Spoken Dialogue Translation Systems Evaluation: Results, New Trends, Problems and Proposals

Hervé Blanchon, Christian Boitet, Laurent Besacier

Laboratoire CLIPS
BP 53
38041 Grenoble Cedex 9, France
{Herve.blanchon, Christian.boitet, Laurent Besacier}@imag.fr

Abstract

It is important to evaluate Spoken Dialogue Translation Systems, but as we show by analyzing evaluation methods in the Verbmobil, C-STAR II, and the Nespole! projects, the current state of the art is not fully satisfactory. Subjective methods are too costly, and objective methods, although cheaper, don't give good indications about usability. We propose some ideas to improve that situation.

1. Introduction

MT evaluation is a hot topic since 1960 or so, it may have several goals [3]. Speech-to-speech translation has been an active research field since 1986. We focus here on spoken dialogue translation, started with the C-STAR I project (1990-93), and its evaluation, which has become an important issue.

In the automatic text translation community, first evaluation techniques proposed were subjective and relied on human judgment about the translation quality [9]. In the subjective evaluation setting, an candidate target translation is compared with the original source utterance or a handcrafted translation reference. The main practical problem with subjective evaluation is that it is a time consuming task.

Hence, the text translation community moved towards automatic, objective, evaluation techniques in order to overcome this problem. In the objective evaluation setting, system outputs are compared with an handcrafted translation reference and several paraphrases.

However, these techniques are not as useful as expected to judge quality and seem useless to judge usability. In this paper, we try to find better ways to evaluate Spoken Dialogue Translation Systems.

For this, we analyze evaluations conducted within three projects: Verbmobil, C-STAR II and Nespole!. We also give an overview of the current objective evaluation metrics used by the community. We report and comment, as well, the first C-STAR III pilot evaluation that used those objective metrics on the BTEC corpus. In the last section of this paper we comment some limits of the current evaluation paradigms, and, in order to go further, we make some proposals.

2. Current SDTS evaluation methods

2.1. Evaluation within verbmobil

The verbmobil system [27] is a large demonstrator providing English, German and Japanese mobile phone users with simultaneous dialog interpretation services for appointment scheduling, travel planning and remote PC maintenance.

2.1.1. Data and protocol

In 1999 [23], potential users with English and German as mother tongue were put in a realistic end-to-end situation about negotiating an appointment. A supervisor listened to the conversation and solved all technical problems. The users and the supervisor had each to fill a form describing their (un)successful interactions on all the topics (13 topics were proposed) they touched on. 45 dialogues were collected and transcribed to provide references for the evaluation.

The forms were used to compute a dialogue success rate. For the linguistic evaluation both monolingual and bilingual data were considered. Monolingual inputs were evaluated according to their syntactic and semantic correctness, and possible misunderstandings. For the translations themselves, – mismatch (yes/no), – soundness (yes/no), and – quality (good, intermediate, poor) were evaluated.

2.1.2. Results

The average percentage of successful task completion reached 86.8%. If the results are weighted by the frequency of the attempted tasks in the dialogues, the success rate reaches 89.6%.

As far as the quality of the translation is concerned, results have shown that the quality is more sensitive to the insertion of information elements than to their deletion. Thus, if at least 50% of information elements are preserved in translation, scores over "poor". On the opposite, an insertion of more than 20% of information elements results in "poor" scoring over "good+intermediate".

In 2000 [23], a mass evaluation was conducted using 5069 German and 4136 English input turns. The percentage of 'approximately correct' translations was mapped against the word accuracy rate in speech recognition. The term 'manual selection' refers to a manual selection of the best translation produced by the different translation engines implemented in the system.

Word Accuracy Rate	>50%	>75%	>80%
GER-ENG *# of turns*	*5069*	*3267*	*2723*
Automatic selection	57	66	68
Manual selection	88	95	97
ENG-GER *# of turns*	*4136*	*3254*	*2291*
Automatic selection	53	58	60
Manual selection	86	92	94

Table 1: results of the Verbmobil mass evaluation

2.1.3. Comments

The evaluation conduced in 1999 gives a precise idea of the translation quality according to speech recognition errors, insertions, and deletions. Using a less fine-grained approach, the 2000 evaluation has also confirmed that, the better the speech recognition hypothesis the better the translation. We will see in section 3 that the NESPOLE! project reached almost the same scores for the same recognition qualities. The excellent results in manual selection announced great potentialities of the system.

2.2. Evaluation within C-STAR II

The C-STAR II systems [2] provided English, German, Italian, French, Korean and Japanese videoconference users with simultaneous dialog interpretation services for the tourism domain (transportation and accommodation booking, and sightseeing). Both users were able to play a customer or a tourist agent. Evaluation was not a key point of the project. However, ATR conducted an original and interesting evaluation [21].

