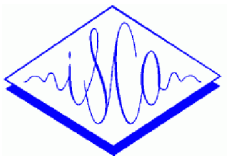

International Workshop on Spoken Language Translation (IWSLT) 2006
Keihanna Science City, Kyoto, Japan
November 27-28, 2006

IWSLT-06: experiments with commercial MT systems and lessons from subjective evaluations

Christian Boitet, Youcef Bey, Mutsuko Tomokiyo, Wenjie Cao, Hervé Blanchon

† GETA, laboratoire CLIPS, IMAG-campus
BP 53, 385 rue de la Bibliothèque
38041 Grenoble Cedex 9, France
{first.last}@imag.fr

Abstract

This is a short report of our participation to IWSLT-06. First, we let 2 commercial systems participate as fairly as possible (SYSTRAN v5.0 for CE, JE, AE, & IE, Atlas-II for JE), taking care of preprocessing and postprocessing tasks, and tuning as many "pairs" as possible by creating "user dictionaries" and finding a good combination of parameters (such as dictionary priority). Second, we took part in the subjective evaluation of CE results (fluency and adequacy). Details on experiments and methodological remarks are provided, with a perspective to introduce less expensive and more objective human- and task-related evaluation methods.

Introduction

MT evaluation is a hot topic since 1960 or so. The literature on evaluation may even be larger than that on MT techniques proper. MT evaluation may have several goals (i) help buyers buy MT system best suited to their needs (ii) help funders decide on which technology to support (iii) help developers measure various aspects of their systems, and measure progress.

The MT evaluation campaign organized by the C-STAR III consortium falls in the latter category. Its aim is to measure the "quality" of various MT systems developed for speech-to-speech translation when applied to the BTEC corpus [12]. Another goal is to compare the MT systems developed by the C-STAR partners not only between them, but also with other systems, notably commercial systems.

In past similar campaigns, the commercial systems used as a "baseline" were tested in quite unfair ways, shedding serious doubts on the results. According to reports, experimenters submitted input texts to free MT web servers, instead of running a commercial version tunable by various parameter settings and building of "user dictionaries".

For example, long ago, IBM CANDIDE system was trained intensively on the *Hansard* corpus, and then compared with an off-the-shelf version of SYTSRAN, without any tuning. SYSTRAN clearly won, but the margin might have been far bigger (or perhaps not, this should have been studied!), if SYSTRAN had been tuned to this totally unseen corpus, at the level of its dictionaries, of course, but perhaps also of its (procedural) grammars.

The Microsoft group also compared its French-English MTS system with SYSTRAN [9]. MTS was highly tuned to their documents (indeed, the transfer component was 100% induced from 150,000 pairs of sentences and their associated "logical forms" or deep syntactic trees). In this case, SYSTRAN was slightly tuned by giving priority to SYSTRAN dictionaries containing computer related terms[1]. However, MSR apparently did not invest time to produce a user dictionary containing Microsoft computer terminology. Considering that technical terminology varies a lot from firm to firm and even from product to product, what is then the value of the conclusion that their system was (slightly) better than SYSTRAN? Indeed, when they performed the same comparison on the HANSARD, SYSTRAN ("general") won, although they induced the transfer part from about 400,000 tree pairs.

Our interest in IWSLT-06 was also to progress towards better and cheaper evaluation methods, both "objective" and "subjective". It was quite interesting to participate in the "subjective" evaluation of CE, because it proved beyond doubt our suspicion that the current setting induces evaluators to sort the results on the same input instead of grading them independently, thus increasing the human evaluation time considerably (about N $\log_2$ N comparisons for N outputs).

Because of delays in August, we were unfortunately unable to postedit raw MT results, measure the human time, and run again the n-gram based measures after adding the postedited MT outputs to the reference translations.

Section 1 describes the experiment with commercial MT systems, section 2 the subjective evaluation, and section 3 contains some suggestions for better and cheaper task-related objective and subjective measures.

1. Experiments with commercial MT systems

We used SYSTRAN 5.0 for all IWSLT-06 pairs (CE, JE, AE, & IE). We used it already for JE at IWSLT-04 [3]. Since then, the CE and JE pairs have been slightly improved as a byproduct of a contract with CISCO on E-CJK. Unfortunately, only IE among the 4 pairs is one of the "good SYSTRAN pairs".

We contacted several firms producing reasonably good JE pairs (Fujitsu, IBM-Japan, Sharp, Toshiba…), some accepted to send us up-to-date versions, but in the end none did, so that we could only use a version of ATLAS-2 acquired about 2 years ago.

1.1. SYSTRAN v5.0

Architecture

The SYSTRAN architecture is a "descending transfer" sort [4].

a) Source language analysis step (MA and SA)

The morphosyntactic analysis module (MA) examines each sentence in the text input, noting all uncertainties and errors. It is based on finite-state transducers and produces a lattice of possible solutions, with one path selected by default (on the basis of statistics or preferences). This allows for reanalysis and decision-making on alternate paths in later processing (interactive disambiguation facility) and for user tuning.

