A First-Pass Approach for Evaluating Machine Translation Systems

Pamela W. Jordan Bonnie J. Dorr*
John W. Benoit
Mitre Corporation
C^{3}I Artificial Intelligence Center
7525 Colshire Drive
McLean, Virginia 22102
(703)883-6685
and
University of Maryland*
UMIACS
A.V. Williams Building
College Park, Maryland 20742
(301)405-6768

Abstract

This paper describes a short-term survey and evaluation project that covered a large number of machine translation products and research. We discuss our evaluation approach and address certain issues and implications relevant to our findings. We represented a variety of potential users of MT systems and were faced with the task of identifying which systems would best help them solve their translation problems.

1 Introduction

During 1991, The MITRE Corp.[1] surveyed and evaluated machine translation (MT) systems across the U.S. and, to a lesser extent, in Europe and Japan. The intent of the study was to recommend software purchases and R&D support that would address the near-term,

[1]MITRE is an independent, not-for-profit organization that provides technical assistance, systems engineering, and acquisition support to U.S. government agencies.

medium-term and long-term translation needs of the users we represented. We found we had three types of potential MT users: those who needed to scan material to estimate its relevance, those who wanted to know the content of the material, and those who wanted publication-quality translations.

Initially we identified over 20 MT efforts in the U.S. alone that we should investigate. This included systems in operational use, systems still in development, and both academic and commercial research systems. Since it is too costly to do in-depth evaluations of so many MT efforts, we decided to gather just enough information to limit the field to the best systems for addressing a particular user's translation problems. What is described here is our filter approach for narrowing down the possibilities and an assessment of its success. Deeper evaluations can now be done on this smaller set but planning and conducting the in-depth evaluations will take place at a later time.

The requirements upon which we based our evaluation criteria were the following: (1) whether the MT systems and research projects provided the necessary functionality; (2) whether the parent organization of the vendor or research group was stable enough financially so that we could reasonably expect them to continue their work and support the user; (3) whether the system would be a good fit for the user's current and future concept of operations; (4) whether the system could be upgraded and maintained at reasonable costs; and (5) whether the system performed well enough to increase user throughput. Figure 1 shows the mapping of these five broad requirements categories to the evaluation criteria we selected.

2 Approach to Evaluation

Although we support the basic idea of black-box and glass-box evaluation [GF88] that is being pursued for NLP systems [PF90], this survey was of such a short time frame that test suites could not be built [KF90] nor customized tests performed. Our approach to evaluation (given the time restrictions) was to interview developers, researchers and current users of MT, participate in MT demonstrations, survey the literature for additional details about the software, and collect (for further evaluation) sample inputs and outputs for each language handled by the software. Detailed questionnaires were developed to guide the information collection process. The information collected included both glass-box and black-box types of data.

The next three subsections are a description of the type of information we collected from vendors and researchers as it relates to the five requirements categories.

