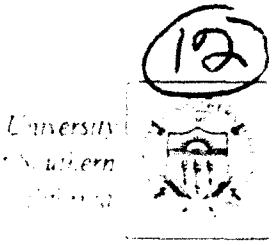
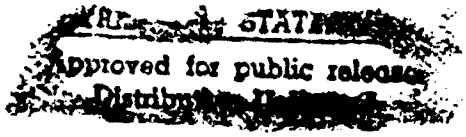


Good Applications for Crummy Machine Translation

Kenneth W. Church
AT&T Bell Laboratories

Eduard Hovy
USC/Information Sciences Institute

July 1991
ISI/RR-93-352



INFORMATION SCIENCES INSTITUTE

310/822-1511
4676 Admiralty Way/Marina del Rey/California 90292-6695

Good Applications for Crummy Machine Translation

Kenneth W. Church*
Eduard H. Hovy**

*AT&T Bell Laboratories

**USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292

Abstract

We have recently begun work in machine translation and felt that it would probably make sense to start by surveying the literature on evaluation. As we read more and more on evaluation, we found that the success of an evaluation often depends very strongly on the selection of an appropriate application. If the application is well-chosen, then it often becomes fairly clear how the system should be evaluated. Moreover, the evaluation is likely to make the system look good. Conversely, if the application is not clearly identified (or worse, if the application is poorly-chosen), then it is often very difficult to find a satisfying evaluation paradigm. We begin our discussion with a brief review of some evaluation metrics that have been tried in the past and conclude that it is difficult to identify a satisfying evaluation paradigm that will make sense over all possible applications. It is probably wise to identify the application first, and then we will be in a much better position to address evaluation questions. The discussion will then turn to the main point, a discussion of how to pick a good niche application for state-of-the-art (crummy) machine translation.

1. Introduction

We have recently begun work in machine translation and felt that it would probably make sense to start by surveying the literature on evaluation. As we read more and more on evaluation, we found that the success of an evaluation often depends very strongly on the selection of an appropriate application. If the application is well-chosen, then it often becomes fairly clear how the system should be evaluated. Moreover, the evaluation is likely to make the system look good. Conversely, if the application is not clearly identified (or worse, if the application is poorly-chosen), then it is often very difficult to find a satisfying evaluation paradigm. We begin our discussion with a brief review of some evaluation metrics that have been tried in the past and conclude that it is difficult to identify a satisfying evaluation paradigm that will make sense over all possible applications. It is probably wise to identify the application first, and then we will be in a much better position to address evaluation questions. The discussion will then turn to the main point, a discussion of how to pick a good niche application for state-of-the-art (crummy) machine translation.

Why work on machine translation now, and what kind of MT is most likely to be commercially and theoretically profitable? Though the ALPAC report concluded in the sixties that there should be more basic research in MT, it stated clearly that this basic research could not be justified in terms of short-term return on investment.[1] In particular, when compared with human capabilities (still the ultimate test), MT systems of the time were not deemed a success, and might never be.

This belief may help explain the resistance of many MT researchers to take evaluation questions seriously. The EUROTRA project, for example, consciously decided to delay evaluation discussions as long as possible: "Exact procedures for evaluation will be decided by the programme's management committee toward the end of each phase..." (Johnson *et al.*, 1985, p. 168). Others argue against any human-related evaluations as follows:

> "Performance of operational MT systems is usually measured in terms of their cost per 1,000 words and their speed in pages per post-editor per hour vs. the relative cost and speed of human translation.... In my opinion, it is becoming increasingly uninformative to compare the performance of MT systems with that of human translators, even though many organizations tend to do that to justify their MT investments." (Tucker, 1987, p. 28)

We believe that these attitudes hurt the cause of MT in the long run. As is proved by the increasing availability of commercial MT and MAT systems (such as Systran, Fujitsu's Atlas, Logos, IBM's Shalt, and several others, for less than $100,000), MT today is beginning to find areas of real (commercial) applicability. Thus, to the questions "Has anything changed since ALPAC? How can one build MT systems that make a difference?", we answer that the community needs to find evaluation measures and applications that highlight the value of MT research in those areas where systems can be employed in a real (and economically measurable) way. Human and machine translation show complementary strengths. In order to design and build a theoretically and practically productive MAT system, one must choose an application that exploits the strengths of the machine and does not compete with the strengths of the human. This point is well put in the following:

---

1. "The Committee recommends expenditures in two distinct areas. The first is computational linguistics as a part of linguistics-*studies of parsing, sentence generation, structure, semantics, statistics, and quantitative linguistic matters, including experiments in translation, with machine aids or without. Linguistics should be supported as science, and should not be judged by any immediate or foreseeable contribution to practical translation... The second area is improvement of [human] translation [with respect to practical issues such as speed, cost, and quality].*" (Pierce *et al.*, 1966, p. 34)

> "The question now is not whether MT (or AI, for that matter) is feasible, but in what domains it is most likely to be effective.... The object of an evaluation is, of course, to determine whether a system permits an adequate response to given needs and constraints." (Lehrberger and Bourbeau, 1988, p. 192)