Syntactic analysis (SA) is procedural and heuristic, leading to a unique solution expressed by a kind of multililevel dependency graph grounded on the path selected after MA. The program flow and basic algorithms for the SA module are essentially the same for all systems sharing the same

[1] This is not in the paper but what answered to a question.

source language, and the system design and architecture are the same for all language pairs. However, in the case of lexical and syntactic ambiguities, decisions are often taken with respect to the target language.

b) "Descending" transfer step (TS+TL+SG)

This step is different for each language pair. It is a combination of structural transfer at surface syntax level (ST) and of lexical transfer (LT). It seems that it first restructures the syntactic structure (a kind of chart) as necessary, and then selects the correct target lexical equivalents of identified words and expressions. Regardless of the fact that restructuring and selection are different, the basic architecture and strategy are similar for all language pairs. The output is a target surface tree.

c) Target language morphological synthesis (MG)

We call such architecture a "descending transfer", because there is no source language independent structural and syntactic generation phase (SG) — there are actually very few real "horizontal" transfer systems.

The morphological generation module (MG) performs all necessary string transformations to generate case, tense, number, etc. according to the rules of the target language.

d) Dictionaries

Two kinds of dictionaries are used: stem dictionaries and expression dictionaries. A stem dictionary contains terminology and base forms. An expression dictionary contains phrases and conditional expressions.

There is a good dictionary manager tool which has a level for helping naive users (not SYSTRAN lexicographers) develop (possibly multilingual) "user dictionaries", which are collections of subject-specific terms that are analyzed prior to being integrated directly into the translation process.

e) XML workflow

As a text undergoes the translation process, its various representations (initially plain text or html or xml or rtf etc., then linguistic graphs and trees, then again annotated text) are serialized in XML [10].

Tuning done on Systran

Preprocessing to be done to the training, development and test batches of turns including changing the encoding, and separating the turn ids from the text.

We used SYSTRAN batch facility, and tuned some parameters (list and priority of SYSTRAN-provided dictionaries, default politeness level, default gender of addressee, way of handling words beginning with a capital, having multiple translations, and unknown or without translations).

Finally, we developed user dictionaries from the list of unknown words obtained by running the system on the available corpora. Due to lack of time, this was done only for the CE and IE pairs.

a) Dictionary update for the Chinese to English system

SYSTRAN with original dictionaries found 400 NFW in the Chinese training corpus. We created a Chinese *user* dictionary containing these words and their English translation with the aid of a Chinese native speaker. The SYSTRAN system associated with this user dictionary found 12 unknown words in the test corpus. These words were further added to the user dictionary.

b) Dictionary update for the Italian to English system

We also created an Italian user dictionary with the same method. After the translation of the training data, the system detected around 1200 unknown words in the Italian training corpus and some tens of unknown words in the Italian test corpus. However, we did not have time (and competence) to handle all missing entries, so that our IWSLT-06 Italian user dictionary was not complete not of very high quality.

Comments

The IE training corpus seems to contain at some places English turns instead of their Italian translations.

For lack of time, the user dictionary was only partially constructed.

The structure of the user dictionaries is really too "string-oriented": for example, one must translate "potrebbe" by "s/he/it could", it is impossible to translate the lemma (infinitive form), indicate its conjugation code, and let the system do the rest. Even naive users should have access to that level. The reason lies in the adopted strategy (first pass in the user dictionaries at string level, before MA).

1.2. ATLAS-II

Architecture and general presentation

The ATLAS system has a "semantic pivot" architecture. It was designed and developed in the late 70's and 80's at Fujitsu by H. Uchida and his team. Its pivot is the "grandfather" of the UNL anglosemantic pivot [5]. There is no bilingual "transfer" step, only an analysis and a generation phase for each language[2].

Around 1982, it was put on the market for the EJ and JE pairs, while components for French, Spanish and German were also developed and demonstrated[3]. At that time, each dictionary contained only about 70,000 entries.

Since about 20 years, that system has been rated among the very best, or the best, for JE and EJ. Components for other languages have been developed[4], but not marketed. The size of dictionaries has gone up tremendously, thanks to corpus-based techniques, from 586,000 at MTS-01 to 1.5 M at ACL-03, to 5,440,000 technical terms categorized into 28 fields in ATLAS V13 (2006).

While ATLAS translation quality depends on the documents to be translated, high performance is obtained on well-structured sentences such as those of manuals, technical writings and articles.

Up to 32 dictionaries can be specified at the same time.

The MT system uses rule-based engines as well as a small Translation Memory (probably for personalized translations) where approximately 5,000 examples can be registered.

ATLAS is a standalone product, and Accela BizLingo is its intranet version. We did not use ATLAS V13, but a previous version with 890,000 bilingual dictionary entries.

Tuning done on ATLAS

a) No user dictionary nor translation memory

As said earlier, due to the period of the campaign, M. Tomokiyo did not have time to produce a user dictionary. Also, operating ATLAS on a PC with a Japanese interface was not easy without her, so that in the end we did not translate the full 40,000 training turns (about 1000 "standard pages"), although that would have taken only 4 hours (3 mn for 500 turns).

b) Preprocessing and post processing

There was quite a lot of preprocessing to perform, mostly related to segmentation problems.

---

[2] As they involve a change of "lexical space", they are more aptly called "enconversion" and "deconversion" in the UNL project.

[3] Prof. Hirakata from Stuttgart University developed an interesting layer of high-level programming language above the "native" tools.