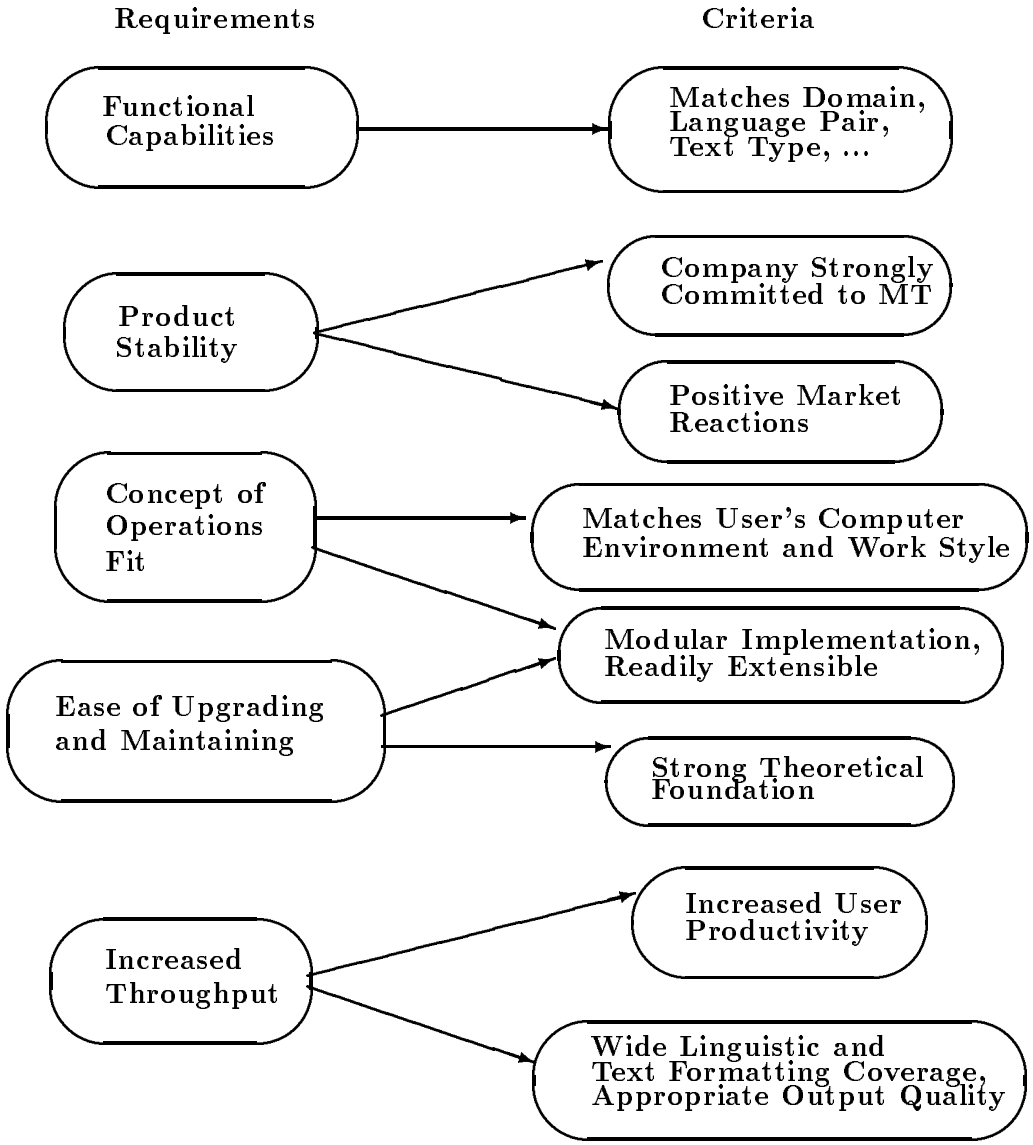

Figure 1: Mapping Categories of Requirements to Evaluation Criteria

2.1 Environment-Dependent Considerations

The first three requirements categories depend on the environment in which the user works. Users have particular language pairs, text types, and domains they need to have translated and this is what we mean by functional capabilities requirements. Developers of MT systems target their systems for particular values along these three dimensions in order to make the translation problem more approachable. If a system exists that matches the user's needs along these dimensions then there is a good chance that system will be a cost-effective near-term solution. A system that is not an exact match could be a good medium-term solution if the system can be extended cost-effectively or it is the only one available in the near future.

The next requirements category emphasizes the importance of the system support that cannot be obtained from the user's own organization and therefore must be obtained from the developer. This support could include maintaining and extending the lexicon and the grammar as well as general system maintenance. For this reason, it was useful to gauge the business health of the developer. We did this by determining the number of systems the developer had sold and the level of customer satisfaction.

The final environment-dependent requirements category relates to the user's computer environment and the user's mode of operation. We call this the concept of operations fit. If the system resides on hardware and interfaces with software that the user already has, then the system will be a more cost-effective near-term solution. If the system can be easily ported to the user's computer environment, or the system's native computer environment is inexpensive and integrable into the user's environment then this will make the system a possible medium-term solution. As for the user's mode of operation, a system that already matches the user's concept of operations (e.g., one that provides optical character recognition) will be a more cost-effective near-term solution. However, a system that is not a perfect match may be a cost-effective medium-term solution if it can be altered with minimal disruption to the core system (e.g., the linguistic knowledge is easily updated or the user interface is easily replaceable).