2.2.1. Data and protocol

The test set consists of 330 Japanese turns extracted from 23 dialogues. Manual transcriptions are used as references. The goal is to compare human translations (from Japanese to English) with the translations produced by the system. Knowing the TOEIC scores of the Japanese translators, the TOEIC score of the system is measured.
Japanese subjects were asked to produce a written translation from each spoken turn. Evaluation sheets were produced with the Japanese transcription and, in random order, the system and human English translations. Native English speakers able to understand written Japanese were asked to evaluate the translations quality, using a 4-point scale (perfect, correct, acceptable, un-understandable), and to select the best one among them.

2.2.2. Results

Evaluation results show that the system performs better than human translators with a TOEIC score between 300 and 400 and performs worse than humans with a score equal to 800. A regression analysis confirmed that the system TOEIC score is 707.6. The same experiment was conducted using the speech recognition hypothesis as the Japanese sentence. In this latter case, the TOEIC score of the system was 548 (a 150 points loss).
For the same system, an automatic (objective) evaluation was conducted using a set of English paraphrases (14.4) for each Japanese turn. The TOEIC scores computed, using a regression method, were 682.9 for the references and 547.3 for the hypothesis.

2.2.3. Comments

ATR subjective evaluation proved a strong correlation between a costly subjective evaluation and a cheaper objective evaluation both based on the TOEIC scale. The problem is that the TOEIC scoring is not directly related with the usefulness of a SDTS.

3. Evaluation within NESPOLE!

The NESPOLE! system [10] provides English, German, French clients and Italian tourist agents with simultaneous dialog interpretation services for the tourism domain over Internet. Two showcases were evaluated en 2001 and 2002 in order to evaluate the performances of the demonstrators and the progress accomplished.
The lingware architecture used within the project is a pivot-based approach. Our pivot, called IF (Interchange Format), is based on domain actions (DAs) that consist of a speech act and concepts. In addition to the DA, an IF representation contains arguments. When analyzing a speech turn, several IFs can be produced to represent roughly each sentence. Each segment of a turn mapped to a unique IF is called an SDU (Semantic Dialogue Unit).

3.1. 2001's showcase-1 on "restricted tourism"

3.1.1. Data and protocol

Four dialogues (2 for summer vacations and 2 for winter vacations) were randomly picked-up from the first NESPOLE! data collection [5] for each language. For Italian, tourist agent turns were used. For English, French and German, client turns were used.
Evaluation was done at two levels: (1) *Hypos*, which are the automatic transcriptions produced by the Automatic Speech Recognition (ASR) modules, and (2) *Refs*, which are the manual transcriptions of the same speech signals. Each turn was also manually split into Semantic Dialogue Units (SDU) in order to get a SDU-based (and not a turn-based) evaluation of the translation quality.
ASR modules were evaluated using the Word Accuracy Rate (WAR) score. However, WAR does not allow to measure precisely how speech recognition errors influence translation quality. We also graded the *Hypos* as paraphrases of the *Refs*, at the SDU level, to measure the loss of semantic information due to recognition errors.
We performed monolingual evaluation (where the generated output language was the same as the input language), as well as crosslingual evaluations. For crosslingual evaluations, translation from English German and French to Italian was evaluated on client utterances, and translation from Italian to each of the three languages was evaluated on agent utterances.
For each set, we used three human graders with bilingual abilities. Each SDU was graded as either "Perfect" (the meaning is translated correctly and output is fluent), "OK" (the meaning is translated almost correctly but output may be disfluent), or "Bad" (the meaning is not properly translated). We calculated the percentage of SDUs in each of these three categories. "Perfect" and "OK" were also merged into a larger category of "Acceptable" translations. Average percentages were calculated for each dialogue, each grader, and separately for client and agent utterances. Combined averages for all graders and for all dialogues were then computed for each language pair.

3.1.2. Results

Table 2 combines all the results (in %) for acceptable translations using average score. Majority is reported when computed.

3.1.3. Comments

Performances of the ASR modules for producing *Hypos as paraphrases* are almost the same regardless the WAR.
The results indicate acceptable monolingual translations (clients and agent turns) in a range of 40-48%. of SDUs on *Hypos*. On *Refs*, the scores are, not surprisingly, better (46-

61%). For crosslingual translation towards Italian (on clients turns only), there is a performance drop (higher on *Refs* than on *Hypos*) compared with the monolingual systems. It shows that either the Italian generator does not handle properly some IFs produced by the French, English and German analyzers (problem of coverage) or that there in an intercoder agreement problem across sites. The same problem occurs for crosslingual translation from Italian (on agent turns only). The performance drop is higher than on client turns. The same reasons explain the phenomenon. However, the problem of coverage is probably dominant in this case. For the French generator, we could indeed check that the latter is true.

As references simulate a 100% speech recognition success rate, the translation scores on *Refs* for the four monolingual end-to-end systems must be considered as upper bounds for the scores on hypothesis. However, we found out that the behaviour of the systems is not a linear function of the hypothesis as paraphrase rate. If it had been the case, for French, the percentage of acceptable translations on *Hypos* would have been 35% for 65% of *Hypos as paraphrases*. The actual score (41%) is 6 points higher than "expectation". Figures are almost the same for all the four monolingual systems.