[4] notably for Malay, Indonesian, Chinese and Thai, during the ODA CICC project (1987-93)

1. For some unknown reason, ATLAS inserted sentence breaks after some numbers if placed at the beginning of a turn and written in Japanese script and not with Arabic digits. They had to be removed.
2. The encoding of character set was tuned to support the default system encoding.
3. The output was filtered manually to produce clean English by removing the annotations and NFW which appeared in the raw translation.

1.3. Overview of tasks and supplied data

We had to handle 4 language pairs. For each, the CSTAR-3 consortium supplied a training corpus and a test corpus. The first was delivered 3 months before the second. The training corpuses were extracted randomly from the BTEC corpus. They consisted of 40,000 turn pairs for CE and JE, and 20,000 for AE and IE. They were encoded into UTF-8. Test corpuses contained 500 turns sent for translation.

The main goal of the campaign was to shift to the evaluation of the effects of spontaneity on the speech dialogs. Systems had to translate a variety of inputs, ranging from audio content collected from spontaneous dialogs to read BTEC utterances to ASR transcriptions (NBEST, 1BEST and word lattice) to preprocessed and unprocessed (w.r.t. segmentation, punctuation and use of case) written BTEC turns.

2. Evaluation

2.1. Objective (n-gram based) evaluation

Remarks about the quality of the source references

We translated all the training turns for CE and IE (40,000 and 20,000 respectively) by SYSTRAN and then used one of its commands to automatically add the NFW[5] to a new user dictionary. We were surprised by some strange expressions detected in some "source" Japanese and Italian turns (see Table 1 and Table 2). As a matter of fact, a high proportion (perhaps all in the case of Italian) has been produced by a human translation process from English or Chinese.

JE reference translations

For example, Table 1 shows two wrong reference turns: in the first, the word **高校** should be **後方** ; in the second，**帰る** has to be changed to **使える**.

Reference	ATLAS translation	Correct
トイレ は 機内 **高校** です ご 案内 致し ます.	It will be a guide of the rest room that an in-flight high school has (*O).	**後方**
はい クレジット カード を ご 利用 頂け ます し **帰る** カード は ビザ マスター アメリカン エクスプレス です.	The yes credit card can be had to be used and the card where (*S) returns is visa	**使える**

Table 1: Wrong kana-kanji conversion in source turns

We found many wrong reference source turns, apparently more than in the target English turns, which are for the BTEC part mostly original English turns found in travel books for Japanese tourists, and for the rest read-speach recordings of human translations of Chinese spontaneous utterances.

From a methodological point of view, that is really problematic, because translations are not reversible. Indeed, there is an "expansion factor" Exp12, when translating from L1 to L2, and another, say Exp21, when translating from L2 to L1, and they are always larger than 1 (typically 1.1—1.15), even is one language is supposedly more "terse"[6].

Hence, language pairs are directional, not reversible, and to develop an L2-L1 system on the basis of L1-L2 translations cannot lead to "good" MT systems. From the point of view of the evaluation, that is bound to unduly decrease the "objective" scores of MT systems not trained on that kind of data, another bias against commercial MT systems.

Subjective evaluation of adequacy is also biased, because the "golden translation" is actually a source utterance, the only one with a chance to be spontaneous. In particular, for the CE and JE pairs, the choice of articles ("a", "the", or none) and of number might be quite free if the real original turns were in Japanese or Chinese, which, having no articles and no obligatory mark for plural, are underspecified with respect to determination and number. To alleviate this problem, evaluation of adequacy should be performed only by bilinguals having access to the source utterance[7].

IE reference translations

IE_TRAIN_12108\Sì, abbiamo la Where, and The City Guide.	IE_TRAIN_12108\Yes, we have the Where, and The City Guide.
IE_TRAIN_01045\Congratualzioni, Henry. Sono felice di sentire del Suo fidanzamento con Jane.	IE_TRAIN_01045\Congratulations, Henry. I'm delighted to hear of your engagement to Jane.
IE_TRAIN_01049\Deve essere stato un grande shock per Lei.	IE_TRAIN_01049\It must have been a great shock to you.
IE_TRAIN_01726\Potrebbe pagare alla reception, prego?	IE_TRAIN_01726\Could you pay at the front desk, please?
IE_TRAIN_02516\Sono contento di averLa conosciuta. Grazie.	IE_TRAIN_02516\I'm glad I met you. Thank you.
IE_TRAIN_06501\Qui parla l'operatore dell' International Telephone Call Service.	IE_TRAIN_06501\This is the operator for International Telephone Call Service.
IE_TRAIN_09747\Facendo lo spelling è G-O-R-O-H.	IE_TRAIN_09747\It's spelled G-O-R-O-H.

Table 2: examples of wrong "source" Italian references

Conditions of the "objective" evaluations

For runs submitted to the official participation, automatic evaluation is carried out in a case-sensitive way, with punctuation. An additional evaluation is also carried out without punctuation, all MT outputs being preprocessed for tokenizing and de-punctuation before evaluation.

a) Punctuation and case reconstruction

The results produced by the ASR engine did not contain any punctuation. Their translations by all MT system also had no punctuation and no uppercase. The SRI Language Modeling Toolkit (SRILM) was used to extract a language model (LM) from the training data that was then used to reconstruct the punctuation and case of the English MT outputs.

