INTERNATIONAL COORDINATION: International Working Group on MT Evaluation

Margaret King
ISSCO, University of Geneva

Back in 1990, at the Third Conference on Theoretical and Methodological Issues in Machine Translation, in Austin, Texas, there was a panel on evaluation, organized by Sergei Nirenburg, that generated quite a lot of heat, even if it didn't generate all that much light. It was that discussion that finished up with saying "well, what we really need is open discussion that goes on and doesn't just stop here at the end of this workshop." And that's how this thing called the International Working Group for the Evaluation of Machine Translation Systems got started.

In one way, that working group has gotten overtaken by events; other kinds of organizations have started to do *some* of the things that the International Working Group was meant to do. That's not a complaint. I think that's wonderful. I think forming the group actually helped to trigger some of these other events and, therefore, the group was serving some of its purposes indirectly.

One of the things that the International Working Group did do, however, was to organize what we called an Evaluators' Forum in the spring of 1991. The basic idea was to bring together people who have experience in evaluation in one place and get them to talk to one another about it. Experience in evaluation wasn't limited to people who had actually *done* evaluations. It also could mean people who had been evaluated or people who had some vested interest in getting evaluations done. It was a very intense discussion that just didn't stop; it started with the first papers and was still going on when people went away at the end.

I said that this type of activity has, in some ways, gotten overtaken by events. Subsequently, after the International Working Group got started, the European, Asia-Pacific, and American Associations for Machine Translation got started as well. The International Working Group has moved itself over into the associations and their activities. Within the context of the European Association for Machine Translation, the International Working Group has sort of changed its hats around and said, "Please, let us be part of your activities, too," and I hope that's going to be in collaboration with the other two associations.

That's one of the groups I wanted to talk about. There is another group that is quite independent. Recently the European community started an initiative with the name of Expert Advisory Groups for Language Engineering Standards (EAGLES). One of the EAGLES groups works on evaluation and assessment. The idea is to try—over a fairly long time, but with an initial short-term plan—to produce some consensus as to what counts as a reasonable evaluation. In that sense it's a group that is interested in producing standards for language engineering and language engineering work. That's important, I think, because one of the things that Yorick *didn't* say, but was underlying a lot of what he *did* say, was how many appallingly *bad* evaluations there have been.

It's been a funny state of the world for the past 25 years. When individual evaluations have been happening, they've tended to be on behalf of particular funding agencies or particular companies who were thinking about buying a system—some kind of individual effort—that quite often has happened in a vacuum in the sense that there was very little discussion in the literature. Evaluation was something that you did on the side. You didn't become an expert in evaluation. In a way, perhaps you even *couldn't* be, because then you might be thought to be biased.

So nearly all evaluations were done in a closed room, by somebody who was starting from scratch, trying to discover what he could from the literature, discovering that in the open literature there was very little reported, and finding out that a lot of previous evaluations had appeared in the form of reports that were hard to get because they were regarded as confidential. That meant that if you were hired to do an evaluation for somebody, you could go completely wild. You probably didn't intend to. You probably intended to be objective and sensible and more or less reasonable about what you did. But even with the best intentions, people can use crazy methods. They can come to crazy conclusions. And, I think, anybody who's been around in the machine translation world for the last 20-odd years could cite you cases where they're pretty well convinced that somebody did an evaluation with a crazy methodology and came up with crazy results.

Now, that's not in anybody's interest. It's not in the interest of people who are working on machine translation, nor is it in the interest of the customers who are using the results of the evaluation as a basis for their decision. So it is quite important to work together toward some kind of consensus about what counts as a reasonable way to do an evaluation and what counts as a valid basis for a judgment at the end of the day. That's one of the reasons why the Working Group on Standards is interested in evaluation and assessment.

In reality, this group is quite small. Its current membership is only six people, and none of them can dedicate a very great deal of time to it. On top of that, it's a group that's interested in *everything* to do with the computerized treatment of written text. It's also interested in things like natural language interfaces to databases, information retrieval and management systems, authoring aids, spelling checkers, and pretty well everything that deals with natural language. As a consequence, in the short term (the first two years) it is looking at much more modest natural language products—spelling checkers, style checkers, grammar checkers—and trying to work out evaluation methodologies for these.

There's a reason why it has limited its scope. The group has limited resources and, at least in the beginning, a comparatively short lifetime. You want to go for something where you think you can get realistic results in a reasonably short period of time. Also, of course, you want to go for something that's complementary to what other people are doing.

I want to finish up by coming back to the title of this session, "International Cooperation," and emphasizing how important it is that we *do* get collaboration, open discussion, coordination of efforts—that we try to come to some consensus as to what counts as evaluation and what counts as a reasonable way to do it. If we don't, then I think we're all—and all means funding agents, manufacturers, research workers, and customers—committing suicide.

Discussion

- I'd like to ask if there have been any efforts to base evaluation of machine translation systems on the principles of human translation systems or the grading of the output of a human translator.

- *(King)* You've got an assumption there that I find interesting. I also happen to teach in the Translators School, and one of the things that constantly comes up for discussion is how you grade human translation. I haven't yet seen anyone come up with a very good answer. Among translators and people who teach translation, we usually end up saying you kind of lick your finger and put it up into the wind. Of course you go checking for things that are just not right—like when somebody drops the negation out of a sentence, you know that that's a very bad mark against the translation—but you don't take that as your starting point because we're talking about professional translators. You start from the idea that the sense is going to be right, then you judge whether you've got a good translation. It's very hard to come up with any sort of abstract idea of what a good translation is. It seems to me, even with human translation, that you have to start by asking what the translation is for—what its purpose is, what it's going to be used for in the end—before it even makes sense to talk about grading it. If that's true for human translation, it's equally true for machine translation.