What then are appropriate evaluation measures? It would be nice if the evaluations were to identify those (aspects of) MT systems that make them suitable for, and then steer them towards, high-payoff niches of functionality. But in spite of all the literature on MT evaluation, the general evaluation measures that are proposed often fail to pinpoint the strengths of systems and lead them toward real utility; instead, they seem to confound important and less important aspects. Tucker's review of Taum-Meteo and Metal, for example, might give one the mistaken impression that both systems work about equally well (namely, approx. 80%):[2]

> "Taum-Meteo has been operational since 1977, translating about five million words annually at a rate of success of 80% without post-editing." (Tucker, 1987, p. 31)

> "[T]he Metal system is reported to have achieved between 45% and 85% 'correct' translations, using an experimental base of 1,000 pages of text over the last five years." (Tucker, 1987, p. 32)

However, these numbers do not accurately reflect the crucial difference between these two systems. Taum-Meteo is generally regarded as a fairly complete solution to the domain-restricted task of translating weather forecasts whereas Metal is widely regarded as a less complete solution to the more ambitious task of translating unrestricted text. The evaluation measure ought to be able to highlight the strengths and weaknesses of a system. Apparently, the "success rate" measure fails to meet this requirement, presumably because it is too vague to be of much use.[3]

Unfortunately, this failure seems to be characteristic of many of the task-independent evaluation metrics that have been proposed thus far. Since, in our opinion, the blame is to be laid on the desire for generality, we propose that MT evaluation metrics should be sensitive to the intended use of the system. In this paper, we begin by outlining metrics that have been proposed and end by concluding that it becomes crucial to the success of an MT effort to identify a high-payoff niche application so that the MT system will stand up well to the evaluation, even though the system might produce crummy translations.

2. Traditional Evaluation Metrics

2.1 System-based Metrics

We identify three major types of evaluation metrics: *system-based*, *text-based* and *cost-based*. System-based metrics count internal data resources such as the number of words in the lexicons, rules in the grammars, semantic, grammatical, or lexical features, the number of representation elements in the semantic ontology or Interlingua (if any), and the number of translation rules (if any). The literature contains many examples of system-based metrics, for instance:

---

2. According to Isabelle (personal communication), Meteo currently achieves 97% success on a volume of 20 million words per year. The increased performance is largely due to improvements in the communication system; communication noise used to be responsible for a large percentage of the failures.

3. The success rate of 80% reported in (Isabelle, 1984, p. 265) probably should not be compared with the numbers reported for Metal. In addition to translating the input, Meteo also attempts to determine if the translation should be checked by a professional translator. The 80% figure reported in (Isabelle, 1984) refers to the fraction of the input that Meteo handles by itself without assistance from a professional translator. The figures reported for Metal refer to an evaluation of the correctness of the output.

> "At the moment there are about sixty subgrammars for analysis and about 900 rewriting rules in total... number of rewriting rules for transfer and generation processes is around 800, and it will be increased in the coming few months. The dictionary contains about 16,000 items at present, and will be increased to 100,000 items at the end of the project." (Nagao, 1987, p. 276)

An advantage of these metrics is that they are easy to measure, which makes them popular. But since these metrics are tied to a particular system, they cannot be used very effectively for comparing two systems. They are much more effective for calibrating system growth over time. The major disadvantage of these metrics is that they are not necessarily related to utility.

2.2 Text-based Metrics

2.2.1 Sentence-Based Metrics These metrics, the most common class, are applied to individual sentences of target texts by counting, for example, the number of sentences semantically and stylistically correct, the number of sentences semantically correct, but with odd style, the number of sentences partially semantically correct, the number of sentences semantically and syntactically incorrect, and the number of sentences missed altogether. A good example appears in (Nagao *et al.*, 1986), in which sentences are classed into one of five categories of decreasing intelligibility and into one of six categories of decreasing accuracy. Another example is the evaluations developed to measure the results of Eurotra systems (see Johnson *et al.*, 1985).

Given the subjective nature of semantic, syntactic, and (especially) stylistic "correctness", these metrics are impossible to make precise in practice. In addition, their limitation to single sentences makes them too simplistic (for example, it is not clear how to scale the metric when several source sentences are combined in the target text, or when parts of them are grouped into sentences differently).

2.2.2 Comprehensibility Metrics These metrics seek to measure translation quality by testing the user's comprehension of the target text as a whole. They include counting the number of texts translated well enough for full comprehension, the number of texts in which enough could be gleaned to get a reasonably good understanding of the content, though details may be missing, the number of texts in which some content could be gathered, enough to tell whether the text is of interest to the user or not, the number of texts with fatal inconsistencies or omissions, and the number of texts missed altogether.

These evaluation metrics enjoy some significant advantages. First, they can be performed by the intended user of the translation, requiring little or no source language expertise. Second, they take in stride the mis- or even non-translation of text due to certain relatively isolated phenomena which have proven very hard to handle in computational systems in a general way (but which people can figure out themselves fairly easily). A major disadvantage of these metrics is the difficulty of quantifying them. One approach to overcome this difficulty is to create comprehension questionnaires that measure (in SAT-test-like manner) how understandable translations are to their intended users with respect to their intended uses. An example, using a test suite of texts, is proposed in (King and Falkedal, 1990). A second approach is to determine how willing users would be to pay for professional translation of the text, given the translated version. Since professional translation is expensive, the users will be motivated to identify the more useful systems.