For the three crosslingual systems towards Italian (client turns), the situation is the same. We observed a 5 points increase. The analyzers in the monolingual and crosslingual systems towards Italian are the same for each source language. Thus, we may say that those scores are better than expected thanks to the analyzers "robustness".

When checking the scores for the three crosslingual systems from Italian (1 Italian analyzer and 3 generators), we get unclear results: the French and German generators do not reach expectation on hypothesis by 1 or 2 points but the English generator scores over expectation by 5 points. We can only conclude that the generators of French and German are not as robust as those of English and Italian. Maybe this due to the fact that the IF is, in a way based on English and mostly defined by CMU and IRST.

ASR *WAR*	71	62	64	77
Hypos as paraphrases	*65*	*66*	*68*	*70*
Majority vote	64	—	—	—
Mono-lingual trans.	*F-F c*	*E-E c*	*G-G c*	*I-I a*
on Refs/Hypos	*54/41*	*48/45*	*46/40*	61/48
Majority vote	51/38	—	—	60/44
Cross-lingual trans.	*F-I c*	*E-I c*	*G-I c*	
on Refs/Hypos	*44/34*	*55/43*	*32/27*	
Majority vote	38/31	—	—	
	I-F a	*I-E a*	*I-G a*	
on Refs/Hypos	*40/27*	*47/37*	*47/31*	
Majority vote	38/26	46/35	45/20	

Table 2: results of the NESPOLE! first showcase evaluation

3.2. 2002's showcase-2a "Extended Tourism"

The second showcase evaluation methodology has been designed to overcome some problems of our first evaluation.

3.2.1. Data and protocol

For each language, two unseen dialogues were picked up from the second NESPOLE! data collection [16]. The dialogues focused on additional scenarios such as tours of castles and lakes. The evaluation data sets were of the same kind as those used in the first showcase evaluation.

This evaluation also includes a comparison of the Showcase-1 components and the Showcase-2a components. The Showcase-1 components were frozen and saved after the Showcase-1 evaluation. The Showcase-1 components were then run on the Showcase-2a evaluation data in order to have a comparison of the two systems on the same data. In this evaluation, the Showcase-1 system was only run on transcribed input.

In this evaluation, we departed from our previous grading methodology in several ways. First, the 3-point scale (perfect, OK, bad) was replaced with a 4-point scale, based only on meaning preservation, taking neither fluency nor grammatical accuracy into account. Second, whereas we previously reported average scores across graders for each SDU, we calculated majority scores as well as averages. The majority votes are generally close to the averages, except where there is an outlier (a grader who was exceptionally harsh or lenient), but this problem did not occurred. Third, the graders for this evaluation were last-year students in a school for translators. Previously, graders had no special training in translation, and the groups were less homogeneous in terms of education and of second language knowledge than this year.

3.2.2. Study on graders agreement

To establish the stability and coherence of our evaluation scheme, it is important to have a good measure of how well different human graders agree on scoring the same output, and also how consistent the graders are over time.

Graders were first all trained on the same data set. They were given grading instructions and a grading training set. They graded this training set and we discussed their grading in order to finally have them agree on the same score for each SDU in the set.

Graders were also given two copies of the same grading check file they had to grade before and after they graded the actual test sets. This allowed to check: (1) their mutual agreement on this check set before and after the actual grading task, (2) their consistency over time.

In order to evaluate intercoder agreement before and after the task, we made 3 categories: (1) the 3 graders fully agree on the same grade, (2) the 3 graders agree on 2 grades that fall in the same final category (there is a majority vote), (3) the graders do not agree on the same final category (there is not majority vote). We got the following figures.

	(1) Agreement	(2) Majority	(3) no majority
Before	71	28	1
After	73	27	0

Table 3: intercoder agreement – It or Fr towards Fr

	(1) Agreement	(2) Majority	(3) no majority
Before	88	15	0
After	75	25	0

Table 4: intercoder agreement – Fr towards It

As far as consistency over time is concerned, we made 3 categories: (1) the grader gave the same grade, (2) the grader gave different grades that fall in the same final category (acceptable or not), (3) the grader gave different grades that do not fall in the same final category. We got the following figures.

	(1) Same grades	(2) grades in same category	(3) grades in ≠ categories
Grader 1	58	27	17
Grader 2	83	13	6
Grader 3	65	18	19

Table 5: consistency over time – It or Fr towards Fr

	(1) Same grades	(2) grades in same category	(3) grades in ≠ categories
Grader 4	83	13	6
Grader 5	102	0	0
Grader 6	58	27	17

Table 6: consistency over time – Fr towards It

We noticed that over time, graders tend to be more severe. When they changed their mind, either they degraded their grade in the same category or they changed their grade from acceptable to unacceptable.

3.2.3. Results

The following table combines all percentage results for acceptable translations using majority score.

