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

PANEL: THE DARPA METHODOLOGY

Presentation: *John White*
Planning Research Corporation

At this point I am going to briefly review the underpinnings of what we wanted to evaluate in the DARPA machine translation evaluation, the method we developed to do that evaluation, how we actually conducted the test, and the results of the test itself.

As George Doddington pointed out in his introduction, the objective was to have machine translation ultimately come into its own alongside the other speech and natural language initiatives within DARPA—that is, to say, to have an evaluation methodology that could be used year after year not only to show improvement in particular approaches and particular systems, but also to show, insofar as possible, the relative merits of one approach over another.

It turns out that for machine translation it is particularly difficult to develop a methodology like that because the evaluation matrices tend to be extremely subjective for anybody's translation, human or machine or otherwise. The criterion seems to be "Is this okay or not?" without any really well-standardized way of pegging it down.

At the same time, the systems in the DARPA initiative were unlike in terms of the languages they translated: one translates from French into English, one from Spanish to English, and one from Japanese to English. They are unlike in the linguistic approach that they take as part of their core algorithm. And they are unlike in terms of their foreseen end use—that is to say, the kind of a systems they would be if they were fielded systems that people actually used. Thus we had three very different application types, three very different linguistic approaches, and three very different language pairs. So, no pressure at all to come up with a single methodology that evaluates all of them and allows for both internal and external comparison among them! It's fairly trivial to conclude that the only way you can do this is by using some sort of modification of a black-box approach. There's no point in looking inside the individual systems, given how diverse they are.

We figured that for the different application types there were probably different sorts of test methods that you could apply to them. We undertook a dry run test at the end of last year in which we evaluated one of the systems in terms of the comprehensibility of the output, as determined by monolingual speakers, and also in terms of quality of the translation, as determined by translators who were familiar with both the source and target language. Based on the results of that dry run, we modified—and in many instances simplified—the methodology to the one that we used this summer. It involved basically two tests along the lines I have just described: a comprehension test, wherein monolinguals determined the comprehensibility of a translation using a multiple-choice SAT-type test, and a quality panel test, in which texts were compared in the source and target languages by professional translators and representatives of the Government to determine, using a U.S. Government scale for grading translators, the acceptability, or the grade, of these translations, as if they were produced by human translators.

In order to do this, you need two different sources of data. For an evaluation of quality, you need texts that were written in each of the source languages—some texts written in Japanese, some in French, and some in Spanish—that are then translated by the systems or by the human controls into English. These are the ones that are compared by the quality panel. Then you have to have another set that monolinguals are going to see, which have to be translated by all three systems, since the three systems each translate different languages. These passages were chosen from the *Wall Street Journal* and then professionally translated into the source languages—Japanese, French, and Spanish—so that the systems and the controls could then translate them back out into English. So, mixed into the pot is the additional consideration that these sets were back-translations.

Each of the three projects (PANGLOSS, CANDIDE, and LINGSTAT) received a set of 18 passages—newspaper articles in the field of business mergers and acquisitions. You have already seen two of these passages in your packet (Annex I). The ultimate objective was to come up with: translations from the contractor sites (which could be processed either fully automatically, human-assisted, or both); output from a control MT system (either SYSTRAN or SPANAM or both); and a control human set (apart from the set we had had professionally translated from the master passages going from the source back to the target), as done by people who you might expect would be users of this system sometime down the road—that is to say, novice people who were familiar with the two languages but were not professional translators. That's where the phrase "Level 2" comes in.

Each site identified people who would do the human-assisted part of their particular MT system's operation and who would also do manual translation for the other half of the text they didn't do the MT on. Each person did exactly half the texts using MT output and half of them human-alone.

So here is what we got. The French-English system, CANDIDE, provided both an unedited output and a human-assisted output for the test passages. The human-assisted part was done as a mixture of some MT (three outputs) and some human-only of the 18 passages involved. SYSTRAN French was used for the MT control. SYSTRAN French is a fully developed product which is used operationally on a daily basis all over the world and is therefore quite suitable as a commercial benchmark for purposes of comparison.

For Japanese-English there was Dragon's human-assisted LINGSTAT and SYSTRAN's batch MT as provided by the Federal Broadcast Information Service. This is a pilot system, and it has not been trained for the domain of business mergers and acquisitions. It should therefore not be considered a state-of-the-art benchmark for control, but SYSTRAN kindly allowed us to use it through FBIS for purposes of this exercise. There was also one output from the human-only translator-operators working at the LINGSTAT site.