2.2.3 Amount of Post-Editing Metrics in this subclass are based on the amount of work required to turn the translated text into a form indistinguishable from a human translator's effort. Ways of quantizing this include counting the number of editing keystrokes required per page, timing the revision process per page, and counting the percentage of machine-translated words in final text. An example is the keystroke count reported as follows:

> "As an alternate measure of the system's performance, one of us corrected each of the sentences in the last three categories (different, wrong, and ungrammatical) to either the exact or the alternate category. Counting one stroke for each letter that must be deleted and one stroke for each letter that must be inserted, 776 strokes were needed to repair all of the decoded sentences. This compares with the 1,916 strokes required to generate all of the Hansard translations from scratch." (Brown *et al.*, 1990, p. 84)

Some researchers object to keystroke counting because they don't believe that the counts are correlated with utility.

2.3 Cost-based Measures

The third major type of metric concentrates on the system's efficiency in producing a translation, as in:

1. cost per page of acceptable translation (machine, human, or mixed),
2. time per page of acceptable translation (machine, human, or mixed).

One such evaluation was done on Taum-Aviation (Isabelle and Bourbeau, 1985)

Task	*Machine*	*Human*
Preparation / input	$0.014	$0.000
Translation	$0.079	$0.100
Human revision	$0.068	$0.030
Transcription / proofreading	$0.022	$0.015
Total	$0.183	$0.145

The problem with cost-based metrics is that they often don't make the systems look very good. As can be noted from the table above, the evaluation shows that Taum-Aviation is actually more expensive than human translation (HT). If one wants the system to look good, it is important to pick a good niche application.

Some might accuse us of "lying with statistics". There is a fine line between realism and fraud. We would say that it is realistic to pick an "easy" niche application if the application is likely to have real value (e.g., Meteo (Isabelle, 1984)). On the other hand, the strategy does run the risk of raising expectations unrealistically if the application only appears to have real value (e.g., the original GU experiment (see section 7.1)). Of course, we would want to concentrate our efforts on good niche applications that really do have value and avoid the bad ones that look like they ought to scale up to something useful but actually don't.

3. Characteristics of a Good Niche

We believe a good niche application should meet as many of the following desiderata as possible:

a. it should set reasonable expectations,

b. it should make sense economically,

c. it should be attractive to the intended users,

d. it should exploit the strengths of the machine and not compete with the strengths of the human,

e. it should be clear to the users what the system can and cannot do, and

f. it should encourage the field to move forward toward a sensible long-term goal.[4]

4. Extensive Post-Editing (EPE): An Inappropriate Niche

It is not easy to identify a good niche application. One cannot simply take a state-of-the-art MT program and give it to a bunch of salesmen and expect a miracle. One has to find an application that makes sense.

The extensive post-editing (EPE) application would appear to be a natural way to get value out of a state-of-the-art MT system. But unfortunately, the application fails to meet most of the desiderata proposed above.

4.1 (a) Realistic Expectations

One can find claims that EPE either increases or decreases productivity by anywhere from a factor of 1 to 2 or 2 to 1. No matter what the truth is, the application would probably be more successful in the marketplace if expectations were more realistic. One rarely finds disclaimers of the form "your mileage may vary" after some of the claims that have been made on behalf of the EPE application:

> "Although you can expect to at least *double* your translator's output, the real cost-saving in MT lies in complete electronic transfer of information and the integration into a fully electronic publishing system." (Magnusson-Murray, 1985, p. 180)

> "Substantial rises in translations output, by as much as 75 per cent in one case, are being reported by users of the Logos machine translation (MT) system after only a few months." (Lawson, 1984, p. 6)

> "For one type of text (data description manuals), we observed an increase in throughput of 30 per cent." (Tschira, 1985)

Statements such as these run the risk of setting unrealistic expectations, and consequently, in the long run, it is possible that they could actually do more harm than good. (We discuss the dangers of unrealistic expectations in section 7.) If users could really expect even modest gains in productivity, then one would have expected the EPE products that have been offered by ALPS, Logos, Systran, Weidner and others would have been more successful in the marketplace than they have been.

4.2 (b) Cost Effectiveness

In fact, we were rather surprised to discover that there have been a number of trials indicating that EPE might actually be more expensive than human translation (HT). For instance, Van Slype (1979) estimated that EPE costs 475 BFrs. per 100 words, almost twice as much as HT (150-250 BFrs. per 100 words). Similarly, the Canadian government found more or less the same result in their trial of the Weidner product:

4. Many long-term goals have been proposed over the years; FAHQT (fully-automatic high-quality translation) (Bar-Hillel, 1960, p. 94) is perhaps one of the more well-known proposals.

