

PANEL: FUTURE DIRECTIONS IN MT

Introduction: *Sergei Nirenburg*, Moderator
Carnegie Mellon University

I would like to start with a very brief quasi-taxonomy of the issues that we might talk about in regard to future directions for research. We want to introduce research and evaluation here as a discussion point for possible types of future directions.

First of all, What constitutes progress in research in machine translation? What are the dimensions of this progress? In other words, How many different things can be evaluated? We can also talk about improving the evaluation as such—that is, the quality and cost of it. Then there are the core technologies: the various types of theories that we can study, various types of algorithms, what could embody these theories, and knowledge which is not procedural. These are some of the major dimensions of the core technology.

One of the issues that I find very interesting is the applicability question. It's one thing to develop a certain theory, methodology, algorithms, knowledge, and so on, but we also need to discuss how appropriate the approach is to the potential applications of machine translation. In other words, Where is the line beyond which it would not be appropriate to import methods from some other discipline into the machine translation task? In general, it is a question of method-driven versus task-driven approaches. In the task-driven approach, you first look at the task at hand and then you think about what methods are most appropriate for solving it. The other approach is, "I have a great method—it can do Problem A. Why don't we try it on Problem B as well?"

Finally, there is the application itself—and MT *is* an application, by definition; it is *not* a core technology. So we need to think what we can say about configuration and systems work in MT research as an applied field. Also under this heading is the matter of acquisition of knowledge and making it essentially cheaper and more reliable.

Then we also may discuss the idea of new paradigms—i.e., new system configurations and so on, which is different from new theories for MT. And, finally, we can consider the possibility of integrating old, existing paradigms in interesting ways, because there certainly are several options.

I will start the discussion by saying that I believe that a very important future direction—if not *the* most important future direction for us—is trying to automate (or semiautomate or three-quarters-automate) the acquisition of knowledge for MT systems so that we make the breakthrough from limited-size systems to real-size systems. In a nutshell, I believe that's one of the important directions that MT research should take.

Now I would like to ask the panelists for their views.

Panelists

Satoru Ikehara, NTT Network Information System Laboratories: I think the aim of machine translation in the future will be the same as it is now—that is, to improve translation quality. I believe that how to do this depends on the type of input text as well as the translation technology involved. In planning our policy and research effort, we in our laboratory needed to answer two questions: How can we surmount the differences between Japanese and English, and What is the knowledge needed for translation? I want to emphasize the importance of these questions.

In regard to the first one, consider the relationship between language analysis and translation quality. There is no doubt that meaning analysis can be done correctly. A translation rate of over 90% can be obtained in Japanese-to-Korean translation. But our experience shows that, even when the analysis is done correctly, it is difficult to get even 80% acceptable translation in Japanese-to-English machine translation.

However, the remaining 20% cannot be translated without an understanding of the culture in terms of both common sense and specialized knowledge. This example shows that the translation technology is highly dependent on the pair of languages involved.

The knowledge that is needed for translation can be classified into language knowledge, text knowledge, common sense knowledge, and specialized knowledge. Language knowledge is knowledge about conventions or rules about the meaning of a word or syntactic rules. We have classified sentences into five types based on what kind of knowledge is needed for translation, and we are now investigating the relationship between these sentence types and the cultural translation.

With this in mind, we have proposed a new translation method. In order to prepare the language knowledge needed for meaning analysis, we have developed a semantic representation system and large-scale meaning dictionaries, including a semantic word dictionary and a semantic structure dictionary. We have also developed several new functions to assist us in overcoming the differences between the languages involved.

Thus, these two questions—understanding the differences between the languages involved, and identifying the knowledge needed for translation—are vital factors in a research policy, and I believe that taking them into account will lead to further development.

Bonnie Dorr, University of Maryland: My feeling is that we've done a pretty good job here—and previous to this—talking about black box evaluation techniques. I believe that now there should be a move toward studying what it means to have a glass box evaluation technique. We need to use these black box evaluation techniques in conjunction with glass box evaluation techniques, and our future research directions are also parallel to this move.

In order to approach glass box evaluation, we need to spend a lot of effort classifying the kinds of distinctions that can arise between a source language and a target language. A number of people have started to do this now, and it would be nice to take advantage of that. From the interlingual perspective, people have studied the divergence problem; from the transfer perspective, people have studied the problems of the complex lexical transfers that arise in translating from one language to another. What would be nice is to then judge the internal representations that are used in these systems—to have a way of judging them to decide whether they characterize certain source-language-to-target distinctions.

