBMT based, and what this choice implies on the resulting translations. In section 3, 4 and 5 we describe the experiment conducted using different ways of evaluating MT. In section 6 we study the correlation between our human metrics and the chosen automatic metrics. Finally in section 7, we conclude on whether these metrics can be useful for evaluating RBMT in a spoken language translation context.

2 System description

In this study we are using the bidirectional English-Spanish version of MedSLT that was used during tests conducted at the Dallas Children's Hospital in 2008 [3]. This system enables an English speaking physician to communicate with his Spanish speaking patient during a medical examination. Both speech recognition and translation are rule-based. The speech recognition (SR) component uses the Nuance 8.5 platform [9], equipped with grammar-based language models. The experiments carried out in [8] have shown that for a safety critical task such as MedSLT, statistical SR does not give better results than the RBSR, although it adds robustness to a hybrid system.

The workflow for the bidirectional version of the system is as follows [10]: the physician presses the SR button and speaks into the microphone; he can then check the back-translation of his utterance. If the physician accepts the produced string, it gets synthesised to the patient. The patient clicks on the SR button to speak a direct answer to the question; alternatively, to produce an answer he can make use of the help window that displays a set of potential answers that are covered by the system. The patient checks the produced back-translation and launches the synthesis if the sentence is correct.

In either case, the back-translation is the result of the entire processing of the input by the system. This means that at run-time, the recogniser produces a source language semantic representation which is first translated by one set of rules into an Interlingua form. Then a second set of rules simultaneously translates the representation back into: the source language (that is, a back-translation, so that the user can check if the system has correctly understood and translated the spoken sentence) and into a target language representation. Then, a target-language Regulus grammar compiled into generation form turns this representation into one or more possible surface strings.

An Interlingua-based architecture has been chosen to avoid having to multiply the number of Interlingua to target language translation rules. Instead it keeps a unique translation for all utterances that have received the same Interlingua representation [11], which is almost flat, as you can see for the following example sentence in Figure 1.

Source: Do you have a sore throat?
Interlingua: [[body_part,throat], [prep,in_loc], [pronoun,you],
[state,have_symptom], [symptom,pain], [tense,present], [utterance_type,ynq],
[voice,active]]
Backtranslation : Do you experience a pain in the throat?
Target : ¿ Le duele la garganta ?

Fig. 1. Example of Interlingua representation

The resulting representation is an unordered list of semantic elements. The attributes are derived from the canonical English form, removing most of the grammatical information. The advantage of such a representation is its simplicity. It enables us to easily write translation rules that are expressive enough to convey the

nuances in the concepts used in specific domains. We always try to keep the most idiomatic translation. For example, we chose to translate the source sentence “Do you have a sore throat” by “Le duele la garganta” (closer to “Does your throat hurt”) instead of the more literal “Tiene un dolor de garganta.” As a consequence, the translations produced are clearly freer and more coherent than translation produced by direct linguistic MT or statistical translation as you can see in Table 1 below.

The use of this Interlingua avoids surface divergences in order to keep only the meaning of the sentences. As a consequence certain losses in style sometimes occur but most of the time the lost information is not important for the purpose of SLT in the medical domain.

Source sentences	Target sentences by GoogleTranslate
Did the doctor do a strep test?	Dijo el médico haga una prueba para estreptococo
Did the doctor run a strep test?	Dijo el médico realizar una prueba de estreptococos
Did they do a strep test?	Hicieron una prueba de estreptococos

Table 1. Source sentences collected during the tests all producing the same translation “le han realizado la prueba rápida por estreptococo?” with our system.

As mentioned before, the main goal is to achieve coherence and reliability so that a physician can communicate efficiently and without danger with his patient, which explains why the output of the system is often more idiomatic and thus freely translated.