> "[T]he HT production chain was significantly faster than the MT production chain. How much faster depends on which phases of the MT chain are counted. If we count all the steps on the log form, human translation was nearly twice as fast as machine translation. If we discount the time that the machine actually takes to translate (on the assumption that the participants could use this time to do other useful tasks), as well as the time for the second dictionary update (on the grounds that these new or modified entries are not intended for the current text), MT remains 27% slower than HT. If, in addition, we discount the time for text entry, assuming that source texts arrive in machine readable form that Weidner could import, MT still remains 5% slower than HT for all the texts translated during the operational phase of the trial." (Macklovitch, 1991, p. 3)

In fact, there have been questions about the cost effectiveness of the EPE application dating back to the ALPAC report, well before many of these products were introduced into the marketplace:[5]

> "The postedited translation took slightly longer to do and was more expensive than conventional human translation... Dr. J. C. R. Licklider of IBM and Dr. Paul Garvin of Bunker-Ramo said they would not advise their companies to establish such a service." (Pierce *et al.*, 1966, p. 19)

It is difficult to know how to balance the results of these government trials against some of testimonials cited above. It is probably the case that EPE saves time and money in some applications, and hurts in others. No matter what the fact are, though, it is almost certainly the case that the field would be in better shape if expectations were better handled. It would be most unfortunate if a potential user were to buy into EPE, expecting to save a bundle, only to discover the hard way that it may not be cost effective in his or her particular application. We were rather surprised to discover that EPE could actually be slower than HT, but after thinking about it for a little while, it should have been obvious that it can take longer to fix a badly written piece of prose than it would take to start from scratch.

4.3 (c) Attractiveness to Intended Users

EPE has failed to gain much acceptance among the intended target audience of professional translators, because post-editing turns out to be an extremely boring, tedious and unrewarding chore.[6]

> "Most of the translators found postediting tedious and even frustrating. In particular, they complained of the contorted syntax produced by the machine. Other complaints concerned the excessive number of lexical alternatives provided and the amount of time required to make purely mechanical revisions." (Pierce *et al.*, 1966, p. 96)

> "Many, but not all, translators decided, after the first phase of the MT experiment, that Systran was not a translation aid, because they found that it took too long, and was too tedious, to convert raw MT into a translation 'to which they would be prepared to put their name.'" (Wagner, 1985, p. 203)

> "When asked by the consultant if they would like to continue working with Weidner on the same texts after the end of the trial, not a single participant accepted." (Macklovitch, 1991, p. 4)

After reading Macklovitch's description of some of the errors in (Macklovitch, 1986), one can easily appreciate why some of the translators would be frustrated with the post-editing task. Macklovitch observed that approximately half of the errors in one sample involved the overuse of French articles. In

---

5. The cost effectiveness of the EPE application is discussed in more detail in Appendix 14 of the ALPAC report. The appendix observed that postediting tends to "impede the rapid translators and assist the slow translators" (Pierce *et al.*, 1966, p. 94). This would suggest that EPE products might be more appropriate for casual use by an amateur rather than daily use by a professional.

6. Perhaps the task would be less tedious if the user interface were made more flexible and more user-friendly.

translating an English noun phrase into French, it is a pretty good bet that the French noun phrase should begin with an article even if there isn't one in English. However, this rule does not hold in tables, where the French use of articles is apparently somewhat more like English. As it happened, one of the texts used in the trial contained a very long list of crop varieties published by Agriculture Canada, most of which should not have been translated with an article. Unfortunately, the Weidner system did not know that noun phrases work differently in tables, and consequently, the post-editor was faced with the rather tedious task of deleting the article and adjusting the capitalization for each of the crop varieties in this very long list. The professional translator probably would have found it quicker and more rewarding to translate the list from scratch.

4.4 Kay's Characterization of EPE

One can continue to go through the list of desiderata proposed above and find even more reasons why EPE is an inappropriate niche. Rather than beat a dead horse ourselves, we thought we would let Martin Kay do it for us, as only he can:

> "There was a long period -- for all I know, it is not yet over -- in which the following comedy was acted out nightly in the bowels of an American government office with the aim of rendering foreign texts into English. Passages of innocent prose on which it was desired to effect this delicate and complex operation were subjected to a process of vivisection at the hands of an uncomprehending electronic monster that transformed them into stammering streams of verbal wreckage. These were then placed into only slightly more gentle hands for repair. But the damage had been done. Simple tools that would have done so much to make the repair work easier and more effective were not to be had presumably because of the voracious appetite of the monster, which left no resources for anything else. In fact, such remedies as could be brought to the tortured remains of these texts were administered with colored pencils on paper and the final copy was produced by the action of human fingers on the keys of a typewriter. In short, one step was singled out of a faily long and complex process at which to perpetrate automation. The step chosen was by far the least well understood and quite obviously the least apt for this kind of treatment." (Kay, 1980, "The Proper Place of Men and Machines in Language Translation," p. 2)

5. A Constructive Suggestion: The Workstation Approach

Having established that EPE is inappropriate, Kay then suggested a workstation approach. At first, the workstation might do little more than provide word-processing functionality, dictionary access and so on, but as time goes on, one might imagine functionality that begins to look more and more like machine translation.

> "I come now to my proposal. I want to advocate an incremental approach to the problem of how machines should be used in language translation. The word *approach* can be taken in its original meaning as well as the one that has become so popular in modern technical jargon. I want to advocate a view of the problem in which machines are gradually, almost imperceptibly, allowed to take over certain functions in the overall translation process. First they will take over functions not essentially related to translation. Then, little by little, they will approach translation itself. The keynote will be modesty. At each stage, we will do only what we know we can do reliably. Little steps for little feet!" (Kay, 1980, p. 11)

In his concluding remarks, Kay expressed the hope that his approach be implemented by someone with enough "taste" to be realistic and pragmatic.

