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

Keynote: Traditions in the Evaluation of MT

Yorick Wilks
Computing Research Laboratory
New Mexico State University[1]

It has been a cliché in the field for years that machine translation evaluation is a better founded subject than machine translation. The reason why one could say this is that over a substantial period there would seem to have been less fundamental dispute in machine translation evaluation. There has *certainly* been fundamental dispute in machine translation. That is very evident at the moment in MT efforts in the United States, and it's a very interesting situation for empiricists. However, there also is *some* dispute, as you know, in machine translation evaluation, and one of the great virtues and interests of this workshop is that we are going to make it explicit.

Machine translation evaluation shares *some* features with MT, particularly what I like to think of as the easy start-up feature. A slide I've shown for some time says that "Any theory will give you some MT." It's an unfair slide, but there's a great deal of truth in it. We all know, if we've been around for some time, that there is *no* theory, however silly, upon which you cannot base *some* machine translation. That's what makes MT quite different from, say, space and flight, where not just *any* theory gets an object off the ground. And of course, MT evaluation shares this easy start-up feature, because it is not hard to do a naive machine translation evaluation: you just *look* at it. Do the two say the same thing or not? But we know, don't we, that that's not quite enough.

Since U.S. machine translation has had something of a revival in the last few years, this is a very timely moment for evaluation. I'm not sure who currently holds the record for the world's largest MT project. It was EUROTRA for a long time, but I suspect that at the moment it is the DARPA[2] machine translation project, where, because of the U.S. empiricist tradition, evaluation was built in right from the beginning. Also, evaluation has been very successful with speech research here in the United States in recent years, and many believe that if it has been so successful with speech, it ought to be equally successful in MT.

One may doubt that argument, but there can be no question of the value of evaluation. One cannot be in machine translation and say "We know we're right, we know we're doing good stuff, but we'd rather not be evaluated"—although that feeling does linger in some corners of the field. I put it down to the linguistic inheritance in the mixed gene pool of MT. The linguistic tradition has been antiempirical, at least in the last 40 or so years, in this country.

I suggest that there are four traditions of MT evaluation, of which the first and last are the most interesting, though not ones we shall be discussing here. For those of you who don't know the story of PIMA, this was a restricted form of English created during the Vietnam war for the maintenance of helicopters. As I recall, there was a very simple test developed for the adequacy of PIMA that measured the number of helicopters that crashed which had been repaired and maintained using PIMA versus those that had been repaired and maintained using either full English or a (not then available) full Vietnamese manual. That's what I would call a radical behavioral test. In theory, some such methodology *could* be developed for MT. It would have to be a wholly domain-orientated methodology. If it was financial mergers you were interested in translating, for example, there could be a financial test based on it. It's a test staged entirely in terms of the subject matter.

The last type of evaluation is one I was once involved with many years ago—namely, an evaluation of SYSTRAN for the Air Force. Its purpose was not to test SYSTRAN—it was assumed that SYSTRAN

---

[1] Currently with the University of Sheffield, Sheffield, UK.

[2] Since the Workshop, the Defense Advanced Research Projects Agency (DARPA) has changed its name to Advanced Research Projects Agency (ARPA). The earlier name has been used in these proceedings to keep the perspective consistent.

worked—but rather to see the degree to which SYSTRAN could be improved by certain well-defined updates and moved to another subject area, in particular from the areas it had been traditionally concerned with to political and economic types of text. It took SYSTRAN quite a while to get to where that was possible.

This brings me to my real point for this talk. The trouble with MT evaluation is that it is hard is to know exactly what you're testing. As we know, once we get into the detail, it's all very far away from the simple idea of "is this a translation of that or not?"

To me, the interesting question is whether MT and MAT—the second and third types—are essentially different or not. This has a special local relevance, because, as you'll soon find out today, the DARPA MT project contains within it both machine-aided and machine translation systems, and great effort has been put into the evaluation methodology organized by John White to try to compare machine-aided and machine translation together in the same test. This has turned out to be extremely difficult and extremely interesting.

Most of our effort over the last 20 or 30 years has gone into MT, and testing MAT turns out to be very, very hard because it is open to a trivialization that's not available in the case of MT. The notion of a human in the loop is extremely difficult because that person can bring the translation being done up to his or her level of translation either by editing or by starting from scratch and ignoring what's on the screen. It is then very hard to test what the quality of the "aid" was.

This makes it very hard to separate out what the MAT system is providing and what the human inside the loop is providing. This has meant, both traditionally and in the tests that have been done recently, that the only way to cut through and get to something empirical and firm is to go for what you might call ergonomic considerations. These have included the suggestion that you should judge MT evaluation by the number of keystrokes required to change a given form into another one that is more adequate. This is an interesting idea that came from the IBM group some time ago.