2.2 Ease of Upgrading and Maintaining

It was not feasible in the time available to formally test the ease of upgrading and maintaining. Instead, we examined each system's architecture and the theoretical foundations upon which it was based. A well-implemented MT system should be designed as a shell that can be readily customized for new domains, language pairs and text types.

We inquired about customization tools since this eases the process of extending the system. In addition to customization tools, we checked whether the core language knowledge was independent from the domain dependent knowledge. In a well-designed system, the core words of a language should not have to be re-entered when the system is customized for a new domain. To determine if the system could reasonably be integrated with other software, we asked if a programmer's interface had been developed and documented for the system.

Knowing something about the theoretical basis of a system provides additional insights into the facility with which it could be extended to handle new languages, domains and text types. Past experience in software engineering tells us that ad-hoc systems that are not based on some coherent theoretical foundation are difficult to extend and maintain. Additionally, knowing the theoretical foundations of the research is a major factor in predicting whether the work is promising for meeting long-term needs.

2.3 Increased Throughput

Predicting whether an MT system will increase translation throughput is a complex problem that depends not only on the quality of the translation and the intended use of the output (competence) but also on the speed with which the system produces a translation (performance).

2.3.1 Competence

To estimate a system's linguistic coverage as part of judging its competence, we used a fairly comprehensive checklist of linguistic and textual phenomena and asked the developers which of these phenomena their system handles. Presumably a wider coverage means that the system could produce a higher-quality output.

Another factor in judging the system competence was whether the system would produce an output of high enough quality to serve the user's purpose. For example, the MT quality required for a user who wants to produce publication-quality text should be such that he would be inclined to post-edit the MT results instead of doing his own translation directly from the source text. To determine whether a system's raw output would be acceptable to at least one of our three classes of users (see Section 1), we evaluated the English output of the operational systems. The source texts used were independently selected texts that corresponded to the language pairs, domains and text types that the system reportedly handled. Whenever possible, the machine translation of this source text was performed in our presence. By being present during the machine translation, we were able to observe

the modifications that had to be made in order to obtain the output that we evaluated. The modifications that were made (e.g., pre-editing, post-editing, lexical changes, and additions) gave us more insight into the linguistic coverage of the system. We collected 8 samples from 4 of the 10 operational systems we evaluated. For the other systems, an appropriate sample was not located before our visits.

We performed two types of fidelity tests on 6 of the outputs in order to predict their acceptability to our three types of users. The fidelity tests determined whether the meaning of the text was retained in the translation (semantic invariance [CCG81]). Users who wanted to know just the subject area, or who needed just the content of the text, are primarily concerned with semantic invariance; anything beyond this is a secondary consideration. If grammatical errors do not prevent such users from understanding the output, then they will be satisfied with its quality. On the other hand, users who want publication-quality output are concerned about stylistic and grammatical well-formedness as well as semantic invariance.

For both fidelity tests, we used 3 evaluators who were blind to which system produced the output and to what the source language was. All three evaluators were project members with technical backgrounds but none were experts in the subjects covered by the sample texts. In the first fidelity test, the evaluators examined only the raw MT output and then were asked to state the subject matter of the text. This test predicted whether the MT output would be acceptable to those who are scanning for texts in particular subject areas.

In the second fidelity test, the evaluators compared the raw MT output to a human translation (which was assumed to be correct) and rated how well the meaning of the original text was preserved. To rate the semantic invariance of the MT output, we provided the evaluators with a scale to keep the ratings consistent across languages, domains, and evaluators. We used Nagao's seven point scale (see Table 1) for judging accuracy or fidelity [N+85] (also like Van Slype's [vS82] measures of information transfer). We formed two hypotheses as to what the rating scale would mean for our three types of users:

Hypothesis 1 A rating in the range 1-4 would be suitable for users who either wanted to know the content of the material or those who wanted to post-edit and produce publication-quality output. The cutoff was at 4 because a rating of 5 indicated that the meaning was not conveyed adequately.