> "The translator's amanuensis [workstation] will not run before it can walk. It will be called on only for that for which its masters have learned to trust it. It will not require constant infusions of new *ad hoc* devices that only expensive vendors can supply. It is a framework that will gracefully accommodate the future contributions that linguistics and computer science are able to make. One day it will be built because its very modesty assures its success. It is to be hoped that it will be built with taste by people who understand languages and computers well enough to know how little it is that they know." (Kay, 1980, p. 20)

In fact, Kay's approach has recently been implemented by people who understand the practical realities well enough to take an even more modest approach than Kay himself probably would have taken. CWARC (Canadian Workplace Automation Research Center) has undertaken to provide the Canadian government's Translation Bureau with a translator's workstation that could be deployed in the near-term to the bureau's 900 full-time translators (Macklovitch, 1989). For obvious pragmatic considerations, they have decided to use the following off-the-shelf components:

a. a PC/ AT,
b. network access to the Termium terminology database on CD-ROM,
c. WordPerfect, a text editor,
d. CompareRite, a program for comparing two versions of a text file,
e. TextSearch, a program for making concordances and counting word frequencies,
f. Mercury/ Termex, a program for maintaining a private terminology database,
g. Procomm, a program providing remote access to data banks via a telephone modem,
h. Seconde Memoire, a program that deals with French verb conjugations, and
i. Software Bridge, a program for converting word processing files from one commercial format into another.

This is clearly a sensible starting point for introducing technology into the translator's workplace. They will hopefully be able to demonstrate that the PC-based workstation is clearly superior to dictation machines. After they have achieved a trackrecord of success and the new technology has been in place for a while, they will be in a much better position to introduce additional tools, which might be more exciting to us, but also more risky for the managers at the translation bureau.

One might imagine all kinds of exciting tools. For example, the workstation could have a "complete" key, like control-space in Emacs, which would fill in the rest of a partially typed word/ phrase from context. One might take this idea a step further and imagine that it ought to be able to build a super-fast typewriter that would be able to correct typos and fill in context given relatively few keystrokes. Peter Brown (personal communication) once remarked that such a super-fast typewriter ought to be possible in the monolingual case, observing that there is so much redundancy in language that the user should only have to type a few characters per word, or about the equivalent of 1.25 bits per character (Shannon, 1951),[7] which is only slightly more than a byte (ascii character) per English word on average. The user should have to type even less in the bilingual case because the source language should provide quite a number of additional bits of information.

---

7. Shannon's estimate that English has an entropy of 1.25 bits per character is probably too optimistic. In practice, one would probably expect a practical system to have an entropy somewhat closer to 1.76 bits per character (Brown *et al.*, 1991).

The super-fast typewriter may still be a ways off, but we are almost already in a position to provide some very useful but less ambitious facilities. In particular, the Translation Bureau currently spends a lot of resources retranslating minor revisions of previously translated materials (e.g., annual reports that generally don't change much year after year). It would be very useful if there were some standard tools for archiving and retrieving previously translated texts so that the translators would have access to the previous translations, when appropriate. It is also becoming possible to use bilingual concordances to help with terminological issues.

The workstation application stands up to the six desiderata proposed in section 3 much better than the EPE application. It is (a) much more realistic, so it should have a better chance of (b) economic success. After all, it ought to be able to beat dictation machines, at least in many cases. In addition, it has a better chance of (c) being attractive to the intended users and (d) exploiting the strengths of the machine as well as those of the human since it is being developed and tested by professional translators at the request of a translation organization. Since it is so modest it should be (e) fairly clear what it can and cannot do. Finally, there is a (f) clear path plan toward a desirable long-term goal, since the strategy explicitly calls for more and more ambitious tools as time goes on.

6. Another Constructive Suggestion: Appeal to the End-User

The workstation approach is an attempt to appeal to the professional translators; it uses the benefits of office-automation as a way to sneak technology into the translator's workplace. An alternative approach, which also seems promising to us, is to use the speed advantages of raw (or almost raw) MT to appeal to the end-user who many not require high-quality.

6.1 Rapid Post-Editing

After noting the translators were unlikely to support the EPE application because they are unlikely to choose MT over HT, Wagner found that end-users would often opt for crummy quick-and-dirty translation, if they were given a choice.

> "We therefore decided to use Systran in a different way -- to provide a faster translation service for those translation users who wanted it, and were willing to accept lower-quality translation." (Wagner, 1985, p. 203)

The output from Systran was passed through a 'rapid post-editing' service that emphasized speed (4-5 pages per hour) over quality. When the project was first presented to the translation staff, it was well-received and 13 out of 35 volunteered to offer the rapid post-editing service on the understanding that they could opt out if they did not enjoy it. Wagner found that "the option is popular with a number of users and perhaps surprisingly, welcomed with some enthusiasm by CEC [Commission of the European Communities] translators who find rapid post-editing an interesting challenge" (Hutchins, 1986, p. 261).

Wagner's rapid post-editing service is a much better application of crummy MT than EPE because it gives all parties a choice. Both the users and the translators are more likely to accept the new technology, warts and all, if they are given the choice to go back and do things the old-fashioned way. The trick to being able to capitalize on the speed of raw MT is to persuade both the translators and the end-users to accept lower quality. Apparently the end-users are more easily convinced than the translators, and therefore, for this approach to fly, it is important that the end-users be in the position to choose between speed and quality.