Results

For the official translation, we translated the CE pair using SYTRAN tuned with new parameters and our user dictionary. We first sent the *ASR spontaneous speech* and *CRR* (Correct

[5] Not Found Words

[6] The regretted Hans Karlgren made extensive experiments on that phenomenon as he led various large translations tasks, the last being the translation of EU laws into Swedish in the early 90's.

[7] Fortunately, that was the case in IWSLT-06.

Recognition Result), and then we evaluated the *read speech* (ASR output) with the ATR web server.

a) The SYSTRAN CE runs

a.i Objective evaluation (SYSTRAN)

	official (with case + punctuation)				
	BLEU4	**NIST**	**METEOR**	**WER**	**PER**
Spontaneous speech	0.0344	2.7374	0.3178	0.87129	0.743063

	additional (without case + punctuation)				
	BLEU4	**NIST**	**METEOR**	**WER**	**PER**
Spontaneous speech	0.0406	2.8625	0.3184	0.880529	0.720287

	official (with case + punctuation)				
	BLEU4	**NIST**	**METEOR**	**WER**	**PER**
Read Speech	0.0536	3.7390	0.3210	0.805919	0.687017

	official (with case + punctuation)				
	BLEU4	**NIST**	**METEOR**	**WER**	**PER**
CRR	0.0366	2.685	0.3178	0.858339	0.726484

	additional (without case + punctuation)				
	BLEU4	**NIST**	**METEOR**	**WER**	**PER**
CRR	0.0749	4.4256	0.3694	0.780118	0.643764

Table 3: objective evaluation of CE on ASR output and CRR

a.ii Problems with Chinese segmentation (SYSTRAN)

对历史 (history) 感兴趣 (to be interested)	be interested in history
职员 (employer) 会 (can)轮流放假	employee can take several days off by turns
艾凡斯顿	Evanston
我就要替你喝完秋葵芙汤 (soupe)了。	gumbo
雕塑 (sculpture) 感兴趣 (be interested)	interested in sculpture
孟斐斯 (proper name)	Memphis
理察德 (Richard) 波尔曼	Richard Paulman

Table 4: Chinese segmentation errors

b) Additional translation runs

(i) Read speech: J-E translation by SYSTRAN

		BLEU	**NIST**
SYSTRAN	**ASR output (Read speech)**	0.0755	3.7685

(ii) Read speech: J-E translation by ATLAS

		BLEU	**NIST**
ATLAS	**ASR output (Read speech)**	0.1084	4.4295

(iii) Read speech: A-E translation by SYSTRAN

		BLEU	**NIST**
SYSTRAN	**ASR output (Read speech)**	0.049	3.6202

(iv) Read speech: I-E translation by SYSTRAN

		BLEU	**NIST**
SYSTRAN	**ASR output (Read speech)**	0.1368	5.1528

Table 5: Objective evaluation — additional runs

2.2. Subjective evaluation (fluency and adequacy)

Subjective evaluation was conducted on CE only, using NIST protocol (http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess02.pdf).

Both *fluency* and *adequacy* had to be evaluated 3 three judges for each result. Our judges were native speakers of English for fluency, and native speakers or specialists of Chinese for adequacy, because we planned to let our adequacy evaluators first grade for adequacy, then postedit (with automatic timing), with the polished output added to the set of references, and then recomputed all measures with that new set. But lack of human resources prevented us to do it in time for IWSLT-06.

Fluency

Fluency refers to the degree to which a translation conforms to the rules of Standard Written English. A fluent segment is one that is well-formed grammatically, contains correct spelling, adheres to common use of terms, titles and names, is intuitively acceptable, and can be sensibly interpreted by a native speaker of English. A *fluency* judgment is one of the following: 1: Incomprehensible, 2: Disfluent English, 3:Non-native English, 4: Good English, 5: Flawless English.

Judges are instructed to grade between 1 and 3 when translations retain source language characters or words, depending upon the degree to which the untranslated characters, among other factors, affect the fluency of the translation.

Adequacy

Here, the judge is presented with a reference translation and/or the original turn, and its translation by all systems. Judges are instructed to give a score between "1: None" and "4: Most" when English translations retain Chinese characters from the original turns, depending upon the degree to which the untranslated characters, among other factors, affect the *adequacy* of the translation.

Adequacy refers to the degree to which information present in the original is also communicated in the translation. Thus for *adequacy* judgments, if judges don't know the source language, a reference translation can serve as a proxy. The question asked is: *"How much of the meaning expressed in the gold-standard translation is also expressed in the target translation?"* and the answer is one of the following:

5: All, 4: Most, 3: Much, 2: Little, 1: None.

Grading CE for adequacy

From the point of view of the human evaluator, the evaluation decision must be based on how much a real person could understand the original meaning of the Chinese speaker through the output English translation. During this evaluation process, we hence tried not to pay too much attention to the form of the outputs: although the output turn may not be very correct, if people could still understand the meaning, we still give it a relatively high grade.

As for the evaluation interface, we warned at the preparatory meeting against presenting translations of the same input together, arguing that this would (mis)lead evaluators to waste a lot of time trying to rank them. The argument was that, if the examination of 1 output costs some unit u, if a comparison costs v, if a typical set (screen) contains 20 outputs, then about $20 \log 2\ 20 \approx 100$ comparisons are needed to rank them and the total time T to grade one screen can rise from $20\,u$ to $20\,u + 100\,v$, with probably $1.5\,u \leq v \leq 2u$, hence $170\,u \leq T \leq 220\,u$, an 8- to 11-fold increase.

