[*MT evaluation: basis for future directions*. Proceedings of a workshop...2-3 November 1992, San Diego]

Results of the Warmup Exercises

Marjorie León
Pan American Health Organization

A total of 25 participants turned in the Warmup Exercises. Of these, 12 people evaluated the French-to-English translations, 3 people evaluated the Japanese-to-English, and 10 people evaluated the Spanish-to-English.

The exercises involved two of the texts that were used by DARPA, one on Borland acquiring Ashton-Tate, and one on an airline reservation merger. One text was used for the quality evaluation and the other for the comprehension test. Eight individual sentences were used for the intelligibility and fidelity evaluation. Some packets had machine-alone translation, some had human-assisted machine, and, in the case of Japanese, some had human translations because we only had one version of human-assisted machine translation.

Many participants found the quality evaluation to be very difficult and time-consuming. Eight people did not do the quality tally at all. There were only three translation versions that were evaluated by more than one respondent. The general opinion was that the categories did not always lend themselves to what needed to be said about the translation.

Everyone did the comprehension test, which was the easiest to do. Almost all those who took the test on the airline reservation merger scored 100%, regardless of which translation they were evaluating. On the other hand, some people got every answer wrong on the text about Borland and Ashton-Tate. The main problem seemed to be that there were errors in the translation from the original English.

The sentences for the intelligibility and fidelity test included, in addition to the machine-alone and the human-assisted machine translation, some human translated examples also. The first sentence was taken from each of eight texts, except for a mistake in the last sentence on the Spanish test.

The participants were more satisfied with the rating scales for intelligibility and fidelity than with the quality criteria. They felt that these scales allowed them to quantify their opinion about the sentences. Some people did get confused because the best score was the highest number for intelligibility and the lowest number for fidelity, but that was how ALPAC did it. In fact, one person got so confused that the scores could not be included in the final results.

Even though the participants liked the intelligibility and fidelity evaluation, there was a wide range of scores for many of the sentences. It would seem that these scales still leave a lot of room for subjectivity. Some of the variation may also be due to different backgrounds of the evaluators. Several participants mentioned that a specialist in the subject area of mergers and acquisitions might react quite differently to some of the translations.

Conclusions

The evaluation exercise provided an opportunity for the participants to get some hands-on experience with several evaluation metrics. The real purpose of the exercise was not to evaluate the translations, but to elicit opinions about the evaluation methodologies. Those who completed the exercises are now more aware of the large amount of effort that is involved in a fair and comprehensive evaluation. They also pointed out some problems that were caused by the use of translated material rather than texts that were originally written in the source language.

Summary of General Comments

French

Scoring sentence by sentence (as in both DARPA and ALPAC methods} is limiting. Sometimes a translation has to be rearranged and recomposed.

A sample of each type of exercise result would have helped those of us who were pressed for time and/or novices at the techniques employed.

Don't put the texts to be compared on the front and back of the same page, forcing the user to keep flipping the sheet. (I hope that that's true only of this Warm-up Exercise, not the evaluation itself.)

It is interesting to look at a language I don't usually translate (French) and see where the difficulties with MT come up. They are at very different places than in German translation.

Japanese

This is a good start.

The background reading material and exercises were very useful.

Japanese to English translation test sentences are limited to the are that is literally translatable. I think, more Japanese original expression should be tested to develop the method of MT.

Good use of cross-comparative evaluation. The examples demonstrate the basic problems we should be aware of and show that much thought and preparation went into the creation of the workshop. However, the scales of measurement should be kept roughly the same from one type of evaluation to the next.

Spanish

The first two approaches presented major problems. Surprisingly, the ALPAC method was the best one for capturing how "good" the translation is—how effectively it achieve its purpose. None of them tell very much about what an MT system is capable of producing.

Translation evaluation tests should not be based on back translations, especially when these are not as carefully done as they should have been. It should always be clear whether we are evaluating a translation or a translation system. Criteria generally uninformative.

Comments on the Quality Evaluation Exercise

French

Perhaps you could use a few more parameters. What is there doesn't seem quite adequate for all situations.

The subcategories of syntactic versus lexical errors are not comprehensive enough.

Instructions hard to follow, in that certain types of error were not easily classified. It doesn't seem right for all lexical errors to count 2 points.

French—Quality Evaluation (cont.)

Some errors did not match the categories given, e.g. sentence 11 is total garbling of nested collocations. More serious, a single mistranslated word (lexical, 2 points) could reverse the meaning of the sentence, which is a worse problem than the misalignment of a phrase (syntactic, 4 points). The middle of sentence 11 is absolute nonsense but countable only as 1 or 2 lexical or style errors.