Hypothesis 2 A rating in the range of 1-5 would be suitable for users who needed to scan material for relevance. The cutoff was at 5 because a rating of 6 indicated that the meaning was not conveyed at all.

Nagao	Clarity of Meaning	Well-Formedness
1. Content of sentence conveyed. Needs no rewriting	1. Meaning clear, needs no rewriting	0. No syntactic errors
2. Content of sentence conveyed, needs some rewriting	2. Meaning clear, needs rewriting	1. Minor corrections needed
3. Content of sentence conveyed, but word order errors	2. Meaning clear, needs rewriting	2. Word order errors
4. Content of sentence generally conveyed, but attachment, tense, and number errors	2. Meaning clear, needs rewriting	3. Attachment, tense, and number errors
5. Content not adequately conveyed, expressions missing, and attachment problems	3. Meaning not clear but can guess	3. Attachment, tense, and number errors
6. Content not conveyed, clauses and phrases missing	4. Totally lost as to meaning	4. Phrases and clauses missing
7. Content not conveyed, subjects and predicates missing	4. Totally lost as to meaning	5. Subjects and predicates missing

Table 1: Evaluation Scales for Rating Fidelity

We separated Nagao's scale into two, one for rating clarity of meaning and the other for rating well-formedness (Table 1). First, the evaluators rated the clarity and wrote a paraphrase for each sentence in the MT output. Next the evaluators rated the well-formedness of the raw translation by comparing it to the correct translation. They also scored whether their paraphrase was right, nearly right, or wrong according to the correct translation. By examining clarity of meaning before well-formedness, the evaluators' judgements were not influenced by the correct translation. We later combined the two ratings so that we had one rating per sentence. Since the scale is for individual sentences, the ratings for each sentence were combined by taking the average.

2.3.2 Performance

Since it is generally accepted that publication-quality translations cannot be automatically generated by current MT systems without some form of human assistance, the question is whether the combined ef-

	M(8)	E(9)	LC(7)	FC(10)	F(10)	CS(6)	S(9)	C(10)
Excellent	12.5%	0%	14%	0%	10%	16.67%	0%	0%
Good	12.5%	22%	29%	30%	30%	50%	45%	40%
Average	62.5%	33%	14%	20%	10%	16.67%	22%	60%
Poor	12.5%	45%	43%	50%	50%	16.67%	33%	0%

Table 2: Evaluation Results for Systems Currently in Use (10 systems)

forts of the post-editor and the MT system are more productive than the translator (and to be fair, any other MAT tools that the translator chooses) without the MT system. To answer this question timings would have to be made in a well-controlled environment, but again this was too expensive to do with the number of systems we evaluated. Instead, we relied on any productivity measures independent users were able to provide. Unfortunately, these were often just rough estimates of translator productivity increases.

3 Findings and Assessment of the Evaluation Approach

3.1 Findings

We distilled the collected information into a rating for each of the criteria. Since there were so many systems and all of the evaluators were not present for all of the interviews, we had to normalize the ratings relative to 3 classes of MT systems: operational systems, systems in development, and research systems. Tables 2 and 3 show the ratings for the operational systems and the systems in development, respectively.[2] We did not include this information on research systems because it was of little practical value in evaluating the systems. We were not able to rate every system for every criterion because the information we needed was sometimes unavailable.

The evaluation of MT outputs served as a test for the two hypotheses proposed in section 2.3.1. Regarding hypothesis 1, we were unable to obtain data relevant to post-editing, but we arrived at some potentially informative results with respect to the suitability for understanding the content. Table 4 shows the correlation between the Nagao rating and

[2]For brevity, the following abbreviations are used: M = Modularity, E = Extensibility, LC = Linguistic Coverage, FC = Formatting Coverage, F = Friendliness, CS = Customer Satisfaction (applicable to table 2 only), S = Stability of Co., and C = Concept of Operations Fit. The numbers in parentheses are the number of systems for which we obtained this information.