Our experience confirmed these initial fears. It typically takes about 3 minutes to grade a screen without ranking, hence $u \approx 9$ *sec.*, and an average comparison takes about $v = 20$ *sec.,* so that the increase should be 12-fold if evaluators

really tried to establish a full ranking. In practice, trying to grade in this way takes anywhere from 20 to 40 minutes, because one never makes all comparisons.

We then told our judges to *never make comparisons between the different output English turns*. Following that simple rule, our main evaluator for CE graded 5,400 turns in about one day and half (270 screens, 13.5 hours, u=3 sec.), while the organizers' estimation (based on IWSLT-05 figures, where the interface was the same) was 4 to 5 days.

2.3. Remarks on types of errors and their sources

Systran JE

All Japanese source turns seem to be polished transcriptions of oral dialogues in the tourism domain. The language level is rather polite. Here are the main problems in the outputs.

- When the utterance is euphemistic (が), the particle is always translated by "but", which is quite wrong.
- Some of the utterances do not make sense without context (e.g. 切りますよ。➟ "it cuts" ?).
- When the first person subject is omitted in Japanese, it is always translated as "it" (ここで降ります。➟ "It gets off here.").
- The test set contains many interrogative utterances. In the translations, the interrogative pronoun or adverb is always (incorrectly) shifted at the end of the translation (e.g. オペラ座はどこですか。➟ "Is the opera house where?").
- A lot of spoken Japanese daily life idiomatic expressions are not contained in the SYSTRAN dictionaries (e.g. どういたしまして。➟ "How doing." もしもし。➟ "It does." さようなら。➟ "Way if.").
- Requests or invitations are not always well translated (e.g. 注文したいのです。➟ "It is to like to order." 一緒に行きましょう。➟ "It will go together.").
- When the valency of the verb for two expressions in Japanese and English is different, the translation is almost always wrong (e.g. 寒気がする。➟ "Chill does.").
- The aspect of Japanese predicates is not correctly rendered in English (e.g. 航空券を家に忘れて しまいました。➟ "The air ticket was forgotten in the house.").

On the other hand, a positive point is that lexicalized Japanese politeness is correctly handled (e.g. そのまま切らずにお待ち下さい。➟ "Without cutting that way, please wait.").

Atlas JE

We estimate that the ATLAS system produced 35% correct translation at the grammatical, syntactical and semantical levels. Wrong translations are due to (1) segmentation errors, (2) lack of resources to handle phenomena specific of spoken language, and, surprisingly enough, to (3) the large proportion of quite bad source texts, which cannot be understood even by human native speakers (37%). Here are some more details about errors observed and their causes.

Table 6 shows how wrong segmentations lead to quite bad translations.

申し訳ありません 離陸 して から でないと テレビ を 御 使い 頂け ません.	The television cannot be had to be used after the take off which apologizes and not is.	The turn is composed of 3 turns, but ATLAS has translated it as two turns with a relative clause"
これ は 無鉛 で は あり ません ね が ご 希望 なら 御 取り替え 致し ます.	If sleep which is not no lead is hope, I will change this.	The turn is composed of 3 turns "これ は 無鉛 で は あり ません ね", "ご 希望 なら" and "御 取り替え 致し ます", but ATLAS has translated it as two turns with a relative clause, because the sentence final particle "ね" is not recognized.

Table 6: Segmentation errors (ATLAS JE)

Table 7 shows how some characteristics of spoken language, not handled by ATLAS, diminish translation quality.

申し訳ありません 離陸 し て から **でないと** テレビ を 御 使い 頂け ません.	The television cannot be had to be used after the take off which apologizes and not is.	Verb "でないと"
やっ てみ ます が から ぞ 予約 できる か 保証 し 兼ね ます.	Whether からぞ can be reserved cannot be guaranteed やってみます.	Verb "やる"
えーっと それ は 六 百 円 です.	Food っとそれ is 600 yen.	Phatic "えーっと"
以前 は 野球 を する の が 好き でし た でも 今 は スキー の 方 が 興味 が あり ます	It was liked to play baseball and skiing is interesting yet now before.	Conjunction "でも"
切っ て 今 手 が ございます どうぞ ご覧 下さい.	(*S) cuts (*O), and there is a hand now and (*S) sees please.	Polite expression
結構 です けど ね でき ます.	(*S) sleeps though it is excellent.	Modal particle "ね"
ドイツ語 の が ある と 一 番 良い の です が 英語 は 読め ない の です.	English cannot be read as German がある though it is the best.	Referential noun "の"
はい 洗濯機 の 着席 優しく 払わ なければなりません ので ご 注意 下さい.	Please <払 わなければなり> note (*O) <sit-down> nice of the tile washing machine.	Modal expression "なければなりません"
通常 一 週間 です でも 天気 が 悪い わ えー 少し 遅れる こと も あり ます.	The weather for one usual week it yet might be late of <badness> いわえ least	Phatic "えー"
かしこまりました 少々 御 待ち 下さい.	Please wait a little standing on ceremony.	Polite expression "かしこまりました" and Honorific expression "御"
陶器御茶の方御酒を買いましたこれらは全てねでーです.	These by which person 御酒 of earthen 御茶 is bought are all sleeps.	Honorific expression "御"
そうですね あと 一 時間 位 で 着陸 し ます.	(*S) <aspect> has, (*S) sleeps, and (*S) will land in about another hour.	''ね'' in dialogues

御 客 様 こちら です 口頭 その ビル の 男性 の 角 に ございます.	It is in the corner of the man in guest こちらです oral その building	Deictic expression ‘’こちら’’

Table 7: spoken language phenomena

Table 8 shows that the dictionary, even very large, is not large enough if the system is applied to an unforeseen type of language (or sublanguage).