Seems to be a good translation, but my commercial French isn't what it should be. Output reads well, but it does interpret the original a bit.

As a (former) translator, I have plenty of experience teasing a meaning out of a string of words, so "finding the meaning" is too easy for me. A subject specialist starting cold will come with very different results, I think. I probably saw this exercise as a repair job (can it be fixed, or should we start again from scratch?).

Only one sentence was entirely correct; syntactic errors tended to compound into extreme failure to render the meaning. Some punctuation/orthography mistakes were not explicable (lack of spaces after a period, or "CP" for "PC"). Fair rating takes lots of time and precision.

Japanese

The criteria are difficult to follow at first. Suggestion: include bullet items on the evaluation sheet. Note: There is an error in sentence 3 of the Japanese text.

I cannot mark these translations because of my lack of ability. However I can say that all of these translations can be marked successful (above 6 points in full 10 points) from our standard. I would like to point out that the original Japanese sentences seem to be translation-oriented Japanese and very suitable to literal translation. Any way, our MT system cannot produce these results by a blind test. After preparing some words into dictionary and brushing up minor rules, I think more than half of these sentences may reach this quality.

This must be a human translation. Too creative for a machine.

I could not mark them because of lack of my English ability. All of these results seems to me successful by our J to E evaluation standard. The test sentences seem not to be original Japanese but to have been rewritten into literally translatable Japanese.

Spanish

The categories, which appear to have been devised to test human translation, are seriously inadequate for capturing the kinds of mistakes that MT makes. Most of the error in the sample defied classification under this scheme. They also fail to reflect the interpretive errors that all translators make, which are the main problem.

Very uninformative. Mixing of criteria and strict point assignment means that a lexical error may be very serious and change the whole meaning, but only gets 2 points, whereas a syntactic error may not modify the essential meaning and still gets 4 points. Difficult to apply criteria consistently. Syntactic criteria artificial in general.

No problems with this exercise except that columns one and two on the Evaluation Tally Sheet are reversed and that slowed me down. I didn't get any sentences with the 12 point cap so it is possible that my evaluation was not critical enough.

Comments on the Comprehension Test

French

Not particularly useful when the translation is quite poor. Simply a guessing game.

F->E gave the needed info in comprehensible form. The J->E and S->E are much poorer, but could be deciphered using the info just gained from F->E.

Not bad, but just 4 questions seems shallow. Maybe it's hard to generate more?

This translation tends to go for complex English equivalents (amalgamate, envisage, etc.) for reasons that are not clear.

Easy to do and perhaps the best form of evaluation for this type of translation since reader comprehension is ultimately the goal here.

Japanese

Very good, but tricky.

May be a machine?

This test sentence set seems to have variety of expressions. The translation results are good for current level of MT. I wish our system could obtain this level of translation.

Spanish

The questions have almost nothing to do with identifying translation problems and demonstrating that they were solved. The last one was mined with tricks. It became a game to find the answers, with little attention being given to the text itself.

Generally useful. Asking for a summary of the main ideas in the text could also have some validity.

Comments on the Intelligibility and Fidelity Evaluation Exercise

French

A nice way of pairing the two factors: a well-written, highly intelligible translation may be quite unfaithful to the original. (I also didn't find it difficult to pick out the right number: the verbal descriptions were appropriate.)

The reader is strongly tempted to make I and F add up to 10. Totals should in fact probably lie in the 9-11 range.

I thought this was the best-designed exercise, with the best potential for revealing the important characteristics of a translation.

Extremely interesting use of dual scales and both documents. Together, they constitute a good measure of effectiveness. I liked doing this rating more than the specific content evaluation of the "Quality Evaluation Exercise."

French—Intelligibility and Fidelity (cont.)

Intelligibility was useful, fidelity easy only given native understandings. Also the rating scale was reversed so it was confusing to grade [NOTE: This individual was so confused that this rating sheet had to be eliminated from the general tally.]

Japanese

The grades of these translations seem the same to me. Expressions used in the test sentences are very limited to test the quality of translation, I think.

This is a good procedure for evaluation.

The expressions used in the test sentences are very limited. And results seemed as the same degree.

Spanish

The criteria were not as difficult to apply as I expected them to be. The levels became quite clear. It was sort of fun, but I wouldn't enjoy doing a lot of them. Some of the "originals" were clearly translations and this made the English even murkier.

Criteria too vague, but in general more indicative then "Quality Evaluation" criteria. It's artificial to evaluate sentences without a context. Fidelity exercise more indicative than Intelligibility exercise.