ASR WAR	58	56	51	76
Hypos as paraphrase	60	67	62	76
Mono-lingual trans.	*F-F c*	*E-E c*	*G-G c*	*I-I a*
on Refs (01)| Refs/Hypos (02)	*69*| 77/58	*68*| 68/50	*45*| 61/51	*36*| 51/42
Cross-lingual trans.	*F-I c*	*E- I c*	*G-I c*	
on Refs (01)| Refs/Hypos (02)	*72*| 77/58	*64*| 70/50	*44* x/x	
	I-F a	*I-E a*	*I-G a*	
on Refs (01)| Refs/Hypos (02)	*19*| 37/33	*33*| 33/30	*38*| 45/38	

Table 7: results of the NESPOLE! second showcase evaluation

3.2.4. Comments

The results indicate acceptable monolingual translations (on clients and agent turns) in a range of 42-48% of SDUs on *Hypos*. On *Refs*, the scores are, not surprisingly, better (51-77%). For crosslingual translation towards Italian (on clients turns only), there is no performance drop compared with the monolingual systems. In this second showcase, the Italian generator handles correctly all IFs produced by the French, English and German analyzers. There no problem of IF coverage or intercoder disagreement across sites. Those problems still occur for crosslingual translation from Italian (on agent turns only). The problem of coverage is dominant in this case.

On the verbmobil 2000 mass evaluation (Table 1), for a WAR<75% the results are 75% for German-English and 66% for English-German. For this second showcase, NESPOLE! reached the same level of performance.

3.3. Accomplished progress

One noticeable observation is that translation performance from Italian is significantly lower than translation into Italian. This is mainly due to the characteristics of the evaluation data: agent utterances (translated from Italian) are more complex, and in some cases are actually out of domain, while client sentences (translated into Italian) are on average shorter, easier and in domain.

In order to quantify this difference and assess its effect on our results, we asked system developers to manually classify the SDUs in the test data into three categories: (1) falls within the domain of coverage of Showcase-1; (2) falls within the domain of coverage of Showcase-2a; and (3) out of domain. We then calculated the performance results for the three groups of SDUs separately.

Interestingly, for English, German and French input (client data), we discovered that only a very small number of SDUs were classified in either group-2 or group-3 (less than 5 SDUs for each). Thus, the data is overwhelmingly within the domain of the Showcase-1 system. For the Italian input (agent data), however, 13% of SDUs were classified within the domain of Showcase2a (group-2), and 25% of SDUs were out of domain (group-3).

The difference in system performance on these three separate categories is also quite insightful.

On the group-1 data, we see an improvement in performance between the results of the showcase-1 system and the showcase-2a system: from 56.6% to 63.2% acceptable translation on transcribed input. This demonstrates improvements in domain coverage in the showcase-2a system (within the domain of showcase-1).

On the group-2 data, the difference is much more pronounced. The showcase-1 system achieves only 14.2% acceptable translations, while the showcase-2a system achieves 38.4%. The showcase-1 system was not designed to cover this type of data, so this is not surprising. While the showcase-2 system performs much better on this data, it did not reach the same level of performance as for the showcase-1 domain.

The Italian system has no coverage of the out of domain SDUs. When excluding these SDUs from consideration, the performance figures are 49.3% for the showcase-1 system and 58.9% for the showcase-2 system – more similar to the results we find for the client input data (English, German and French).

4. Current trends, problems and proposals

Objective (automatic) evaluation techniques are now broadly used for automatic text translation.

4.1. Current main trends: a quick overview

The cost of subjective evaluation techniques motivated the shift towards objective techniques. Quite a lot of techniques are available today.

A first family of metrics is based on edit distance [15, 26]. [20] proposed edit distance between the candidate translation and reference translations. One the same track, [18] introduced a length-normalized edit distance, called Word Error Rate (WER) between the candidate translation and reference translations. [1] used a combination of different edit distances to rank the output translation. [11] introduced a related measure called position-independent word error rate (PER) that does not consider word position but use a bag of words instead. As an alternative [17] finally proposed accuracy measures to compute similarity between candidate translation and reference translations in proportion to the number of common words among them.

BLEU [19] and NIST [7], almost standard benchmarks for the domain, compute statistical distances for n-grams

between the translation produced by the system and a gold translation associated with a set of paraphrases.
We can also cite the GMT [24] score which is the harmonic mean (F-measure) of a new proposed precision and recall measures based on a maximum match size between a candidate and a reference translation.
Recently, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [12] framework, proposed to automatically determine the quality of summaries, has also been used for MT evaluation [13]. The same authors also proposed ORANGE (Oracle Ranking for Gisting Evaluation), for evaluating evaluation metrics for machine translation [14].

4.2. A new evaluation framework from C-STAR III

Nonetheless, we would like to promote comparative evaluation on the same data set [22] in the C-STAR III framework. This would enable us to tackle some of the questions raised above.

4.2.1. Pilot closed evaluation (2002)

The C-STAR III partners ran a pilot evaluation experiment in year 2003 on our common BTEC corpus for two conditions. Development and test data were picked up for BTEC. For development purposes every kind of resources could be used. The test set consisted of 500 English sentences that had their translation into Italian, Japanese, Korean and Chinese within BTEC. Subjective evaluation followed the Linguistic Data Consortium evaluation guidelines for the DARPA TIDES[1] project. BLEU and NIST scores were calculated on the rough systems output.
Under the primary condition, systems were going from Italian, Japanese, Korean and Chinese to English. Both subjective and objective (with BLEU and NIST) evaluations were conducted for five systems (labelled S1 to S5) on this condition.