6.2 No Post-Editing

The Georgetown system was used extensively at the EURATOM Research Center in Ispra, Italy, and the Atomic Energy Commission's Oak Ridge National Laboratory from 1963 until 1973. Translations were delivered without pre-editing or post-editing. In 1972-1973, Bozena Henisz-Dostert (now Bozena Thompson) conducted an evaluation and concluded that users were quite happy with raw MT:

> "The users presented a rather satisfied group of customers, since 96 percent of them had or would recommend machine-translation services to their colleagues, even though the texts were said to require almost twice as much time to read as original English texts (humanly-translated texts also were judged to take longer to read, but only about a third longer), and that machine-translated texts were said to be 21 percent unintelligible. In spite of slower service than desired and a high demand on reading time, machine translation was preferred to human translation by 87 percent of the respondents if the latter took three times as long as the former. The reasons for the preference were not only earlier access, but also the feelings that the 'machine is more honest', and that since human labor is not invested it is easy to discard a text which proves of marginal interest. Getting used to reading machine-translation style did not present a problem as evidenced by the answers of over 95 percent of the respondents." (Henisz-Dostert, 1979, p. 206)

It is also interesting to compare the attitudes of the users of this service the with attitudes of the translators mentioned above. Henisz-Dostert found that end-users were generally quite supportive, and would recommend the service to a friend, whereas Macklovitch found that professional translators were generally unwilling to continue using the service themselves, let alone recommend the service to a friend.

> "A grateful word is in order on the users' attitudes, who were most cooperative and friendly, and interested in what was involved in machine translation. They showed their familiarity with the aberrations of the texts, some of which were considered quite amusing 'classics', e.g., 'waterfalls' instead of 'cascades' (the users asked that this not be changed!). Very commonly, and understandably, they were interested in improvements and offered many suggestions. An example of an extreme attitude on the part of one user in this respect was that of 'cheating' on the questionnaire by giving less positive answers than in oral discussions. When subsequently asked about this, he reacted with something like: 'I use it so much, I want you to improve it, and if I show that I am satisfied, you will not work on it any more.'' (Henisz-Dostert, 1979, p. 151)

Why are these users so much more satisfied with MT than the translators involved in the Canadian government's trial of Weidner? We believe the difference is the application. It makes sense to offer end-users the option to trade off speed for quality, whereas it does not make sense to try to force translators to become post-editors. Consider the example of the crop varieties mentioned above. Many end-users might not be bothered too much by the extra articles because they can quickly skim past the mistakes, but the professional translator might feel quite differently about the extra articles because he or she will have to fix them.

6.3 Even More Modest Attempts to Appeal to the End-User

Consider, for example, the problem of reading email from other countries. The first author currently receives several messages a day in French such as the following:[8]

> Pour repondre aux questions de Maurizio LANA, j'ai entendu dire de bonnes choses concernant le programme ALPS de Alan MELBY. C'est au moins le nom de sa societe (ALPS) qui se trouve a Provo ou a Orem (Utah, USA). Il est egalement professeur de linguistique a la Brigham Young University (Provo, Utah).

It might be possible to provide a tool to help recipients whose French is not very good. Imagine that the

8. These messages usually arrive without accents.

email reader had a "Cliff-note" mode that would gloss many of the content words with an English equivalent:

Pour respondre aux questions de Maurizio LANA,
answer *questions*

j'ai entendu dire de bonnes choses concernant
heard *say* *good* *things* *concerning*

Cliff-note mode could be used as a way to sneak technology into the email reader, just as Kay's workstation approach is a way of sneaking technology into the translator's workplace. At first, Cliff-note mode would do little more than table lookup, but as time goes on, it might begin to look more and more like machine translation. In the future, for example, the system might be able to gloss the phrase *le nom de sa societe* as *the name of his company*, but currently the system would gloss *nom* as *behalf* (as in *au nom de*), and *societe* as *society*, because these senses happen to be more common in the Canadian Hansards (parliamentary debates), which were used to train the system. Obviously, the results would be much improved if we started with a more representative sample of general language, but nevertheless, even these results may be useful, at least for users whose French is sufficiently weak.

Cliff-note mode stands up fairly well to the six desiderata. (a) It sets reasonable expectations. (b) It doesn't cost much to run. (c) It ought to be attractive to users. After all, those who don't like it, don't have to use it. (d) It is well-positioned to integrate the strengths of the machine (vocabulary) without competing with the strengths of the user (knowledge of function words, syntax and domain constraints). (e) It is so simple that users shouldn't have any trouble appreciating both the strengths as well as the weaknesses of the word-for-word approach. Finally, (f) the strategy of gradually introducing more and more technology is ideally suited for advancing the field toward desirable long-term goals.

Perhaps, it may already be the case that the field can deliver much more than cliff-note mode. If so, after have bought into cliff-note mode, the marketplace would be well-positioned to appreciate these these improvements.

It may seem perverse to suggest that we should try to deliver much less than the state-of-the-art. However, in the near term, one probably cannot deliver a small, reliable, easy-to-use, inexpensive MT system with broad coverage that would be able to do much better than cliff-note mode. It is probably better to do something modest, than try to do too much and end up accomplishing too little.