赤青緑黄色がございますどの色が御好みですか.	Which color with 赤青緑 yellow is favor?	Special words ”赤,青,緑”
いいえ その ドア を 出 て から 右 に 曲がらなければなりません.	It is necessary to turn right after (*S) goes out of the door of いいえそ.	Deictic and anaphoric word Mots déictique et anaphorique ”その”
こんにちは 御 客 様 の フライトナンバー と 宿泊 を 取る 名前 を 書い て下さい	The name by which the flight-number and staying of 御客様 hello are taken	Honorific word “御 客 様”
ラジオ の 電源 スイッチ は 一人 が ですし の つまみ は 音量 を 調節 する 為 の 物 です.	つまみ of <one person> ですし is a thing to adjust the volume. the power supply switch of the radio	Special word “つまみ”
御 客 様 もう しばらく 御 待ち 下さい 一 週間 以内 に は 御 返事 差し上げ ます.	Guest もうしばらく is waited and I present the answer within one week.	Special word “もうしばらく”
. 御 客 様 こちら です 口頭 その ビル の 男性 の 角 に ございます.	It is in the corner of the man in guest こちらです oral その building	“御 客 様” Deictic word “こちら”
あちらの大きな連中は記念ように保存されています.	A big party there is preserved in the commemoration way.	Deictic word “あちら”
いいえ まだ です.	いいえまだです.	Special word “いいえ”
一 番 近く の レストラン は 車 で もう 三十分 近く 掛かり ます.	The nearby restaurant hangs in the vicinity for 30 another minutes in the car.	Semantic ambiguity of verb “掛かる”

Table 8: problems coming from the dictionary

Table 9_shows examples of bad translations caused by errors in the orthographic transcription. In some cases, even a Japanese native speaker cannot guess what it could possibly mean.

精神は三名ドルほどです.	The soul is about three person dollar.	?
私の国は中国のりんご君日本です.	My country is apple 君日本 of China.	?
離陸 を 三十分 以内 に は 昼食 を 御 出し 致し ます.	The take off is served and I will serve lunch within 30 minutes.	離陸 を → 離陸 後
トイレ は 機内 高校 です ご 案内 致し ます.	It will be a guide of the rest room that an in-flight high school has (*O).	高校 → 後方
はい クレジットカード を ご 利用 頂け ます し 帰る カード は ビザ マスター アメリカンエクスプレス です.	The yes credit card can be had to be used and the card where (*S) returns is visa	帰る → 使える
はい 車 で 十分 ほど と 頃 に 一つ ございます.	It is a tile car and there is one every about ten minutes.	と 頃 に → のところに
こちら です 化粧 品 は 二 階 です え で データ で 上がっ て下さい.	Cosmetics which have (*O) <here> must rise by data in placing by the second floor.	えで データ → エレベータ
やっ てみ ます が から ぞ 予約 できる か 保証 し 兼ね ます.	Whether からぞ can be reserved cannot be guaranteed やってみます.	から ぞ → 必ず
申し訳ありません 今 の 所 に を 五 チャンネル は ございません.	There are no place にを five channels now since (*S) apologizes and (*S) does not exist.	にを → には

Table 9: problems in the input Japanese text

Table 10 shows two interesting characteristics of ATLAS:

4. when the subject or object is absent in Japanese, ATLAS generates placeholders for them (instead of awkward and often misleading pronouns or pronoun lists such as he/she/it or him/her/it).
5. it can produce an output showing the English equivalents (in context) of Japanese words or expressions inserted after them.

まっすぐ行って下さい一度物理木と調和サービスデスクの隣にあります.	(*S) goes straight and (*S) exists once in the vicinity in a physical tree and the harmony service desk.	A subject missing in Japanse is indicated by (*S), an object by (*O).
この道をまっすぐ行ってご指定の近くですそこ迄行くには徒歩で五分位です.	If sleep which is not no lead is hope, I will change this.	この(this)道(road)をまっすぐ(straight)行っ(go)てご指定(specification)の近く(near)ですそこ迄行(go)くには徒歩(on foot)で五分位です.

Table 10: placeholders in output for missing subjects and objects & "bilingual" output

Finally, Table 11 shows examples of phenomena typical of spoken language and not handled by ATLAS, which has been mainly developed to handle written texts in technical domains.

