[*Translating and the Computer 19*: Papers from the Aslib conference… 13 & 14 November 1997]

MT EVALUATION: SCIENCE OR ART?

Derek Lewis
School of Modern Languages, University of Exeter, Devon, UK

Some developments in the current state of MT evaluation are reviewed. Factors in the assessment of linguistic performance of MT systems and also types of evaluation tools are briefly described. Experiences of implementing MT within a university curriculum are presented. Finally, the paper discusses samples of output from a general purpose test suite, used in a teaching environment to illustrate the problems of evaluating MT output from various PC systems.

The Current State of MT Evaluation

The famous 1966 ALPAC Report, for which evaluators rated output MT on scales on scales of speed, cost and quality, was one of the first major comparative evaluations of MT systems. John Lehrberger's guidelines for evaluation, compiled in 1981, took account of output quality, cost, time, and system improvability (Lehrberger and Bourbeau (11)). In the early 1990s the Essex University MT evaluation group discussed various methodologies, tools and approaches, which appeared in a number of reports. An indication of the stage evaluation has now reached is the 1994 draft report by the EAGLES Subgroup on the evaluation of natural language processing systems. The report does not focus on MT, which is considered as a translation aid alongside multilingual dictionaries and thesauri, terminology management systems and translation memories. Indeed the only specific MT product mentioned is ALPnet, chosen because it was at the time (March 1994) the only package that was both available commercially and familiar to the report's compilers.

The report does not aim to supply a league table detailing the advantages and disadvantages of specific products. Instead it lists criteria for evaluating certain types of translation aids. It defines what is involved in the evaluation process and sets up feature checklists of translators' tools according to what it calls the 'consumer report paradigm', a set of specific questions asked of the software which may be answered by yes/no or by values on quantifiable scales. The result is a set of worked out

checklists which corresponds to an approach suggested by Steven Krauwer (10). Although the 1994 report devotes little attention to the specifics of MT evaluation (which is the object of attention of a different subgroup), it indicates the position of MT within an increased range of electronic tools for translations and shows how evaluation has become increasingly orientated to benchmark measures of users' requirements.

The report pays particular attention to user profiles as a factor in evaluation. These include the quantity and quality of translation work required, the nature of texts to be translated, and the characteristics of the organisation in which translation is carried out (e.g. its policy on language, or the number of languages it uses in-house). Various developments in the translation industry are noted. Firstly, the number of languages in which translations are required is increasing; at the same time certain languages are emerging as universal focal languages (e.g. a text is produced directly in English, and then translated into, say, Finnish and Japanese). Secondly, source texts are becoming more repetitive (mass produced manuals are a typical example); at the same time translation involves more revision, updating, and layouting than before and it is proving difficult to keep up with burgeoning terminology. Thirdly, organisations are tending either to outsource translation or to turn their translation departments into independent business units. The conclusion is that there is indeed no such thing as a typical user profile; the main determinants are an organisation's resources and its policy towards translation and languages.

What effect do these factors have on MT evaluation? Since MT is more often found in large, well resourced organisations, smaller companies and freelance translators are still less likely to use MT or to have the resources to evaluate, integrate, and customise MT to their needs. A sample of users and organisations undertaken by the EAGLES subgroup indicated that usage of high end electronic tools such as memories, archives and dictionaries is currently low, although increasing. MT is seen to be at the very top of the scale of technical sophistication and is either used or being considered by large translation companies or concerns with internal translation units. The most likely environment for any translation help tool is an integrated document production and translation facility, a computer network, a high level of computer literacy, and a positive framework for exploring new systems; MT remains a sophisticated translation aid demanding significant resources for development and fine-tuning. The report still sees MT as the province of large scale users with extensive IT and human resources. At the same time it recognises the emergence of a market for small users as systems become more compact and affordable. While larger companies may increasingly outsource translations to freelance translators, these are less likely to have technical aids such as MT; very occasionally companies will make resources (such as standardised term banks and in-house dictionaries) available to the contracted translator in order to ensure consistency in the product.

but cannot ensure that a particular linguistic phenomenon is tested. The TSNLP (Test Suites for Natural Language Processing) project reported in 1994 that it was in the process of developing methods of automatically constructing test suites for NLP applications such as MT (Balkan (2), Balkan et al (3)). There are a number of issues in constructing and using test suites for evaluating MT.