For Spanish there were three outputs: human-assisted output from PANGLOSS, fully automatic unedited MT from SYSTRAN, which is in a pilot stage, and fully automatic unedited SPANAM.

We did not time the fully automatic translations. The control systems—namely, SYSTRAN and SPANAM, agreed to a 48-hour turnaround time, so from the time we sent the passages out to the time we got them back was a two-day turnaround. This would ensure that there was some comparability between the systems by themselves and also between these systems and CANDIDE, which provided unedited output as well.

I mentioned that there were two evaluations of the output of these 18 documents. The 12 passages that were originally English were used in a comprehension test evaluation. The purpose of the comprehension test was, of course, to determine the comprehensibility of the individual outputs, and also to have a direct measure of comparability—to whatever extent that that's ultimately feasible—because all the systems translated all the same passages (if you accept the fact that a back-translation is translating the same passage). This allowed a means of comparing all the different outputs against one another—and also, in the case of CANDIDE and, hopefully, in the future, of the other systems as well—alternate methods of operating the same system.

The 12 texts (numbered according to their occurrence in the sequence of 18) were arranged so that each monolingual test-taker took a comprehension test like the one in your packet. We looked for educated people who were literate enough to read about mergers and acquisitions but had never studied Japanese, French, or Spanish, or, ideally, any foreign language at all. Some of them were upper management staff at my company.

Each of 12 test-takers saw each of the passages and each translation type once. Twelve different test packets were made up from the 12 outputs: machine-assisted output from each of the systems, unedited machine-alone output from each of four control systems (SPANAM and SYSTRAN Spanish, French, and Japanese), and human-only output for each of the three languages done by the Level 2 translators at the sites—adding up to a total of 12 versions for each passage.

The other test involved a quality panel. In this case we took the six original passages from each language plus the language-specific outputs from six of the master passages (randomly selected) and put them together in another 12-passage packet which we presented to native English-speaking professional translators (Level 4 and Level 5 competency) who were experienced, respectively, in the three languages in question. So there was a Japanese team, a French team, and a Spanish team. Each of these teams consisted of one Government person, who essentially coordinated the process, and two outside professional translators, who assessed the quality of the various outputs by applying the criteria used to grade translators in the U.S. Government— namely, syntactic, lexical, stylistic, and orthographic errors. So this part of the exercise is externally motivated in the sense that it was originally developed for grading human translators.

The test scores were dimensions of, in the case of the comprehension test, the comprehensibility score against the amount of time it took to produce the translation in the first place. If the unedited machine-alone systems are considered to take zero time, then what's really being measured on the time dimension are the human-only translations by the Level 2 translators and the machine-assisted times. The quality panel was the same sort of thing, except that there's a quality score based on the criteria we just mentioned as the vertical axis, with time as the horizontal axis.

As you might expect, on the comprehension test there were fewer errors when the test-taker was reading the original English as it was printed in the *Wall Street Journal*—although there were still 13 errors, or an average of one per passage. Of the MT systems, SYSTRAN Spanish did quite well—i.e., led to very few errors—which was sort of a surprise given that it's considered a pilot system, while SYSTRAN Japanese, as you might suspect from the same caveat, led to a large number of errors. Among the machine-assisted systems, it can be seen that the CANDIDE human-assisted version did quite well in terms of errors, while the LINGSTAT human-assisted Japanese system did quite well on time and PANGLOSS was in the same clump in terms of comprehensibility and at the extreme end with regard to time.

Now for the results of the quality panel. Fortunately, the results look quite similar to those from the comprehension test, which may help us answer a number of questions about the validity of using back-translations and that sort of thing—I would like to think it would. Half the texts are original passages and the other half are back-translations. The only real changes now are that SPANAM did best in terms of quality (with the assumption of zero time for unedited translation); the CANDIDE human-assisted system did best in terms of quality and time considered together; and PANGLOSS and SYSTRAN Japanese remained approximately where they were in the comprehension test.

The next round is still under negotiation and will probably take place sometime next year. One of the things we learned from this evaluation was that finding lots of categories of people to do lots of different small tasks is very daunting from a logistic point of view. If we can find an equivalent methodology that uses fewer categories of people, we'll be much better off. So we are trying to reduce the number of experts needed for the evaluation.

We are now working on a methodology that will improve the objectivity of measurement, increase granularity, and at the same time be simpler for an untutored monolingual to score.

This concludes my brief presentation. I would like for the members of the panel to spend a minute talking about particular issues that have been important for them in this evaluation.