• *(John White,* PRC) I'd like to respond to that question. The DARPA evaluation adopted the method used by the U.S. Government for grading translators on the basis of translation samples that they do. We use that same methodology on the output of machine translation.

• *(Vasconcellos)* I expect that those of you who did your Warm-up Exercises will have become aware that the criteria are different. I had this problem. I'd say, "Oh, that makes sense for grading a human translation, but I don't see that it's really capturing the problems the machine is producing." So there are going to be differences.

On another point, I don't think it's safe to assume that professional human translators don't misinterpret the sense of a text. It happens that even the best of them can misunderstand the original language. Natural text is a real challenge!

• *(Jackie Murgida,* FBIS) That was my point, too. Especially in languages like Arabic, which I'm familiar with, you will often get people who are not translating into their native language, and they *do* miss very basic things.

• *(King)* Hold on! There are two separate situations here. One of them is where your *source* text is a source of confusion, and it would be forgivable for anybody to misunderstand it. The other situation is where there *isn't* any problem in understanding the source text, and *still* somebody gets the translation wrong in the sense of misrepresenting its content. Now, I'm not saying that that doesn't happen, but I'm saying that it *ought* not to happen.

• *(Doris Albisser,* Unibank) I would like to emphasize the point that Maghi King just made. I'm from the Union Bank of Switzerland, and we're working with a machine translation system. When we were evaluating MT systems we realized that you just cannot compare human translation with a machine-translated text. In the first place, machine translation is never intended to fully replace a human translator. Moreover, the mistakes a machine translation system makes are totally different from the mistakes a human makes, be it a translator or a bilingual person. For these reasons we decided to apply different criteria for evaluating machine translation systems: How fast could you postedit a text? How easy was it to understand the target text? and What linguistic criteria the translation system failed to meet?—that is, vital or crucial linguistic things that should be right. So definitely, I think you should apply different criteria from those that you use to grade the work of human translators.

• *(Klaus Schopmeier,* Translingua) I'm here to find out at what point a company such as ours—we are very heavily involved in packaging, labeling, and documentation for biomedical products—should join the game. It seems to me, coming from a very practical side, that what makes a *good* translation is a vastly different question, and the parameters are basically dictated by the requirements of the customer. In two different fields, different translations may be required, both of them equally good.

• *(King)* I totally agree with you that the purpose for which a translation is going to be done and what it's going to be used for are of critical importance. I find it very hard to talk into empty space about what counts as a good or a bad translation. I need to know what it's for and what the criteria are in that particular situation before I can even *talk* about evaluating a translation. And that's one of *my* problems with the DARPA methodology, just to cause some trouble.

• *(Wilks)* We have had two completely opposed positions set forth in the last five minutes, and they can't both be true. If the evaluation of machine translation is to be a function of the needs or function of the translation, then there *can't* be different criteria for human and machine translation.

• *(King)* No. Before you start saying something is good or bad, you have to say what it's good *for* or bad *for.* It doesn't follow from that that judging a machine product and judging a human product has to be the same process just because the criteria you're trying to meet are the same.

• *(Mark Mandel,* Dragon Systems) We are one of the DARPA contractors. I would like to propose a mediating position somewhere in between: the ultimate criteria may well be, and I think should be, the same—namely, What is the translation for? Does it suit the purpose? However, since the state of the art doesn't yet make the two comparable, then, perhaps it's not useful, in evaluating a machine translation, to make the same demands of it that you would of a human translation. However, I don't think that that's universally true. Especially if we are examining different technologies for achieving machine translation, we will find many different kinds of mistakes being made—kinds of things that need to be improved in the machine translations. As I seem to recall, the whole question arose here out of whether we could gather useful experience for developing evaluations for machine translation from the experience in evaluating human translation, and I think that still needs to be looked into and can't be settled simply by a priori judgments.

• *(King)* I'm sorry, I wasn't intending to imply that. First, of course, I'm not suggesting that everybody stop reading the machine translation literature. I'd be the last person to suggest that. And I'm not suggesting that everybody stop looking at how human translations get evaluated. The actual question was: Has anybody used the grading schemes that are sometimes used for human translation? And John replied that some of the DARPA methodology is based on this approach.

I'd like to take as an example the type of translation problem that Klaus Schopmeier mentioned: a package insert inside a medicine box that tells you how many of the pills you have to take every day, what the indicating symptoms are, what the counterindications are, and so on—which follows a rigid style in a closed domain. In that case your criterion for acceptability is going to be total accuracy. So the criteria are going to be the same, and, up to that point, Yorick is absolutely right, I think. But then, you might want to impose some additional criteria on top, which may well vary in the two cases of human translation and machine translation. In this particular case, for example, you might want to put on an additional criterion for a machine system that has to do with how many of the things it can churn out every day, which might be less relevant for a human translator who is doing other things as well.

But it does seem to me that, even though some criteria might be the same some of the time, finding out whether a system meets that set of criteria doesn't *have* to be the same in the human case and in the machine case.

• *(Chris Montgomery,* LSI) The gentleman from Translingua had a very interesting problem because he needed high accuracy/high quality, on the one hand, which is more an attribute of human translation than of MT, and, on the other hand, he needed consistency, which is not an attribute most of the time of human translators, and in a lot of cases isn't even desirable. So there's an unusual set of circumstances. Perhaps in this case we can bring in an MT type of attribute—namely, consistency—into human translation. Maybe the whole thing can work both ways. Anyway, that was an interesting and challenging problem he brought up, which needs thought on both sides and criteria for both sides.