	Adequacy [0..5]	Fluency [0..5]	BLEU [0..1]	NIST [0..∞[
S1	4,00 (1)	3,76 (2)	0,6620 (1)	10,5706 (1)
S2	3,92 (2)	4,03 (1)	0,5820 (2)	6,5565 (2)
S3	3,01 (4)	2,81 (4)	0,3153 (4)	5,8889 (3)
S4	2,59 (5)	2,30 (5)	0,2733 (5)	5,6830 (4)
S5	3,21 (3)	3,74 (3)	0,5542 (3)	3,4013 (5)

Table 8: results of the C-STAR III pilot evaluation under the primary condition

Under the secondary conditions, systems were going from Chinese to English. Objective evaluation only was conducted on three systems on this condition.

	BLEU [0..1]	NIST [0..∞[
S1	0.5542 (1)	3.4013 (3)
S2	0,3884 (2)	8.1383 (1)
S3	0.2733 (3)	5.6830 (2)

Table 9: results of the C-STAR III pilot evaluation under the secondary condition

[1] http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess02.pdf

Free Systran® MT systems available on the web have been evaluated on the same test set. All systems scores were inferior to those of the C-STAR partners. However, nothing was done to tune them to that particular task, although that is possible on systranet: choice and order of dictionaries, handling of capitalized and unknown words, polite and direct forms of address (please … / I would like you to…).

4.2.2. Lessons learned

For the primary condition, there is an inconsistent ranking for system 5 with BLEU (3rd) and NIST (5th). This system outputs are significantly shorter than the output of the other systems. It impact strongly on the results with a brevity penalty for NIST.
A fourth parameter came to us when checking the actual strings produced by the each system. We found different use of case ("Tokyo", vs. "tokyo"), punctuation ("juice, please" vs. "juice please"), digits (spelled-out vs. numerals), abbreviations ("OK" vs. "okay"), compound words ("duty-free" vs. "duty free"), sentence boundaries, and special characters.
The MT outputs and the references translation were normalized. We observed differences of ±0.15 on the BLEU scores and ±1.8 on the NIST scores. Those differences are quite important. In the future, system outputs and references will be normalized for common evaluation.
Under the secondary condition, the ranking is still inconsistent with BLEU and NIST.
For both conditions, C-STAR III systems outperformed Systran systems available free of charge on the web.

4.2.3. Open evaluation in 2004

Following this pilot evaluation, the C-STAR III consortium decided to initiate an open evaluation campaign on the BTEC corpus. We hope this opportunity will provide answers to some of the questions we raise in the next section.

4.3. Problems and proposals

4.3.1. Problems

4.3.1.1 Current metrics don't measure linguistic quality

A first problem is that the figures produced with these techniques are not directly interpretable in terms of translation quality[2]. A lot of work has been done to correlate objective evaluation results with subjective evaluation results [6, 7, 19], but the results are inconsistent. BLEU is said to correlate well with human judgments of quality, NIST is said to be better than BLEU for theoretical reasons, but BLEU and NIST give contradictory rankings (see above). Hence, if correlation with human judgments is a measure of the quality of a metrics, NIST cannot be better than BLEU… or the correlation is too weak to be meaningful.
Another trouble is that, as reported in ACL-03, these metrics often give quite bad scores to high quality human translations. An experiment has been reported, where each of 15 (human) paraphrases had been tested, with the 14

[2] The question of linguistic quality is related to what FEMTI [8] calls Functionality (section 2.2.1 of the classification), and more precisely to the subsections 2.2.1.1 Suitability, 2.2.1.2 accuracy, and 2.2.1.3 Wellformedness.

other proposals used as Refs. The scores were not as bad as those of MT outputs, but bad enough, although they were perfect translations. The problem here is that the Refs are always quite sparse in the space of all possible perfect or very good translations, if distance is measured by these metrics (or by WER, WPER etc.).
A solution would be to ask human posteditors to review MT outputs with the least possible effort, leading to (one of) the "nearest" possible correct equivalent translation. The scores would certainly better reflect the quality of the MT output, in that they would in some way measure the minimal effort needed to produce a polished translation from the output. With that approach, good human translations would always get a perfect score.
A disadvantage of that solution is that, although the scoring process would remain automatic, the evaluation process would require more human work: not only produce a certain number of Refs, but postedit each MT output.
A frequent answer to our objections is that those metrics are in any case useful to evaluate progress during the development process. We would like to challenge that opinion on the ground that "progress" is defined in a circular way. What is measured is simply the progress according to the abovementioned metrics, which don't measure linguistic quality. Also, in any case, measuring the linguistic quality is certainly good for developers, but not for users. And the ultimate goal of building SDT systems is to have humans use them, isn't it?