3 Evaluation of RBMT vs. SMT output

Given the non-literal nature of our translations, we believe that the classical automatic metrics that are based on the resemblance of a MT with one or more references would give low scores on our RBMT system and higher scores for the SMT system that produces more literal and similarly long translations as the original. Hence, purely reference oriented metrics as BLEU [12] and WER should prove less suited for our system. In evaluation campaigns such as WMT09 [13], there is no real comparison between RBMT and SMT. This is why we specifically want to compare the results obtained for RBMT and SMT outputs using the same metrics. Our main objective remains to find more appropriate metrics for RBMT and especially Interlingua based RBMT similar to our system.

According to [14], BLEU shows a favorable bias towards SMT, so we would like to verify this claim. In recent studies some new “less literal” metrics have emerged, such as translation edit rate (TER) [15-16] and METEOR [17]. In [16], TER is described as reaching a higher correlation with human judgments because it assigns lower costs to phrasal shifts than BLEU, which implies that it might be better suited for RBMT than the classical n-gram metrics. However, our corpus is quite different

from classical written MT, because our sentences are very short and often syntactically quite remote from a literal translation as mentioned in section 2, we have thus decided to run both types of metrics on our test corpus. As a comparison point, we are going to study the relation between our tailor-made human metrics and more classical human metrics, and will also analyse their correlation with the chosen automatic metrics, namely BLEU, WER and TER.

We will now explain the experimental framework by describing the data collection. Then we will give a detailed description of the human and automatic metrics applied.

4 Data collection

The data we are using in this experiment has been collected during a test-phase in 2008, where our aim was mainly to compare two versions of the system [3]. We had organised a data collection with English speaking physicians and Spanish speaking standardised patients at the Dallas Children's Hospital. The aim of the task was to determine whether the patient suffered from a bacterial infection (strep throat) or not. Eight physicians and 16 patients participated. The patients were acted by native-Spanish in-house interpreters of the Dallas Children's Hospital. We asked the patients to simulate viral sore throat or strep throat symptoms, described in eight different fixed scenarios. None of the participants had used the system before. Our test corpus for this study consists of 222 English to Spanish translated diagnosis questions from our Dallas data collection.

5 Human evaluation

In our research, we wanted to focus on the end usage of the produced translations and get away from linguistic issues. In our particular case, what is most important is that the message comes across and this is why the scale chosen focuses on the meaning, in a specific context of use: communication between a doctor and his patient while asking diagnosis questions. As suggested in [18], we are aiming at a metric directly related to the final use of the produced translation rather than using the classical metrics that are commonly applied to evaluate the degree of adequacy and fluency of a translation.

5.1 Scale description

Our scale is focused on evaluating if the produced translations are useful for our task or if they could be dangerous. Therefore, this evaluation scale tried to leave purely linguistic aspects on the side, that is, instead of judging the syntactic or linguistic aspects of the translations, the evaluator's task consisted on indicating whether the message from a patient was correctly sent to the doctor. For this purpose, the 4-point scale chosen relates the meaning of a sentence to its potential to create

misunderstandings or false communication between a doctor and his patient. The scale is described as follows:

- 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.

In the evaluation form sent to the judges we included the description of the scale and we provided them with the following examples, by way of tutorial on how to proceed with the evaluation:

Source	Target	Score
Do you experience pain?	Le duele ? (<i>Does it hurt</i>)	MEAN
Do you have a headache	Tiene tos ayer (<i>Do you have a cough yesterday</i>)	NONS
Are you having fever?	¿El dolor está aliviado cuando tiene fiebre? (<i>The pain is decreasing when you have fever</i>)	DANG

Table 2. Evaluation examples for annotators

As mentioned before, this scale is clearly focused on meaning and you could thus wonder why some trace of grammar and style remains present in the category CCOR: this is only to reflect the difference between sentences that are clearly correct in all aspect and sentences that are slightly different but have most of the meaning present. One of the typical examples for the MEAN category is the following sentence, where the meaning is similar although the sentences are syntactically distant: “*do you experience pain*” vs. “*does it hurt*” for the Spanish sentence “*le duele*”.