The ergonomic consideration that was used in the current DARPA MT test has been the time required to produce an augmented translation by the human in the loop. This has led to an extraordinary feature of that test. Here you have a range of systems: some of them require days to run in order to produce the translation, some require hours, some require minutes. You stick a human in the loop with all of them, and you then produce an impossible situation. Humans will not sit for days waiting to see what they will want to augment. What then happens, of course, is that the material is run in advance so that you have the test with the translation pre-prepared for the human to upgrade. The starting conditions can be radically different for different groups, and that is in fact what happened with the DARPA MT groups. Some have *enormous* time lags because they're running in real time, and some have very small times because the work was all done in advance and provided to the human pre-compiler. It seems to me that, although the DARPA team methodology has been extremely interesting, it is a unique departure to compare different things.

My last point concerns the correlation of intelligibility with accuracy. It's an empirical discovery of this field that intelligibility and accuracy are strongly correlated. We all know from everyday life that fluent and intelligible people are often accurate and possibly telling us the truth. It carries a certain authority with it. But, as we also know, it's an extremely dangerous belief. Fluent and intelligible people are frequently wrong.

The important practical point for MT evaluation is that if intelligibility and accuracy *are* strongly correlated, it means that monolinguals can do MT evaluation because all they have to judge is the intelligibility, which they can judge. The accuracy will come with it if the two are sufficiently correlated.

Ever since the ALPAC Report, machine translation has been seen fundamentally as an economic question. It's not "Can you do it?" but "Is it worth doing?" Similarly, of course, machine translation evaluation itself is an economic question. It turns out, as we've seen in the recent DARPA exercise, to be an extremely expensive business. To divert funds to evaluating the claim of a strong correlation of accuracy and intelligibility may not be something that people want to do. However, it is important.

Recent statistical MT, one of the new major players in the game, has thrown up the question in a new way. You could describe the IBM statistical MT system, CANDIDE, as one which gives intelligibility based on an n-gram analysis of the target language and then drags a certain amount of fidelity with it—about 50%—which is, by chance, the redundancy rate of the character strings of English. Or, to put it crudely, as a final question, what we need to know is whether machine-aided translation and machine translation *can* be effectively compared without reducing either one to the other. If they *are* radically different, you cannot reduce one to the other. That is a question we have to test, and I would like to see the answer to it.

Discussion

• *(George Doddington,* DARPA) What is a quick and easy definition of "intelligibility?"

• *(Wilks)* Does the string in front of you make some kind of clear sense? Can you assign a clear interpretation of it without thinking about where it came from or whether it's a translation of anything else? People make this kind of judgment in a very consistent manner.

• *(Doddington)* Where would one go to find out about tests on intelligibility?

• *(Wilks)* I know, from what little evaluation work I've been involved in, that you do see a strong consistency among intelligibility judges, as you do in most human performance tests. Take a set of 10 judges, ask them to make intelligibility tests, and you'll find that, given standard statistical criteria, they cluster together in the way you'd expect. Possibly one judge sticks out, and you do the usual thing and drop that judge. But there's the *standard* consistency between judges that you'd expect in human performance tests. That seems to be the evidence that intelligibility is a consistent, testable faculty.

• *(Henry Thompson,* University of Edinburgh) There is a tradition that started in the 1940s or earlier—George Miller and various other people did work on this—of intelligibility testing under noisy conditions. From the information theory point of view, they were looking at the what human judges could do with perturbed language of various sorts. I'm less confident than Yorick that there's really robust intersubjective reliability of those measures. There certainly is literature to be looked at.

• *(Wilks)* I think, from the sound of it, that you're referring to something quite different: the ability to interpret *perturbed* language, like how well you can understand when someone's shouting to you across a waterfall. That's not the same, I would have thought.

• *(Thompson)* George Doddington was asking for a definition of intelligibility. I think the point is that when you're asking people to make an intelligibility judgment, there's an a priori assumption that what you're looking at is *not* standard, clean English, or there wouldn't be any point.

• *(Muriel Vasconcellos,* PAHO) Who can tell us about recent work in the evaluation of translations or machine translations where intelligibility has been a criterion?

• *(Wilks)* You're going to hear today about the DARPA system. Also, the Air Force evaluation of SYSTRAN in 1979-1980 involved intelligibility judgments. It was largely done by monolinguals who couldn't speak Russian (see J. Newton, ed., *Computers in Translation* London: Routledge, 1992).