4.3.1.2 Current metrics don't measure practical usability

A sad fact about translation is that linguistic quality does not correlate with practical usability. As early as 1972, a usability report on Systran (Russian-English) used at Euratom (Ispra) gave it a quality score of about 1/5 and a usefulness score of 4.5/5. In the case of speech, it is notorious that transcripts from interpreted monologues or dialogues are judged as very poor translations, although interpreting is very difficult and well paid… because the result is very useful.
Usefulness can only be measured on a working system. During development, we can still aim at measuring usability, to estimate future usefulness. But, in the current evaluation paradigms using exclusively objective evaluations methods,

- usability is not and cannot be measured,
- the real data is not accessible to the final users of the evaluation, who cannot form their own "subjective judgment" and then interpret the scores according to their perception (of quality or usability).

In a few talks where the authors were present and real translation examples were given, a sizable part of the audience did not agree with the judgment of quality, either because it was wrong, or because the feeling was that the translations, although wrong, were felt to be perfectly adequate for users to understand the meaning and take appropriate action.
It is then be important to concentrate also on usability issues, and to try to correlate usability with objective and/or subjective metrics. To our knowledge, this problem has never been tackled for SDT except in Verbmobil.

4.3.1.3 Problems of comparisons with commercial systems

It is also common that systems are evaluated against other systems, in particular off the shelf ones. The reasoning is that, if the systems under development fare better, they are linguistically better, and hence more usable in the future. First,one can argue against the fairness of such comparisons, because the results of objective evaluation are very sensitive to several parameters. The commercial systems used for comparison should at least be parameterized to give the best they can on the corpus at hand, or not used for comparison if they are clearly specialized for quite different tasks: while METEO by far outperforms junior translators on weather bulletins, it will always be very bad on tourist sentences!
The first parameter is the “style” of the reference translations which the output of the systems will be compared with. For example, how can we compare a verbose (wordy) system and a concise one? Thus, what are the “good” references? How many do we need?
The second parameter is the domain the system has been developed for. In particular, how can we compare a system tuned for a particular domain with a broad coverage system applied to a test set from this particular domain?
The third parameter is the granularity of the comparison. MT output and reference translations can be compared at various levels (words, POS, inflectional attributes). At which level do we want to evaluate, and why?
Second, the hope to measure usability in this way is quite vain. Usability depends crucially in the ergonomics of a whole, integrated system or service. For example, brute force (string to string) statistical MT is extremely slow compared to commercial systems based on procedural programming or on quasi-deterministic ATNs or DCGs or the like, and processing time grows very fast with the length of the translated utterance. But computing time is not taken into account. Another important feature is the flexibility; can the system in some way learn from usage, accept user dictionary items, etc.?

4.3.1.4 Other features should also be evaluated

In the MT evaluation study prepared by H.Nomura for JEIDA (now JEITA) in 1993, many other features of MT systems were evaluated. Some of them concerned the potential of the systems, such as the linguistic and computational internal workings of the system, and the ease to maintain and improve the linguistic data, be they symbolic or numerical or both.
Two features of the linguistic architecture of speech-to-speech translation systems are important for evaluating their future potential: the use of the context and the richness of the data structures manipulated.
The Verbmobil architecture includes these two features: the system makes extensive use of the dialogue context, and uses a rich interface structure between the different modules.
By contrast, the architecture of other systems (C-STAR II and III, NESPOLE!) is fairly trivial. Each spoken utterance is translated in isolation, ignoring the context. The information passed from one module to the other is minimal within a black-box integration approach.
We would like to stress the need for component integration just like Verbmobil did.

4.3.2. Proposals for evaluation

As said above, good evaluations should mix several criteria, hence several metrics, and the evaluation process cannot be fully automatic, even with "objective" methods (such as BLEU).

4.3.2.1 Evaluating potential and actual linguistic quality

Measuring architectural features (and hence, in a way, upper quality limit) can only be done by specialists, but its cost is limited.

Measuring actual linguistic quality in a meaningful way requires a lot of human time, but a large part of it could be "mutualized". We are indeed building a web-oriented platform, PolyphraZ, to display, translate (by several MT engines), edit (human postedition or simply translation), and then grade parallel corpora of multilingual "polyphrases". A polyphrase corresponds to the meaning of one original utterance in some language. It is a 1-line table where each column contains a set of homogeneous "proposals", such as Refs (paraphrases) in each language, or results of some MT systems (where the proposals correspond to different versions and/or different parameters), or reobjective scores (BLEU, NIST, etc.), or various edit distances, etc.

4.3.2.2 Evaluating usability

Perhaps the only possibilty to measure usability is to put prototypes on the web, integrated in a bare-bone but usable service, constantly available, and offering either MT or HT (done in Wizard of Oz mode, or in "real person" mode), or combinations. It will then be possible to measure the efficiency of each setting as the time it takes interlocutors to achieve their goal, using that setting (ST), as compared with a human interpreter: $\mathrm{Eff}_{\mathrm{rel}} = \frac{\mathrm{TimeTaskHum}}{\mathrm{TimeTaskST}}$.