The order can also appear as surprising since a nonsense sentence (NONS) receives a higher score (2) than a DANG sentences (1). This can simply be explained by the fact that in the context of a medical dialog, a nonsense sentence, that clearly appears as such is more easily recognised and rejected than a sentence that “looks” correct but the meaning is in fact totally different (for example: false negative sentence). This kind of sentences could produce serious diagnosis errors. The main aim of this scale is to encourage the evaluators to forget about linguistic differences and focus on the meaning. But as we will see in section 4.2, where we compare the results of the evaluation by translators and non-translators, we noticed that this is clearly difficult for translators as they continue to rate more severely than non-linguists. While conducting previous studies we also noticed the impact of the attitude towards

machine translation and technology in general on the severity of the evaluation [19]. We had at that time already noticed how difficult it is for “classical” translators to take a certain distance with grammar and style issues in order to focus solely on the meaning, compared to the results of non-translators on the same task.

We thus asked two groups to evaluate the output of our system. The group of translators is composed by a sub-set of the Spanish language Interpreters of the Dallas Children’s Hospital who had participated in the data collection and by a number of professional English-Spanish translators. The second group is composed by non-translators, with a pro-technology background, since most of them happen to be Spanish speaking computer scientists. We asked each group to evaluate a set of 222 sentences translated by our system and by GoogleTranslate applying our human metric.

We will first present the results for our human metrics and then we will pass on to the automatic metrics before studying the correlation between them.

5.2 Results

Table 3 below shows that the average for both types of systems is quite close. The scores are only slightly higher for the RBMT system when it is evaluated by non-translators. The difference between non-translators and translators is clearer in favour of the RBMT. But as a whole the difference between the two systems is not significant if we consider only the averages.

	RBMT	SMT
Translators	3.40	3.43
Non-translators	3.62	3.46
All	3.51	3.44

Table 3. Average using our scale (4=highest, 1=lowest).

In order to get a better idea of the actual quality of each system in Table 4 we show the percentage of each category of the scale, in a majority wins perspective rather than by calculating the average score as in Table 3.

Cat	RBMT Trans.	SMT Trans.	RBMT Non-trans	SMT Non-trans
1=DANG	4.5%	3.2%	3.2%	2.7%
2=NONS	2.3%	6.3%	0.0%	2.3%
3=MEAN	25.2%	20.7%	16.7%	15.3%
4=CCOR	68.0%	65.8%	76.1%	70.7%
No Agreement	N/A	4.1%	4.05%	9.0%

Table 4. Translation quality by category, by majority wins

In this table we can see that in fact our RBMT obtains better results, again especially with our group of non-translators since they evaluated 76.1% of sentences produced as totally correct, compared to only 70.7% for the SMT system. When using human metrics the problem of agreement between judges always arises, so we

decided to calculate the inter-rater agreement using the *AgreeStat Excel VBA program* [20].

Kappa estimate	RBMT	SMT
Translators	0.1758	0.3698
Non-translators	0.0973	0.2591

Table 5. Kappa estimate for our 4-point scale

Table 5 shows that our Kappa estimate is particularly low for the RBMT system. This can quite simply be explained by the fact that this scale is more difficult to apply consistently especially without previous training on how to interpret the scale. However, since these Kappa estimates remain quite difficult to interpret, we decided to follow [21]’ and to calculate the percentage of total agreement between judges, that is the number of times all three judges agreed on the choice of our 4-point scale. As you can see in Table 6, the overall percentage of both categories of evaluators is very low especially for our RBMT system. For the RBMT it is interesting to note the difference between the translators and the non-translators; the latter group is more coherent, while we get the reversed trend for the SMT. It is very interesting to see that the non-translators get a much higher agreement for the RBMT than the translators.

	RBMT	SMT
All 6 evaluators	18.5%	27.9%
Translators (3)	33.8%	49.5%
Non-translators (3)	41.9%	46.4%

Table 6. Agreement between evaluators

The question that arises at this point is if our tailor-made metric has removed all fluency and linguistic differences in quality, leaving us with two quite different output sets that get almost equal results. In order to further study this observation we decided to conduct an extra study using a more classical human metric, namely a ranking evaluation.