	M(5)	E(6)	LC(4)	FC(6)	F(6)	S(6)	C(6)
Excellent	60%	0%	0%	0%	33%	0%	0%
Good	20%	66.67%	50%	50%	50%	67%	33%
Average	0%	16.67%	50%	0%	0%	33%	67%
Poor	20%	16.67%	0%	50%	17%	0%	0%

Table 3: Evaluation Results for Systems in Development (6 systems)

the correctness of the evaluator's paraphrases. If we take the correctness of the paraphrases as an indication of how well the sentences were understood, we see that at least 80% of the sentences with a rating of 4 or higher were understood. Thus, there is a systematic relation between the assignment of ratings from Nagao's system and the degree to which individual sentences were understood. But a question still remains: what percentage of sentences must be understood in order for the meaning of the text as a whole to be adequately conveyed? If the answer is 100% then the rating must be in the range of 1-2 in order to fully understand the content of the text. However, it is not clear that even 100% is the appropriate percentage: if we look just at the clarity rating in Table 5, we see that 8% of the time, the evaluators thought they understood the sentence when they actually did not. This raises a question we do not have an answer for: what is the user's tolerance level for being misled?

Regarding hypothesis 2, the evaluation results indicated that the Nagao fidelity rating was not a predictor of which outputs were suitable for scanning purposes. The subject area of a translation that received a rating of 7, was still correctly identified.

3.2 Assessment

Users are certainly capable of evaluating MT output without additional technical help but this is rarely the case when evaluating the system's engineering aspects. The systems we considered to be well-designed and based on strong theoretical foundations, all produced high quality outputs but not all high quality outputs were produced by such systems. Understandably, output quality is not a good predictor of system maintainability and extensibility.

We found that the information we collected was adequate for doing a first-pass evaluation of the operational and developing systems. However, our approach was of little use in evaluating research systems. The prototypes are typically too narrowly focused to translate a complete text and, although the researchers were able to answer many of

Fidelity Rating	**Correctness of Evaluator's Paraphrases**		
	Right	Nearly Right	Wrong
1	0%	0%	0%
2	82%	18%	0%
3	56%	25%	19%
4	27%	53%	20%
5	5%	63%	32%
6	0%	50%	50%
7	0%	4%	96%

Table 4: Correlation Between Average Nagao Rating per Sentence and Correctness of Each Evaluators' Paraphrases per Sentence

Evaluator's Rating of Understanding	**Correctness of Evaluator's Paraphrases**		
	Right	Nearly Right	Wrong
Meaning Clear	79%	21%	0%
Meaning Clear, needs rewriting	56%	36%	8%
Meaning Not Clear but can guess	12%	45%	43%
Meaning Unclear	0%	32%	68%

Table 5: Correlation Between Clarity Rating and Correctness of Evaluators' Paraphrases Compared to an Acceptable Translation

our questions, many of the questions were not useful in determining whether the research work was promising enough to support.

At the beginning of the survey, we were initially concerned that the subjects of the MT interviews would not be willing to spend as much time as it took to go through our lengthy questionnaire. Fortunately, a large majority of the groups were extremely cooperative. The biggest difficulty we encountered was in finding out enough about the user's requirements to make an evaluation possible. Users cannot easily tell us what makes a translation acceptable to them and this is one of the key elements in evaluating an operational MT system.

References

[CCG81] Jaime G. Carbonell, Richard E. Cullingford, and Anatole V. Gershman. Steps toward knowledge-based machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-3(4), 1981.

[GF88] Tom Gilb and Susannah Finzi. *Principles of software engineering management.* Addison-Wesley Pub. Co., Reading, Mass., 1988.

[KF90] M. King and K. Falkedal. Using test suites in evaluation of machine translation systems. In *COLING 90*, volume 2, pages 211 – 216, 1990.

[N+85] Makoto Nagao et al. The Japanese government project for machine translation. *Computational Linguistics*, 11, April-September 1985.

[PF90] M. Palmer and T. Finin. Workshop on the evaluation of natural language processing systems. *Computational Linguistics*, 16(3), 1990.

[vS82] G. van Slype. Conception d'une méthodologie générale d'évaluation de la traduction automatique. *Multilingua*, 1(4):221 – 237, 1982.