An important point here is to develop a very flexible platform, so that developers can experiment with a large variety of settings.

4.3.2.3 Don't concentrate only on evaluation!

A last proposal on evaluation is… not to concentrate only on evaluation. If we consider the history of MT, it seems that a disproportionate amount of efforts (and of money!) has gone into evaluating MT systems, prototypes or even mockups than into building and improving systems. Metaevaluation (evaluating evaluation methods) has also become a hot research topic, generating many papers like this one, or at least a part of it or [14].

Returning to a multicriteria approach, and hence reintroducing measures of usability by building SDT prototypes, putting them to actual use, and evaluating many criteria, seems to be a good way to restore some balance between development and evaluation.

This approach should also lead developers to try alternate internal designs, rather than to converge (as now) on the designs maximizing some scoring methods. For example, real usage might show that a particular combination of automatic processing, user control, and interactive disambiguation, would be more efficient than any current setting.

4.3.3. Proposals for overall system architecture

To improve the overall quality of the system, several classes of improvements have to be considered: tighter component integration and using the context.

4.3.3.1 Component integration

Component integration may be achieved, for both handling the source input utterance and producing the target output utterance.

As far as the input is concerned, more rich information should be passed to the analysis module by the automatic speech recognition (ASR) module. We even propose a bidirectional exchange of information. The ASR module may pass a word lattice (possibly augmented with prosodic information) to the analysis module, or even a partially analyzed output [25]. On the other direction, the analysis module may pass the current topic of the discourse allowing the ASR module to switch, on the fly, among different language models. The analysis module may also pass a set of previously used words (form the previous utterance of the speaker and form the generated output of the other speaker last utterance). The weight of these words may be increased during the decoding stage.

As far as the output is concerned, the generator should provide a textual stream with annotations to the speech synthesis module. It seems counter productive to let the speech synthesis module compute information the generator knows.

4.3.3.2 Using the context

We distinguish between three kinds of context, global, dialogic and linguistic [4].

The global context contains at least:

- the general type of dialogue (reservation, enquiry, chatting, request for help…),
- the characteristics of the participants, in particular their names, sex, ages and relative politeness level,
- the roles of the participants (agent/client, doctor/patient, host/guest) and their current relation (unknown, friends, former teacher-student…),
- the names of their locations, beause they can be personified (as in "But Taejon has just told me that…").

Ideally, the dialogue context should contain:

- a representation of the past dialogue,
- the present stage of the dialogue if it follows some known script,
- and some predictions about the future.

In the short term, much could already be achieved if the analyzer could access a sorted list of speech acts predicted by a suitable dialogue model.

In this framework, analyzers should produce not only their usual output (IF or linguistic structure), but also the identified speech act, if not explicit in the output.

Utterances in natural dialogues contain many instances of anaphora and ellipsis. This considerably limits the output quality of the current analyzers, which handle utterances without information about the previous utterances even if the missing elements are present in a previous utterance of the same turn.

Context is also important for lexical disambiguation (e.g. "Je prendrai un express" → "I will take an espresso/an express train") and for consistent lexical selection from an utterance to the next.

The most necessary part of the linguistic context seems to be the list of possible "centers", that is, possible referents for anaphoric elements or ellipses (main context words such as nouns and verbs).

5. Conclusion

It is important to evaluate Spoken Dialogue Translation. We tried to give an overview of interesting results obtained using subjective evaluation techniques. We also showed that current state of the art objective evaluation

techniques are not fully satisfactory because they do not measure linguistic quality or practical usability. We also pointed out problems of comparisons with commercial systems. We proposed ideas for better evaluation and better SDTS architecture.

6. Acknowledgements

The authors would like to express their grateful thanks to their NESPOLE! and C-STAR partners for the richness of our cooperation on spoken dialogue translation during the last 8 years.

7. References

[1] Akiba, Y., Imamura, K., *et al.* (2001) *Using Multiple Edit Distances to Automatically Rank Machine Translation Output.* Proc. MT Summit VIII. Santiago de Compostela, Spain. 18-22 September, 2001. vol. **1/1**: pp. 15-20.

[2] Blanchon, H. and Boitet, C. (2000) *Speech Translation for French within the C-STAR II Consortium and Future Perspectives.* Proc. ICSLP 2000. Beijing, China. Oct. 16-20, 2000. vol. **4/4**: pp. 412-417.

[3] Blanchon, H., Boitet, C., *et al.* (2004) *Towards Fairer Evaluation of Commercial MT Systems on Basic Travel Expressions Corpora.* Proc. IWSLT'2004. Kyoto, Japan. September 30 - October 1, 2004. vol. **1/1**: pp. 6 p.

[4] Boitet, C., Blanchon, H., *et al.* (2000) *A way to integrate context processing in the MT component of spoken, task-oriented translation systems.* Proc. MSC-2000. Kyoto, Japan. October 11-13, 2000. vol. **1/1**: pp. 83-87.