Results of the Quality Evaluation

Not enough responses were received to do any meaningful comparison of results. There were only three texts that were evaluated by more than one individual. The points deducted in each subcategory do not add up to the total, because of the application of the 12 point cap. Also, some totals were assigned without providing the breakdown by subcategory. The data are given below.

Text	Time (min)	Lexical	Syntactic	Stylistic	Punc./ Orth.	Total
French 1	33	56	8	18	1	83
French 1	50	46	20	4	1	79
Spanish 1	25					178
Spanish 1	90	54	28	17	10.5	88.5
Spanish 2	25	30	32	6	2	70
Spanish 2	25	30	32	6	2	70
Spanish 2	45	20	12	5	0	37

DATA
French Intelligibility and Fidelity

Sent. 1		Sent. 2		Sent. 3		Sent. 4		Sent. 5		Sent. 6		Sent. 7		Sent. 8	
I	F	I	F	I	F	I	F	I	F	I	F	I	F	I	F
6	3	8	1	5	2	5	6	9	1	4	6	3	7	9	1
8	6	8	5	6	2	8	3	8	5	7	3	7	3	9	2
7	4	8	1	6	2	6	3	8	1	5	5	4	7	9	1
5	6	8	2	8	3	6	3	8	2	8	3	5	6	9	1
6	8	8	3	6	3	5	6	9	4	4	6	3	8	9	1
6	8	8	1	7	1	3	9	8	1	6	8	3	6	9	1
4	6	9	1	7	1	6	5	9	5	6	6	3	8	9	1
4	7	6	3	5	8	6	2	9	1	4	9	2	8	9	1
7	2	8	0	6	3	8	1	8	1	5	2	4	6	8	1
6	6	8	2	6	2	7	3	8	6	4	8	3	9	9	1
7	3	8	2	7	2	6	5	7	2	7	4	6	6	9	1
66	59	87	21	69	29	66	46	91	29	60	60	43	74	98	12
6.0	5.4	7.9	1.9	6.3	2.6	6.0	4.2	8.3	2.6	5.5	5.5	3.9	6.7	8.9	1.1

DATA
Spanish Intelligibility and Fidelity

Sent. 1		Sent. 2		Sent. 3		Sent. 4		Sent. 5		Sent. 6		Sent. 7		Sent. 8	
I	F	I	F	I	F	I	F	I	F	I	F	I	F	I	F
9	1	8	6	7	1	6	4	9	6	8	2	7	2	9	9
9	1	8	6	7	4	6	2	8	4	8	3	8	2	9	9
8	3	8	4	6	6	9	2	9	4	8	2	8	2	8	9
8	1	8	3	6	2	6	4	7	6	7	1	6	1	9	8
9	1	8	9	5	2	4	1	9	1	8	1	6	0	9	9
9	1	8	4	7	2	6	2	8	1	6	2	7	1	9	9
8	2	7	4	6	4	6	6	8	9	6	3	6	8	9	9
60	10	55	36	44	21	43	21	58	31	51	14	48	16	62	62
8.6	1.4	7.9	5.1	6.3	3.0	6.1	3.0	8.3	4.4	7.3	2.0	6.9	2.3	8.9	8.9

SUMMARY
French Intelligibility and Fidelity
11 Responses

Sent. No.	Translation Mode	Intelligibility			Fidelity		
		Average	High	Low	Average	High	Low
1	MO	6.0	8	4	5.4	8	2
2	HT	7.9	9	6	1.9	5	0
3	MO	6.3	8	5	2.6	8	1
4	HA	6.0	8	3	4.2	9	1
5	HT	8.3	9	7	2.6	6	1
6	MO	5.5	8	4	5.5	9	2
7	MO	3.9	7	2	6.7	9	3
8	HA	8.9	9	8	1.1	2	1

MO = machine only HA = human assisted HT = human only

SUMMARY
Spanish Intelligibility and Fidelity
7 Responses

Sent. No.	Translation Mode	Intelligibility			Fidelity		
		Average	High	Low	Average	High	Low
1	HA	8.6	9	8	1.4	3	1
2	HT	7.9	8	7	5.1	9	3
3	MO	6.3	7	5	3.0	6	1
4	HT	6.1	9	4	3.0	6	1
5	HA	8.3	9	7	4.4	9	1
6	MO	7.3	8	6	2.0	3	1
7	MO	6.9	8	6	2.3	8	0
8	HA	8.9	9	8	XX	XX	XX

MO = machine only HA = human assisted HT = human only