7. Conclusion

We have identified six desiderata for a good niche application. Two marketing strategies appear to meet these six desiderata fairly well:

1. . use the benefits of office-automation to sell to the professional translator, or

2. use the speed advantages of raw (or almost raw) MT to sell to the end-user who many not require high-quality.[9]

9. Other possibilities have also been successful in the past. Xerox for example, has obtained impressive results by introducing a restricted language into the document preparation organization (Hutchins, 1986, p. 294). Smart Systems has also exploited the use of a restricted language in organizations that generate text. Limiting the domain is another formula for success. The classic example is Meteo (Isabelle, 1984). Unfortunately, however, it is very hard to find very many other naturally-occurring limited domains that people care about, and consequently, this strategy is unlikely to be repeated very many times in the future.

The discussion has stressed pragmatism throughout. The speech processing community, for example, has been somewhat more successful recently in making it possible to report crummy results. It is now quite acceptable in the speech community to work on very restricted domains (e.g., spoken digits, resource management (RM), airline traffic information systems (ATIS)) and to report performance that doesn't compare with what people can do. No one would even suggest that a machine should be able to recognize digits as well as a person could. Because the field has taken a more realistic approach, the field now has a fairly good public image, and is appearing to be making progress at a reasonable rate:

> "Slowly but surely, the technology is making its way into the real world." (Schwartz, 1991, *Business Week*, p. 130)

But there was a time when speech researchers were much more ambitious. According to Klatt's review (Klatt, 1977), the first ARPA Speech Understanding project (Newell *et al.*, 1973) had the objective of obtaining a breakthrough in speech understanding capability that could then be used toward the development of practical man-machine communication systems. Even though Harpy (Lowerre and Reddy, 1980) did in fact exceed the specific goals of the project (e.g., accept a thousand word-vocabulary connected-speech with an artificial syntax and semantics and produce less than 10% semantic error in a few times real time on a 100 mips machine), it didn't matter because Harpy had failed to obtain the anticipated breakthrough. And consequently, funding in speech recognition and understanding was dramatically reduced over the following decade. When activity was eventually resumed many years later, the community had learned that it is ok to strive toward realistic goals, and that it can be dangerous to talk about breakthroughs.

7.1 The GU Experiment

The experience in machine translation is perhaps even more sobering. The 1954 Georgetown University (GU) experiment was a classic example of a success catastrophe. In Zarechnak's 1979 review of early work on machine translation, he recalled that the GU experiment was originally seen as a huge advance:

> "The result of GU machine translation was given wide publicity in 1954 when it was announced in New York. The announcement was greeted by astonishment and skepticism among some people. L. E. Dostert summarized the result of the experiment as being an authentic machine translation which does not require pre-editing of the input nor post-editing of the output." (Zarechnak, 1979, p. 28)

But now, we can look back and see that the 1954 GU experiment probably did more harm than good by setting expectations at such an unrealistic level that they could probably never be met. Ten years after the GU experiment, the ALPAC report compared four then-current systems with the earlier GU experiment and suggested that there had not been much progress.

> "The reader will find it instructive to compare the samples above with the results obtained on simple, or selected, text 10 years earlier (the Georgetown-IBM Experiment, January 7, 1954) in that the earlier samples are more readable than the later ones." (Pierce *et al.*, 1966)

Zarechnak, a member of the Georgetown effort, complained rather bitterly that the comparison was unfair. In reality, the 1954 GU experiment had been a canned demo of the worst kind, whereas the four systems developed during the 1960s were intended to handle large quantities of previously unseen text.

> "When ten years later a text of one hundred thousand words was translated on a computer without being previously examined, one would expect a certain number of errors on all levels of operations, and the need for post-editing. The small text in 1954 has no such random data to translate." (Zarechnak, 1979, p. 56)

In fact, the ALPAC committee had also appreciated the "toy"-ish aspects of the 1954 GU experiment, but they did not feel that that was an adequate excuse. They criticized both the 1954 experiment as well as the four systems in question, the former for setting expectations unrealistically high, and the latter for failing to meet those expectations, unrealistic as they may be.

> "The development of the electronic digital computer quickly suggested that machine translation might be possible. The idea captured the imagination of scholars and administrators. The practical goal was simple: to go from machine-readable foreign technical text to useful English text, accurate, readable, and ultimately indistinguishable from text written by an American scientist. Early machine translations of simple or selected text, such as those given above, were as deceptively encouraging as 'machine translations' of general scientific text have been uniformly discouraging." (Pierce *et al.*, 1966, pp. 23-24)

If expectations had been properly managed and the waters had not been poisoned by the 1954 GU experiment, it is possible that we would now look back on the MT effort during the 1960s from a much more positive perspective. In fact, one of the four systems in question later became known as *Systran*, and is still in wide use today. In this sense, early work on MT was much more successful than early work on Speech Understanding; the first ARPA Speech Understanding Project did not produce any systems with the same longevity as Systran.

For some reason that is difficult to understand, the two fields currently have entirely different public images; on the one hand, the laymen can readily recognize that it is extremely difficult for a machine to recognize speech, while, on the other hand, even the manager of a translation service will blindly accept the most preposterous pretensions of practically any MT salesman. Perhaps we can change this perception if we succeed in focusing our attention on good applications of state-of-the-art (i.e., crummy) machine translation.

References

[1] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., Rossin, P. (1990), "A Statistical Approach to Machine Translation," *Computational Linguistics*, 16:2, pp. 79-85.