Several people have studied this problem, including people like Elaine Rich and J. Tsujii and their colleagues at ACL and COLING this year. We need to look at the work that has been done to probe more deeply into the adequacy of representations that are used for machine translation. I call this “coverage adequacy”—ensuring that the internal representation used for translation covers the concept that underlies the source and target language. This is somewhat the analog to fidelity in the black box evaluation procedure, but it's also a little different.

Essentially, you want to ensure that information is not lost during the translation. There are criteria that we can apply based on knowing what the different divergence classes are, for example, that are defined in the literature. The divergences that occur take place at different levels. I can cite a couple of examples at the lexical semantic level. You also have to think about this at the syntactic level, at the knowledge level—at all other levels as well. For example, there is the head-swapping case, where in English you have “swim across” and in French the head has to be the “crossing” component and “swimming” becomes the argument of that head. Another example is conflation: in English you conflate manner with motion, whereas in Spanish those come out as two separate constituents. These types of problems have been studied, and it turns out that there are a small number of such distinctions that can arise across languages.

What we need to do is to characterize the representation and the translation process in such a way that it allows us to prove, perhaps formally, that each divergence category is fully covered—in other words, that coverage adequacy is met. This is an area that I think should be investigated in the future. Ultimately, what we would like to be able to do is automate this process. As Henry Thompson said

yesterday, and I fully agree, it would be nice to have an automatic way of having a glass box evaluation so that we don't have to do that ourselves. Even the types of glass box evaluations that we do now are almost black box evaluations in that we say, "Gee, does this system handle ellipsis and anaphora? Let's feed it a sentence that has ellipsis and anaphora in it and see if it comes out the other end." In other words, we should go down to a level beneath that and look at a way to do glass box evaluation by formalizing the representations we use and actually proving, formally, that the concept is covered in the source and target language.

Virginia Teller, Hunter College and the Graduate School, the City University of New York: From my perspective as an academic researcher, I have just one or two points I'd like to add to what has been said in the last two days.

The comments this afternoon suggest that one of the most important breakthroughs could be in the area of scanning technology (optical character reading) so that users of MT systems could input data with good results and also handle, for example, multi-column printed input and the formatting and reformatting for desktop publishing, so as to eliminate all that from the translation cycle. As for the evaluation process in the terms that it's been discussed over the last two days, evaluation is something of a mixed blessing for researchers. This point has been made repeatedly by Eduard Hovy and others. If evaluation spurs research, it can be of great benefit. But if it drives research to the extent that research focuses only on the criteria for evaluation, the result can be a highly standardized set of research results: a massive amount of redundant work among the groups that are involved in the evaluation and a focus on easier problems that can be solved by the time of the next evaluation cycle instead of harder problems that take a longer time to work out.

The MUC-type evaluations, which typically take a team as much as three to six months to prepare, would seem to be excessive for the machine translation community right now. A different approach has been adopted by the Association for Computational Linguistics, which earlier this year formed a special interest group on the evaluation of broad-coverage parsers. The group met last summer and also this fall at the University of Pennsylvania for two days. The people involved tested the output of their parsers against parses produced by the Perm Tree Bank and compared their results during the meeting in September, but without great import on the outcome. The teams spent something like a week on preparation, since essentially very little was at stake. In certain cases, this level of evaluation might be more appropriate than something along the DARPA lines.

Finally, in the context of evaluation and research, if the criteria for evaluation are, for example, how many sentences can be translated in a minute, or the quality and fidelity of a large amount of output, the work going on in my own research laboratory probably wouldn't pass muster. We're working on various aspects of multilingual natural language processing in the context of machine translation. While I can't necessarily say that everything we're working on would lead to a research prototype, it could ultimately be integrated into such a system if particular projects are successful—and of course not every project under way right now will ultimately lead to success.

In conclusion, evaluation is absolutely necessary from the researcher's point of view, but it shouldn't be allowed to stifle innovation.

Masaru Tomita, Carnegie Mellon University and Keio University: I would like to make two quick comments. First, it would appear that there are quite a number of commercial MT systems now, but not many people are using them. I think that these systems deserve to be used more. I will continue to use MT systems in my class in order to have students get used to working with them, and hopefully, after they graduate, they will continue to use some of them.

The second point is that in my research in the next couple of years I will be working on a voice translation system, especially the translation of dialogue and spontaneous speech. People never speak a grammatical sentence, so that is what I will be working on. Speech translation is not just a concatenation

of the speech recognizer, then translation, and finally the speech synthesizer. It is much more than that. A public demonstration will be given next January in three countries—Japan, Germany, and the United States—using telephone lines and a 400-word vocabulary in the domain of conference registration tasks.