The second human evaluation task clearly shows that the output by our RBMT is preferred in 61.1% cases to the output produced by the SMT (34.5%). The Kappa estimate for this task is of 0.5564 which is much higher than the results obtained for the 4-point scale displayed in Table 5. The reason for this is probably that the ranking scale is easier to apply and gives less variation possibilities.

We will now explore the possibility of using automatic metrics in order to finally achieve objective MT evaluation suitable for RBMT.

6 Automatic metrics

As mentioned in section 3, we chose to evaluate standard classic automatic metrics such as Word Error Rate (WER) and BLEU [12] compared to newer metrics like the Translation Edit Rate (TER) [15] computes the number of edits needed to change the output so that it semantically corresponds with a correct translation. Although another potentially suitable automatic metric is METEOR [17], we have not run it in this

experiment because we are lacking the Spanish language resources needed by this metric; this is clearly one disadvantage of this metric preventing its wide use in evaluation.

6.1 Resource description

Since the above cited automatic metrics are very dependent on the reference, we have run the tests with three different reference sets for our 222 source sentences: (1) three human translations provided by the interpreters of the Dallas Children’s Hospital themselves and completed by translations produced by professional English-Spanish translators, (2) a set of translation used as corpus reference for our system and (3) a mix of the two first sets of reference translations in order to provide both more literal human translations and translations that we as developer aimed at in our Interlingua perspective. We are well aware that the corpus is quite small but this is due to the cost of creating such a pool of human references.

We will now analyse the results obtained for the automatic metrics before studying their respective correlation to the human metrics described in the previous section.

6.2 Results

As you can see in Table 7, the average obtained for all sentences are quite similar for both types of systems when we use human translations only (columns 4 and 5) and with an equal number of translations from translators and the developer’s corpus (columns 6 and 7). There only appears a clear difference in favour of the RBMT for all metrics if you use as only reference the developer’s corpus (columns 2 and 3). The latter result is coherent with our second human evaluation task.

Metrics	RBMT-ref_dev	SMT -ref_dev	RBMT-ref_trans	SMT-ref_trans	RBMT-ref_all	SMT-ref_all
BLEU	0.84	0.17	0.35	0.33	0.35	0.33
WER	0.12	0.80	0.59	0.67	0.55	0.58
TER	0.10	0.67	0.53	0.65	0.65	0.66

Table 7. Result for the automatic evaluation

The RBMT corpus has served as reference before in a similar evaluation task using BLEU [22], but we think it is fairer to use a mix of the two types of references (columns 6 and 7). These results point out the importance of the choice of references, since they are totally different according to the translation references used. In order to have a better grasp of why the results are so close in columns 4-6, we decided to check the results by applying the metrics at the sentence level.

Source	Target	Bleu4	Hum.	TER	WER	bleu2
Are you coughing?	¿Tiene tos?	0	4.0	0.0	0.0	1
Do you have a cough?	¿Tiene tos?	0	4.0	0.0	0.0	1
Do you have a fever?	¿Tiene fiebre?	0	4.0	60.0	75.0	0

Have you vomited?	¿Ha vomitado?	0	4.0	1	1	1
-------------------	---------------	---	-----	---	---	---

Table 8. Sentences level evaluation sample

This analysis makes immediately clear that the overall scores cannot be used as such, as you can see in the sample provided in Table 8. To illustrate this, Table 8 shows the scores for Bleu (4-grams), human evaluation, TER, WER and Bleu (2-grams) for sentences 19, 48, 50 and 145 of our corpus. As we mentioned in our system description, our sentences are very short and actually 20% of sentences (46/222) are shorter than 4 words for RBMT and 14% for SMT (32/222), which explains how the scores for these sentences using the classic BLEU based on 4-grams does not suit well for our test corpus. For almost 10% of our sentences, we get a score of 0 while our human evaluators rated with a 4.