[2] Brown, P., Della Pietra, S., Della Pietra, V., Lai J., Mercer, R. (1991) "An Estimate of an Upper Bound for the Entropy of English," submitted to *Computational Linguistics*.

[3] Church, K. and Gale, W. (1991) "Concordances for Parallel Text" Seventh Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, Oxford, England.

[4] Henisz-Dostert, B. (1979) "Users' Evaluation of Machine Translation," in Bozena Henisz-Dostert, R. Ross Macdonald, Michael Zarechnak (eds), "Machine Translation," Mouton Publishers.

[5] Isabelle, P. (1984) "Machine Translation at the TAUM Group," in King, M. (ed.) *Machine Translation Today: The State of the Art*, Edinburgh University Press.

[6] Isabelle, P. and Bourbeau, L. (1985) "Taum-Aviation: Its Technical Features and Some Experimental Results," *Computational Linguistics*, 11:1, pp. 18-27.

[7] Johnson, R., King, M., and des Tombe, L. (1985) "Eurotra: A Multilingual System under Development, *Computational Linguistics*, 11:2-3, pp. 155-169.

[8] Kay, M. (1980) "The Proper Place of Men and Machines in Language Translation," unpublished ms., Xerox, Palo Alto, CA.

[9] King, M. and Falkedal, K. (1990) "Using Test Suites in Evaluation of Machine Translation Systems," COLING, pp. 211-216.

[10] Klatt, D. (1977) "Review of the ARPA Speech Understanding Project," *Journal of the Acoustical Society of America*, reprinted in Waibel, A. and Lee, K. (eds.) (1990) *Readings in Speech Recognition*, Morgan Kaufmann Publishers, Inc., San Mateo, California.

[11] Lawson, V. (1984) "Users of Machine Translation System Report Increased Output," *Language Monthly*, 11, pp. 6-10.

[12] Lehrberger, J. and Bourbeau, L. (1988) *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation*, John Benjamins Press.

[13] Lowerre, B. and Reddy, D. (1980) "The Harpy Speech Understanding System," in *Trends in Speech Recognition*, Prentice Hall, reprinted in Waibel, A. and Lee, K. (eds.) (1990) *Readings in Speech Recognition*, Morgan Kaufmann Publishers, Inc., San Mateo, California.

[14] Macklovitch, E. (1986) "MT Trial and Errors," presented at the International Conference on Machine and Machine-Aided Translation, April 7-9, Aston University, Birmingham, United Kingdom.

[15] Macklovitch, E. (1989) "An Off-the-Shelf Workstation for Translators," *Proceedings of the 30th American Translators Conference*, Washington DC, 1989.

[16] Macklovitch, E. (1991) "Evaluating Commercial MT Systems," paper presented at the Evaluators' Forum on MT systems, organized by ISSCO at Ste. Croix, Switzerland.

[17] Magnusson-Murray, U. (1985) "Operational Experience of a Machine Translation Service," in Lawson, V. (ed.) *Tools for the Trade, Translating and the Computer 5*, Alden Press, Oxford.

[18] Nagao, M. (1987) "Role of Structural Transformation in a Machine Translation System," in Nirenburg, S. (ed.) *Machine Translation: Theoretical and Methodological Issues*, Cambridge University Press, pp. 262-277.

[19] Nagao, M., Tsujii, J-I., and Nakamura, J-I. (1986) "Machine Translation from Japanese to English" *Proceedings of the IEEE*, 74:7, pp. 993-1012.

[20] Newell, A., Barnett, J., Forgie, J., Green, C., Klatt, D., Licklider, J., Munson, J., Reddy, D., and Woods, W. (1973) *Speech Understanding Systems: Final Report of a Study Group*, North-Holland/American Elsevier, Amsterdam.

[21] Pierce, J., Caroll, J., Hamp, E., Hays, D., Hockett, C., Oettinger, A., Perlis, A. (1966), "Language and Machines: Computers in Translation and Linguistics," also known as the ALPAC report, National Academy of Sciences Publication 416, Washington D.C.

[22] Schwartz, E. (1991) "A Computer that Recognizes its Master's Voice," *Business Week*, June 3, pp. 130-131.

[23] Shannon, C. (1951) "Prediction and Entropy of Printed English," *Bell Systems Technical Journal*, vol. 30, pp. 50-64.

[24] Tschira, K. (1985) "Looking Back at a Year of German-English MT with Logos," in Lawson, V. (ed.) *Tools for the Trade, Translating and the Computer 5*, Alden Press, Oxford.

[25] Tucker, A.B. (1987) "Current Strategies in Machine Translation Research and Development," in Nirenburg, S. (ed.) *Machine Translation: Theoretical and Methodological Issues*, Cambridge University Press, pp. 22-41.

[26] Van Sylpe, G. (1979) "Evaluation of the f1978 Version of the SYSTRAN English-French Automatic System of the Commission of the European Communities, *The Incorporated Linguist*, 18:3, pp. 86-89.

[27] Wagner, E. (1985) "Rapid Post-Editing of Systran," in Lawson, V. (ed.) *Tools for the Trade, Translating and the Computer 5*, Alden Press, Oxford.

[28] Zarechnak (1979) "This History of Machine Translation," in Bozena Henisz-Dostert, R. Ross Macdonald, Michael Zarechnak (eds), "Machine Translation," Mouton Publishers