Alan Melby, Brigham Young University: My point would be how to improve machine translation using existing technologies. That was one of the items in Sergei's initial remarks. I would suggest that one way to do it would be to go beyond the sentence level. Everyone agrees with that—that's like apple pie in America. But how? How can we do it practically in commercial systems?

One way—and not the only way—is to seek structured input with descriptive markup as opposed to presentational markup. That would mean seeking input to your machine translation system which is marked up with a meaningful SGML DTD. If you had input text which was marked up using the Text Encoding Initiative's DTDs, you could have some idea of how it all fits together. Your system would know whether something is a title or whether it's in the body of a paragraph. You would know relative relationships between paragraphs. And it would be much easier to take advantage of extrasentential information that is explicitly coded. That's my suggestion for the future.

In terms of what's available right now, there's one working group in the Text Encoding Initiative that has developed a terminology interchange format, so that you can more easily input partial records into your system and then add whatever other information your machine translation system requires. This format is ready for use and further comment. It has now gained the acceptance of the original MATER group, INFOTERM, and Technical Committee 37 of the International Standards Organization. It has also obtained the approval of the Localization Industry Standards Association and the Terminology Committee of the American Translators Association—so we have a lot of momentum.

Margaret King, ISSCO, University of Geneva: Leaving aside things like system architectures, toolbox approaches, and things like that, I would like to focus on the issue of evaluation, which has been the theme throughout this workshop. I may be prejudiced, but I think it is really true that evaluation is one of the most important topics.

It has also become increasingly clear to me in the course of the last two days that there's an awful lot of clearing away of undergrowth to do when we start talking about evaluation. Henry Thompson has referred to the distinction between "adequacy" evaluation, "progress" evaluation, and "diagnostic" evaluation. Listening to the discussions, I've been aware of a constant slide between those different types of evaluation. When people thought they were talking about adequacy evaluation, they were in fact talking in terms of progress evaluation, and vice versa. I think that's even true in some of the methodological paradigms that have been pointed out to us, like the DARPA paradigm.

I believe that one of the first and most important things to do over the next few years is to try to get our own heads clear about what context we're talking about when we refer to evaluation: what kinds of methodologies are appropriate to different kinds of contexts, and what kinds of data gathering-techniques are really reliable inside those contexts.

Makoto Nagao, Kyoto University: One thing I would like to point out is that almost all the present-day machine translation systems are essentially rule-based, and that approach poses a certain limitation because it is very rigid compared with human translation, where the translator has a lot of knowledge about the expressions to be translated. Recently, example-based machine translation has become rather attractive. It's fashionable. I believe that example-based translation will improve translation quality.

My other point, as Alan Melby has said, is that there is a kind of discourse problem—a problem of intersentential information processing—which is particularly important for MT word selection and the translation of sentential style. We need to include metaphorical expressions in the dictionary and interpret these expressions much more deeply. Also, we need to specify parameters for local topic domains that will change from sentence to sentence or paragraph to paragraph. That kind of information needs to be

made explicit, and I think that we can do that. We can detect this kind of information.

Also, speaker-hearer relations are quite important for the translation of dialogue, particularly for the Japanese people, who have very sophisticated honorific expressions. Speaker-hearer relations have a great influence on expression. That kind of information should be incorporated into the analysis and generation of sentences.

Another—much more difficult—thing that is also very important is guessing the intention of speakers. We need to be able to interpret why a speaker uses a particular kind of expression. With this knowledge we can then probably produce much better translation overall. That means translation with understanding, which is the ultimate goal for all of us. We should move step by step toward that ultimate goal.

Scott Bennett, Siemens Nixdorf Information Systems: I would like to add one more thing that I think we need to be clear about, and that is, if our systems are going to be used in the real world by real users—away from labs and away from those of us who created them and nurtured them and sort of coddled them—by real users, we also have to somehow build into any evaluation matrix a way of looking at the system from the users' perspective. We need to find out what they would like to see and then react to it. The user perspective has already been mentioned by a number of people here. This is an important dialogue. As we look at our systems, it is important to build in the realization that if they are going to be used by a user, we must somehow represent her or him when we evaluate our systems.

The other point that I would make, building somewhat on what Bonnie Dorr said earlier, is that we need not only a glass box kind of evaluation but an automated glass box. Again, I'll play manager and say that the more work you have to put into the evaluation process, the more costly it gets and the less inclined you are going to be to do it. If you can automate that, I think it will improve the chances of people actually doing evaluation and getting something out of it.