Although they carry the same content, our translation is often quite distant from the original syntax, as shown in the following example. Our translation for “*Do you have a rash*” is “*Tiene una erupción cutánea*” but all human references contain a more regional variation as “*Tiene sarpullido*” or the more familiar variation “*Tiene un picor*” or “*Tiene urticaria*”. Another example of this kind is our translation for “*What are you allergic to?*” which is “*Qué le da alergias?*”. This solution has been adopted in order to avoid ambiguities in the gender (e.g. *alérgica/alérgico*) that our reference translators did not take into account: “*A qué es alérgico?*”. In those cases, only a semantic metric, rich in synonyms and regionalisms could detect that these sentences are equivalent, even if on the n-gram side there is almost no resemblance. These two observations explain how the BLEU score in Table 8 are artificially drawn down for our system.

In order to find the fairest metric for our task, we calculate the correlation with the human evaluation on a sentence basis and added scores for BLEU2 and BLEU3.

Correlation type	RBMT	SMT
bleu vs H	0.127	0.264
bleu-3 vs H	0.205	0.290
bleu-2 vs H	0.331	0.223
Ter vs H	-0.304	-0.208
wer vs H	-0.487	-0.262

Table 9. Correlation between automatic and human metrics on segment level

Table 9 shows that the highest correlation occurs for BLEU-2 and WER and not for TER compared to our initial hypothesis. It is interesting to note that the correlations for the SMT are much lower than those for RBMT and that BLEU (3-gram and 4-gram) correlates better with humans in the case of the SMT. Finally, these results show that TER is not behaving so differently from the classic n-gram metrics and that a possible set of metrics to apply in future evaluations could be our human metrics plus BLEU-2 and WER.

7 Conclusion

The aim of this paper was to find the best suited metrics to evaluate the output of our RBMT system. One of our findings is that the automatic metrics we used did not show a bias in favour of SMT; in fact, correlations are lower for the SMT and also the automatic scores, depending on the set of references used. However, they did not prove adequate either. The results obtained in section 5 prove that, given the nature of our corpus, we still need to find a better metric. It turns out that in our context, the classical BLEU based on 4-grams is not suited at all and should be replaced by BLEU based on bi-grams. It would have been interesting to apply other metrics, such as METEOR, in order to explore other aspects of our translations but, as mentioned before, we need the necessary resources for the language under evaluation, Spanish in this case. According to the study described in [13], the best correlation with human evaluation is achieved with UPC [23], which is a combination of many metrics commonly used but the interesting idea is that the authors aim at not only assess one facet of MT quality, which is in most cases the lexical resemblance but to try to englobe syntactic and semantic aspects. This is the direction we need to take, because what we aim at is a metric that assesses the quality of MT through the semantic equivalence to the reference translation, which is what the authors of [24] propose to do using recognition of textual entailment (RTE). This kind of metric that really includes semantics in its assessment of quality would probably obtain better results on our RBMT system. Ideally, we should use metrics calculating the resemblances of the output not on a surface level but more deeply on the semantic representation inspired from [22] but the drawback of such a metric would be that it can not be generalized to other systems' output.

8 References

1. Bouillon, P., Rayner, M., Chatzichrisafis, N., Hockey, B.A., Santaholma, M., Starlander, M., Isahara, H., Kanzaki, K., Nakao, Y.: A generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation. In: Tenth Conference of the European Association of Machine Translation, pp.50-58. Budapest, Hungary (2005).
2. Bouillon, P., Halimi, S., Nakao, Y., Kanzaki, K., Isahara, H., Tsourakis, N., Starlander, M., Hockey, B.A., Rayner, M.: Developing Non-European Translation Pairs in a Medium-Vocabulary Medical Speech Translation System. In: 6th International Conference on Language Resources and Evaluation, pp. 1741-1748. Marrakech, Morocco (2008).
3. Starlander, M., Bouillon, P., Flores, G., Rayner, M., Tsourakis, N.: Comparing two different bidirectional versions of the limited domain medical spoken language translator MedSLT. In: 12th annual conference of the European Association for Machine Translation, pp. 174-179. Hamburg, Germany (2008).
4. Starlander, M., Estrella, P.: Relating recognition and translation quality with usability of two different versions of MedSLT. In: Machine Translation Summit XII, pp.324-331. Ottawa, Ontario, Canada (2009).