結構 です けど **ね** でき ます.	(*S) sleeps though it is excellent.	Back channel particle “ね” is not recognized, but is interpreted as the verb “寝る”.
ドイツ語 **の が ある** と 一 番 良い の です が 英語 は 読め ない の です.	English cannot be read as German **が ある** though it is the best.	Anaphoric pronoun “の” is not recognized.
はい 洗濯機 の 着席 優しく 払わ **なければなりません** ので ご 注意 下さい.	Please **<払 わなければなり>** note (*O) <sit-down> nice of the tile washing machine.	Modal expression “なければなりません” is not recognized.
通常 一 週間 です でも 天気 が 悪い わ **えー** 少し 遅れる こと も あり ます.	The weather for one usual week it yet might be late of <badness> **いわえ** least	Phatic “えー” is not recognized.
かしこまりました 少々 御 待ち 下さい.	Please wait a little standing on ceremony.	However, politeness expression “かしこまりました” and honorific particle ”御” are recognized.

Table 11: Most Japanese spoken language characteristics are not processed by ATLAS

3. Towards better and cheaper measures

3.1. Towards task-related objective measures

Problems with n-gram based measures

Contrary to "mainstream" thinking, the current "objective measures" based on n-grams don't measure translation quality. It has been proven by experiments and by theoretical analysis. See for example [6]. They measure some kind of similarity with the n-grams in the reference translations, and tend to diverge more from human judgment when translation quality (as judged by humans) grows — to the point of putting a new, perfect human translation last or next to last while human evaluators would put it first.

Another problem is the cost of preparing the reference translations. Building 4 reference translations for a training corpus of 40,000 turns of 6.5 words in average (equivalent to 1,000 standard pages) requires 4,000 hours without machine help, at least 2,000 if there is an adequate translation memory, and at least 1,000 if there is a "good" MT system — we used SYSTRAN EF in this way and translated 4,000 turns in about 24 hours (postedition in "translator setting").

It is true that these reference translations can be used again and again with no cost on successive versions of systems, but (1) objective measures tend to correlate less and less with human judgments of quality when the (task-related) quality increases, and (2), new sets of reference translations have to be built each time a new corpus is tackled. That encourages developers to stick with the same corpuses. Their systems may get better grades, and even perform well on these corpuses, but the ultimate goal of MT is missed: who needs to translate and retranslate the same corpus? It is necessary to evaluate on new types of texts, and on new language pairs.

In the case of IWSLT-06, where 16 reference translations were attached to each of the "source" turns in the development sets (and 7 in the test sets), we suspect that the same set of English turns has been selected, so that this large number (16) has been obtained by adding the 4 reference translations built for each of the 4 language pairs[8]. That would have been impossible if translating from English, and not into English. Even so, the human time necessary to build them can be estimated between 12,000 and 16,000 hours (6 to 8 man-year).

[8] According to one reviewer, 3 of 4 references were produced by English native speakers paraphrasing the original English turn. Cost estimates don't change, but references are even less "true".

Objective measures involving humans can be cheap

We have long advocated the use of *objective* human- and task-related measures. The idea is to go from expensive and inadequate measures such as BLEU, NIST, etc. to low-cost measures, inherently better because task-related.

a) HQ translation

If the task is to produce high-quality translations, a first possibility is to measure the time spent on "postediting" (in "translator's mode", that is, reading first the source text). Translators can easily do it by entering beginning and end times in an Excel sheet, or the tool used (Trados, Déjà Vu, Transit, Similis, etc.) can be equipped to do it.

A second possibility, used by TAUM-METEO (since the early 80's), is to measure the number of actions of each type (insertion, deletion, local replacement, global replacement), to assign a coefficient to each type, and to derive a cost. For METEO, the measure used was simply 100 minus the number of insertions and deletions done to postedit 100 words of output. This quality measure went from about 55% at the beginning (≈1978) to 97% from around 1988.

Comprendium also uses it in assessing the quality of its Spanish-Catalan and Spanish-Galician systems which translate newspapers every day (1 hour/page with no machine help, 30 minutes with translation memory, 5 minutes using MT, a 12-fold increase of productivity).

Another type of measure is based on computing a distance between the raw translation and its postedition. An interesting problem here is to "reconstruct" a sequence of operations of minimal cost. That is almost trivial if global changes are not considered. But if global changes (on a document or set of documents[9]) are considered, this becomes a hard problem.

What about the cost? It is really minimal, because the human work is necessary in any case to perform the desired task, and the "instrumentation" of the translation support tool is done once and does not require any special equipment.

b) Pure understanding

The best way to measure content understanding is perhaps to perform TOEFL-like comprehension tests. Most of them are multiple choice questions. We would then trade reference translations against reference questions!

[9] For example, Systran EF on BTEC translated "please" as "s.v.p.", which is OK in written texts but not in transcriptions of spoken utterances. Changing all occurrences of "s.v.p." (Thousands) by "s'il vous plait" in 168,000 turns (168 files) took only a few seconds. Hence it should be assigned a few cost points, not thousands.

What about the cost? According to R. Mitkov, recent research on computer-aided generation of multiple choice questions to test comprehension (MCQC) has been quite encouraging. A MCQC can be built interactively in about 3 minutes, and about 10 MCQC are enough to test the comprehension of a page of 250 words (40 BTEC turns).

For 1 page, we would go down from 2-4 hours spent on preparing 4 reference translations to 30 minutes spent on assisting an interactive MCQC generation system. That is a reduction by 4 to 8. But we must add some time during the evaluation, because answering the MCQC automatically is still a research problem. However, a human can probably read a page and answer the MCQC relative to it in less than 5 minutes. If 3 "judges" are used as in IWSLT-06, each evaluation would cost 15 minutes of human time. That is quite less than what is spent now on subjective evaluation.

c) Understanding to act

In the case of e-commerce applications, the situation is similar to the production of HQ translations, in that the task is clear: induce buying acts. For a marginal cost, it should be possible to measure some rates of actions based on precise comprehension (e.g., compare the number of buying acts for 100 visits to a web page accessed in its original language, and accessed in an automatically translated version).