[5] Burger, S., Besacier, L., *et al.* (2001) *The NESPOLE! VoIP Dialog Database.* Proc. Eurospeech. Aalborg, Denmark. September 3-7, 2001. 4 p.

[6] Coughlin, D. (2003) *Correlating Automated and Human Assessments of Machine Translation Quality.* Proc. MT Summit IX. New Orleans, USA. September 23-27, 2003. 8 p.

[7] Doddington, G. (2002) *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics.* Proc. HLT 2002. San Diego, California. March 24-27, 2002. vol. **1/1**: pp. 128-132 (note book proceedings).

[8] Hovy, E., King, M., *et al.* (2002) *Principles of Context-Based Machine Translation Evaluation.* in Machine Translation. vol. **17**(1): pp. 43-75.

[9] King, M. (1996) *Evaluating Natural Language Processing Systems.* in Communication of the ACM. vol. **29**(1): pp. 73-79.

[10] Lazzari, G. (2000) *Spoken Translation: Challenges and Opportunities.* Proc. ICSLP 2000. Beijing, China. Oct. 16-20, 2000. vol. **4/4**: pp. 430-435.

[11] Leusch, G., Ueffing, N., *et al.* (2003) *A Novel String-toString Distance Measure with Application to Machine Translation Evaluation.* Proc. MT Summit IX. New Orleans, U.S.A. September 23-27, 2003. 8 p.

[12] Lin, C.-Y. (2004) *ROUGE: A Package for Automatic Evaluation of Summaries.* Proc. ACL 2004. Barcelona, Spain. July 21-26, 2004. 8 p.

[13] Lin, C.-Y. and Och, F. J. (2004) *Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics.* Proc. ACL 2004. Barcelona, Spain. July 21-26, 2004. 8 p.

[14] Lin, C.-Y. and Och, F. J. (2004) *ORANGE: a Method for Evaluation Automatic Evaluation metrics for Machine Translation.* Proc. COLING 2004. Geneva, Switzerland. August 23-27, 2004. 8 p.

[15] Lowrence, R. and Wagner, C.-K. (1975) *An Extension of the String-to-String Correction Problem.* in Journal of the ACM. vol. **22**(2): pp. 177-183.

[16] Mana, N., Burger, S., *et al.* (2003) *The Nespole! VoIP Corpora in Tourism and Medical Domains.* Proc. EUROSPEECH 2003. Geneva, Switzerland. September 1-4, 2003. 4 p.

[17] Melamed, I. D., Green, R., *et al.* (2003) *Precision and Recall of Machine Translation.* Proc. HLT-NAACL 2003, companion Volume. Edmonton, Canada. May 27 - June 1, 2003. 3 p.

[18] Nießen, S., Och, F. J., *et al.* (2000) *An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research.* Proc. LREC 2000. Athens, Greece. 31 May - 2 June 2000. vol. **1/3**: pp. 39-45.

[19] Papineni, K., Roukos, S., *et al.* (2002) *BLEU: a Method for Automatic Evaluation of Machine Translation.* Proc. ACL-02. Philadelphia, USA. July 7-12, 2002. vol. **1/1**: pp. 311-318.

[20] Su, K.-Y., Wu, M.-W., *et al.* (1992) *A New Quantitative Quality Measure for Machine Translation System.* Proc. COLING-92. Nantes, France. 23-28 août 1992. vol. **2/4**: pp. 433-439.

[21] Sugaya, F., Takezawa, T., *et al.* (2000) *Evaluation of ATR-MATRIX Speech Translation System with Pair Comparison Method Between Human and System.* Proc. ICSLP 2000. Beijing, China. October 16-20, 2000. vol. **3/4**: pp. 1105-1108.

[22] Takezawa, T., Sumita, E., *et al.* (2002) *Towards a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World.* Proc. LREC-2002. Las Palmas, Spain. May 29-31, 2002. vol. **1/3**: pp. 147-152.

[23] Tessiore, L. and Hahn, W. (2000) *Functional Evaluation of a Machine Interpretation System: Verbmobil.* in Wahlster, W. (ed.), Verbmobil: Foundation of Speech-to-Speech Translation. Springer-Verlag. Berlin. pp. 611-631.

[24] Turian, J. P., Shen, L., *et al.* (2003) *Evaluation of Machine Translation and its Evaluation.* Proc. MT Summit IX. New Orleans, U.S.A. September 23-27, 2003. pp. 386-393.

[25] Vu Minh, Q., Besacier, L., *et al.* (2004) *Interchange format-based language model for automatic speech recognintion in speech-to-speech translation.* Proc. RIVF'04 (Recherche Informatique Vietnam-Francophonie). *To be published in a special issue of Studia Informatica Universalis [Suger Editor].* February 2-5, 2004. vol. **1/1**: pp. 47-50.

[26] Wagner, C.-K. and Fisher, M.-J. (1974) *The String-to-String Correction Problem.* in Journal of the ACM. vol. **21**(1): pp. 168-173.

[27] Wahlster, W. *ed.*, (2000) *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer-Verlag. Berlin. 677 p.