Eduard Hovy, Information Sciences Institute, University of Southern California: I was sitting and working at my system the other day when I fell asleep (I was waiting for it to run) and an old gentleman appeared to me in my dream and said, "Well, young fellow, my name is Paracelsus. I'm visiting you from the past, and I want to hear some things."

"Oh, my goodness," I say to myself, "Who is Paracelsus?" And then I remember: "Oh yes, the alchemist!"

"Indeed, the most famous alchemist of my time," he responded. "I've come to the future, and I would like to know what has happened. Have you reached the goal?"

"What goal are you talking about?" I asked.

He seemed incredulous. "The goal, of course, to transmute base materials into gold. In my time we put together this whole set of evaluation criteria for the different experiments. For instance, Is the product—you put things together and you get the product—is it more goldlike in color? That's one test. Another one: Has it got the right degree of hardness? Another one: Does it have the right specific weight? And so forth. Any experiment that did not fit nicely within these parameters, of course, was not helping us toward the goal of transmuting metals into gold."

I didn't quite know how to answer him. "This actually— This really— This hasn't really come about. What has happened, though, is that in around 1800 people discovered that if you combine different weights of things, you get different other kinds of combinations of weights. In fact, they've developed now a theory of atoms and so forth. They've got a field of knowledge that they call chemistry, and they can do all kinds of wonderful things in it. Chemistry's a very rich field with many, many subfields. Then later somebody discovered that the spatial orientation of atoms when they get together in large bunches is important. You can make plastics and all sorts of things."

Paracelsus got impatient: "But we go for the *gold*. Can you make *gold*?"

"Well, that's sort of not the point anymore," I replied. "We've got this rich field here ..."

Now, looking into the future (at that point my dream was interrupted—the system came back), I want to fall asleep and say to some young fellow 200 years from now, “Okay, young fellow, so have you achieved machine translation?”

And he’ll say, “What do you mean? Do you mean my e-mail assistant? Or do you mean my product manual translator? Or my newspaper translator? Or my court proceedings/legal translator? Or my business discussions thing (that little thing that I carry with me, like VERBMOBIL in Germany)? Or my information retrieval machine? Or my little box I carry with me when I’m a tourist (I point it at things and take photos and it translates for me)? Or the thing that translates speech to sign language (with a little hand or gestures)? What are you talking about when you say ‘machine translation’?”

And I think to myself, “Well, yes, what *do* I mean when I say ‘machine translation’?” The future is clearly not just one thing—nor, as far as I can see, is it just one technology. There are a lot of different things that go in here, and it doesn’t make sense for us to focus our noses very closely on just one particular idea—like going for the gold. We might never get there.

To wind up, I think what Sergei said at the beginning is very true in the short term. If you look at all these things, they all need words, they all need certain kinds of knowledge. I think our short-term efforts are going to be focused on the collection and systematization of large collections of words and grammars and so forth—to the point where we reach a basis upon which we can exploit their differences—use different algorithms or whatever—to build different kinds of systems.

Sergei Nirenburg: The diversity of opinions here on this panel was really serious. In addition to the points that I made, there were several relating to evaluation proper, and there were several points mentioned in regard to certain particular pieces of knowledge—especially knowledge that we need to put in. And, in the last point in Ed Hovy’s presentation, we also got into the application area. At this point I think it’s appropriate to ask anyone in the audience if they would like to add their comments.

General Discussion

- *Loll Rolling* (European Commission) I would like to emphasize something that Makoto Nagao pointed out a few minutes ago—namely, example-based MT. That is certainly a very important thing for the future, and I was astonished that nobody has talked about it at preceding MT Summits and here today. At the European Commission we have very large parallel text corpora in several languages. One example is the publications of the official gazettes of the member countries, of which over 100,000 pages are available in nine languages, translated by human translators. It is obvious that this is a large resource that can be exploited for the extraction of equivalent terms in all these languages to add to terminology data banks and also MT dictionaries. That is exactly what we intend to do in the months to come. We call it the EQUITEXT Project. The objective is to extract equivalent terms and phrases from parallel corpora. In addition to terms, we can identify equivalent phrases. For a given concept we can easily extract the equivalences in French, German, and English from the corpus, whether it’s a single word or a string of words. I think this is a valuable resource for machine translation, too. Terms can be introduced very easily into dictionaries; phrases must be validated by human coders. But I think we should not neglect this possibility in the future.
- (*Nirenburg*) One point I think I should make is that the type of knowledge you extract from sources of this kind cannot be used directly in a machine translation system unless it is a system built on a statistical basis. Dictionaries in MT systems that are based on the analysis of language require, at the very least, a format that is different from a list of bilingual terminology. But the general point is taken.