5. Callison-Burch, C., Osborne, M.: Re-evaluating the role of BLEU in machine translation research. In: 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 249-256. Trento, Italy (2006).
6. Google translate site, <http://translate.google.com/>
7. European Parliament Proceedings Parallel Corpus site, <http://www.statmt.org/europarl>
8. Rayner, M., Estrella, P., Bouillon, P.: Bootstrapping A Statistical Speech Translator From A Rule-Based One. In: Second Workshop on Free/Open-Source Rule-Based Machine Translation, pp.21-28. Barcelona, Spain (2011).
9. Nuance Communications: Nuance Grammar Developer's guide, version 8.5, Menlo Park, CA, USA (2003).
10. Bouillon, P., Flores, G., Starlander, M., Chatzichrisafis, N., Santaholma, M., Tsourakis, N., Rayner, M., Hockey, B.A.: A Bidirectional Grammar-Based Medical Speech Translator. In: Workshop on Grammar-based approaches to spoken language processing. ACL 2007, pp. 41-48. Prague, Czech Republic (2007).
11. Bouillon P., Rayner M., Novellas Vall, B., Starlander, M., Santaholma, M., Nakao, Y., Chatzichrisafis, N.: Une grammaire partagée multi-tâche pour le traitement de la parole : application aux langues romanes. In: TAL (Traitement Automatique des Langues), vol. 47, no. 3, pp. 155-173. Hermes and Lavoisier, Paris, France (2007).
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318. Philadelphia, USA (2002).
13. Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J.: Findings of the 2009 Workshop on Statistical Machine Translation. In: Fourth Workshop on Statistical Machine Translation, pp.1-28. Athens, Greece (2009).
14. Hartley, A., Popescu-Belis, A.: Évaluation des systèmes de traduction automatique. In: Chaudiron, S. (ed.) Évaluation des systèmes de traitement de l'information, pp. 311-335. Hermès, Paris, France (2004).
15. Snover, M., Dorr, B. J., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: 7th Conference of the Association for Machine Translation in the Americas, pp. 223-231. Cambridge, Massachusetts, USA (2006).
16. Snover, M., Madnani, N., Dorr, B. J., Schwartz, R.: Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In: Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics, pp. 259-268. Athens, Greece (2009).
17. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: ACL-2005: Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65-72. University of Michigan, Ann Arbor, 2005.
18. Boitet, C., Bey, Y., Tomokio, M., Cao, W., Blanchon, H.: IWSLT-06: experiments with commercial MT systems and lessons from subjective evaluations. In: International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation, pp. 23-30. Kyoto, Japan (2006).
19. Rayner, M., Bouillon, P., Chatzichrisafis N., Santaholma, M., Starlander, M.: MedSLT: A Limited-Domain Unidirectional Grammar-Based Medical Speech Translator. In: First International Workshop on Medical Speech Translation, HLT-NAACL, pp. 44-47. Omnipress Inc. New-York, USA (2006).
20. AgreeStat Excel VBA program site , <http://www.agreestat.com/agreestat.html>

21. Hamon, O., Fügen, C., Mostefa, D., Arranz, V., Kolss, M., Waibel, A., Choukri, K.: End-to-end evaluation in simultaneous translation. In: 12th Conference of the European Chapter of the ACL, pp. 345-353. Athens, Greece (2009).
22. Rayner, M., Estrella, P., Bouillon, P., Halimi, S.: Using Artificial Data to Compare the Difficulty of Using Statistical Machine Translation in Different Language-Pairs. In: Machine Translation Summit XII, pp. 300-307. Ottawa, Ontario, Canada, (2009).
23. Giménez, J., Màrquez, L.: The UPC Participation at the Metrics MATR Challenge 2008. In: Metrics MATR Workshop at AMTA'08 Machine Translation, Waikiki, Hawai'i (2008).
24. Padó, S., Cer, D., Galley, M., Jurafsky, D., Manning, C.D.: Measuring machine translation quality as semantic equivalence: A metric based on entailment features. In: Machine Translation 23, 2-3, September, pp. 181-193. Kluwer Academic Publishers Hingham, MA, USA (2009).