3.2. Towards eliminating subjective measures

Evaluation measures are too often called "subjective" because they involve humans. As stressed above, that is inexact: it is quite possible to use humans in cheap "objective" measurements, and it is an old practice in MT. But, to do it, systems must be put to operational use.

There is however no dispute that the current "subjective" measures are indeed subjective and costly. As we have seen, their cost can be drastically diminished if interfaces for judges are built such that they *never present several translations of the same input together*.

Finally, why not eliminate these fluency and "adequacy" measures altogether? First, adequacy would be far better measured by the task-related measures above, which depend on *what* translations are supposed to be adequate for. Second, fluency is often a component of adequacy (depending on the task and on the usage situation). A suggestion, then, would be to suppress the classical fluency and adequacy measures as we know them since 4-5 years or so, and to invent and introduce usage-related measures for deployed systems.

Conclusion

We reported on our participation to IWSLT-06. First, we let the SYSTRAN and ATLAS systems participate as fairly as possible (SYSTRAN v5.0 for CE, JE, AE, & IE, Atlas for JE), taking care of preprocessing and postprocessing tasks, and tuning the MT systems as much as possible by creating "user dictionaries" (for SYSTRAN CE and IE) and finding a good combination of parameters (such as dictionary priority).

Second, we took part in the subjective evaluation of CE results (fluency and adequacy). Details on experiments and methodological remarks have been provided, with a perspective to introduce less expensive and more objective human- and task-related evaluation methods.

Acknowledgements

We would like to warmly thank the participants to the subjective evaluation, Wei Weng, Etienne Blanc, Emmanuelle Esperança-Rodier, and John Kenright, as well as our partners from ATR, especially Michael Paul, and our reviewers, for pertinent comments. Thanks also to SYSTRAN SA for letting us use their systems for these experiments.

References

[1] Blanchon H., Boitet C. & Besacier L. (2004) *Evaluation of Spoken Dialogue Translation Systems: Trends, Results, Problems and Proposals.* Proc. COLING-04, Genève, 23-27/8/04, ACL, 7 p.

[2] Blanchon H., Boitet C. & Besacier L. (2004) *Spoken Dialogue Translation Systems Evaluation: Results, New Trends, Problems and Proposals.* Proc. IWSLT-04, Nara, 30/9-1/10, ATR, pp. 95-102.

[3] Blanchon H., Boitet C., Brunet-Manquat F., Tomokyo M., Hamon A., Hung V. T. & al. (2004) *Towards Fairer Evaluations of Commercial MT Systems on Basic Travel Expressions Corpora.* Proc. IWSLT-04, Kyoto, Japan, 30/9-1/10, ATR, pp. 21-26, 6 p.

[4] Boitet C. (2001) *Machine Translation.* In "Encyclopedia of Cognitive Science", A. Ralston, E. Reilly & D. Hemmendinger, ed., Nature Publ. Group, London, 10 p.

[5] Boitet C. (2002) *A rationale for using UNL as an interlingua and more in various domains.* Proc. LREC-02 First International Workshop on UNL, other Interlinguas, and their Applications, Las Palmas, 26-31/5/2002, ELRA/ELDA, J. Cardeñosa ed., pp. 23—26.

[6] Callison-Burch C., Osborne M. & Koehn P. (2006) *Re-evaluating the Role of BLEU in Machine Translation Research.* Proc. EACL-06, Trento, April 3-7, 2006, ITC/irst ed., 8 p.

[7] Leusch G., Ueffing, N., et al (2003) *A Novel String-to-String Distance Measure with Application to Machine Translation Evaluation.* Proc. MT-Summit X, New Orleans, USA, 23-27/9/03, pp. 8.

[8] Levenshtein V. I. (1966) *Binary codes capable of correcting deletion, insertions and reversals.* Soviet Physics Doklady 8/10, pp. 707-710.

[9] Pinkham J. & Smets M. (2002) *Traduction automatique ancrée dans l'analyse linguistique.* Proc. TALN'02, Nancy, France, 24-27 juin 2002, vol. 1/2, pp. 287-296.

[10] Sénellart J., Boitet C. & Romary L. (2003) *XML Machine Translation.* Proc. MTS-IX (Machine Translation Summit), New-Orleans, 23-28/9/03, 9 p.

[11] Siegel S. & Castellan N. J. (1988) *Nonparametric Statistics for the Behavioural Sciences; 2nd ed.* McGraw-Hill, New-York, pp. 400.

[12] Takezawa T., Sumita E., Sugaya F., Yamamoto H. & Yamamoto S. (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.

[13] Tomás J. & Mas J. À., et al. (2003) *A Quantitative Method for Machine Translation Evaluation.* Proc. EACL-03, Budapest, 14/4/03, vol. 1/1, pp. 8.

[14] Turian J. P. & Shen L., et al. (2003) *Evaluation of Machine Translation and its Evaluation.* Proc. MT-Summit IX, New Orleans, USA, 23-27/9/03, pp. 386-393.