- (Thelma Litsas, Mead Data Central) I'd like to inject another perspective on the type of systems that users might need in the future. We are the providers of LEXIS and NEXUS on-line, and the size of our database is 180,000,000 documents. An average document is 3,000 words. The total number of words would be 550 billion. Most of this is in English, but there are also files in French, Italian, and German. When we count this in bytes, we count in terabytes. This is the amount of text we might, at some point, need to translate. I would like to put a plug in for standards: SGML standards for general markup, which we are espousing, and the markup template—the DTD that you were talking about for dictionaries or terminology.

I'm wondering if there are other standards in the area of representation. I heard a lot of mention of interlinguas: "We put this into an interlingua"; "Every system has its own interlingua"; and so on. I'm wondering if standardization in this area might be a thing to pursue if we're looking at research directions.

- (Nirenburg) From time to time in the research field a suggestion arises to standardize "X." However, very seldom does real standardization take place. I may be wrong, but it seems to me that success in standardization does not depend on good will or quality of research. It depends on support—on resources that have to come from the outside. In other words, we would be very happy to introduce some standards, but the initiative would need to be well-oiled.

- (Hovy) I can speak to that a little, too. There is an effort, not under DARPA sponsorship but under the DARPA umbrella, to standardize knowledge representation languages, and part of this is to standardize terminology. There is something called ONTOLINGUA that has been built by people at Stanford with the idea that you can then take your representation system (something like your interlingua) and write transfer routines that would capture as much as possible and put it into the ONTOLINGUA so that it could be available for everybody else.

As Sergei said, however, these things very seldom have much chance for success. In fact, John Sowa once pointed out that whenever somebody tries to impose a standard, another *de facto* standard pops up—like when some people were arguing for PASCAL or ADA to be a standard programming language, C appeared instead and became the standard in many cases. So maybe there's a forcing function or maybe not, but it doesn't seem likely.

- (King) May I interject at least a very brief plea to not get knowledge representation languages and interlinguas mixed up. There are a lot of us who don't even believe in interlinguas.

- (Nirenburg) There is a technical difference: an interlingua is a means of representing some of the meanings of a text. It could, in fact, be done in English, except that it couldn't be processed by the machine.

- (Beth Sundheim, NRaD) Virginia Teller's comments have led me to want to say a few things about research and its relationship to evaluation from the MUC experience. I particularly liked her terminology when she was talking about *spurring* research versus *driving* research, where the good thing would be research inspired by evaluation and the bad thing would be research and development being done only in order to satisfy the requirement of evaluation imposed on them. In this regard, I think our experience does tell us some things that are negative. In the first place, if it does take six months to do a decent effort, then the evaluations really do need to be held less often. Moreover, it shouldn't be surprising if the evaluators discover at the conference that some of the researchers don't have too much to report.

I've been a little disappointed, in a sense, by people backing off from their natural language research

to explore shallow techniques. However, I do think that they're only trying to do the easy problems *first*. I believe they all acknowledge that the harder problems are still there; they know that they've still got to do them, and everybody's got to get something together to the point that they can get beyond that. We need to keep pushing on that.

If the effort is somewhat redundant at this point, it's just in order to get to that stage of development at which something more interesting can be done. And every year people do come through and show us something new and interesting.

- (Nirenburg) Do you think that stage will ever come?
- (Sundheim) Yes, I believe so.
- (Muriel Vasconcellos, PAHO) In all of this, I don't think we've said enough about fact that we need to understand more about language itself—in particular, discourse. We still don't understand the forces involved in producing truly coherent translations—the many factors that are entailed in patterning information, identifying themes in a text, and so on.

If we're actually going to improve in the areas that are toughest, then we've got to begin to explore much more discourse than we're looking at now. If we're only going to reduce text to pieces of the kind that Bonnie suggested, we're never going to reach our goal totally. I can give you cases where I've taken the same concept and rendered it maybe five or six different ways in the same translation, depending on the informational force that it has at the particular point in the development of that discourse. There is no such thing as a single translation for a single phrase in another language, really, because so much depends on how it is going to occur. We need to be much more sensitive to those things if we're going to “go for the gold.”
- (Nirenburg) So we are all in agreement—we know what we will be doing next! Let me just point out one thing—it's a meta-comment. Someone once said “Science is the art of the possible.” That seems to me to be a very apt quotation right now. In our plans for future research we think in terms of what we need to do next because we actually see a hope to getting there somehow. At the same time, MT is a difficult application for this type of attitude because there are certain results that we must achieve. Thus it might be valid for us to say that we are still artists, in some sense, and not just scientists.