Firstly, test suites are generally regarded as diagnostic tools for system developers but they can also be used to assess adequacy of output, which is what users are interested in. To do this, however, test suites need to take into account the frequency of occurrence of a particular phenomenon and should include a variety of lexical, rather than syntactic, information. The data in a test suite may need to be tailored to an application, such as MT, or even to a domain or text type within MT itself. Tools may be developed to generate suites that test particular phrase types, combinations of structures, and even domain-specific vocabulary; the data is held in a database and the suite generated according to criteria specified by the user (see Nerbonne et al (13)). It has also been suggested that structures in test suites should be matched with structures found in corpora so that the contents of the suite reflect real-life texts. Finally, it is clear that general-purpose monolingual test suites are only of limited use in evaluating a particular MT system in a particular operational environment: they need to be adapted for user's needs and, more generally, for language pairs. As a very simple example, a test suite for English-to-German MT would have to test for nouns whose gender is known to vary in German; the gender of a noun will also determine pronominal constructions elsewhere in the sentence, including, for instance, relative pronouns introducing sub-clauses, as in the following:

English: the table *which* is in the corner . . .
the lamp *which* is in the corner ...
the television *which* is in the corner

German: der Tisch, *der* in der Ecke steht ...
die Lampe, *die* in der Ecke steht ...
das Fernsehen, *das* in der Ecke steht ...

Constructing test suites for MT users is an expensive and complex task. Evaluating the results of test suites is likewise beyond the resources of most users (for references to developed test suites see Nerbonne et al (13) and Way (17)).

Using MT in a University Curriculum: General Issues

This section addresses the question of how MT can be introduced in a university teaching curriculum. It is based on experiences of teaching students in a natural language processing applications module for undergraduate students of foreign languages at the University of Exeter. Although the need for language undergraduates to be familiar with the technological

tools of translation has been highlighted (Clark (4)), a recent survey on tools and techniques for MT conducted by the University of Essex concluded that the time was not right to pursue any coordinated initiative to promote or develop MT teaching; the reasons included lack of interest on the part of suppliers, high cost, and the large amount of work involved. Despite such obstacles the integration of MT remains a worthwhile goal for a university modern languages programme. For a university student the central task should be to demonstrate that he understands the parameters of MT evaluation and can apply these in practice. For a future employer the student must show that he can apply those parameters to MT system in general, not just the package used on the course. On the other hand a university student is not in the position of being able to evaluate MT in a genuinely professional or practical context, i.e. for a company or organisation. He is not handling translation as part of a document production or handling process and does not have access to, say, large volumes of technical texts. Neither is he a professional translator with experience of translating domain-specific texts into his mother tongue. Most UK undergraduates translate into the target language as part of a language learning activity, with all the differences which this approach entails. A further constraint is time: a student has to complete a module in a limited period and alongside demands made by other areas of the course; in a culture-based combined honours modern languages curriculum MT is unlikely to enjoy a high priority. It must also be noted that MT is new to students: unlike other software tools, they have almost certainly never encountered MT before. As a result they are likely to approach it with great curiosity but also with the expectations inherent in the title: a computer program which does the kind of translation that a human being performs, only fully automatically.

This may sound like a long list of reasons for not expecting students to evaluate MT. On the other hand students are not unlike many potential users of MT (who, as noted above, are unlikely to exhibit a standard profile). As a future employee with foreign language expertise in a company (whether or not a full-time translator), a student may well be expected to 'look at' or 'advise on' the wisdom of investing in MT and to produce a recommendation on a fairly informal basis; this, of course, may change when tools for professional evaluation of MT eventually become available. At the moment, however, the question is what are the most appropriate methods for student-based MT evaluation within the constraints outlined?

The first method is for students to evaluate a particular MT package using samples of different text types (e.g. journalistic, literary, technical). The evaluation can be informal (i.e. based on general impressions on the quality of output), or it can be based on surveys of fellow students' assessments, using scaled measures of intelligibility, accuracy, etc. The samples are usually small and manageable; most students in fact rely on their personal, i.e. intuitive, assessment of output. Another approach is for students to evaluate a system using part of a

public machines update and maintain their own dictionaries (as they must do for assessment and evaluation purposes): most MT systems are not designed to be installed and used in this way.

Finally, one can raise the issue of how much linguistic knowledge is required by users in order properly to evaluate errors in output. Leaving aside matters of intelligibility and accuracy, it is clear that a clear knowledge of syntactic categories and phrase structures is a prerequisite; it is also important to be able to bring this knowledge to bear on MT output. As an example, consider the following example, taken from an article on fashion:

SL English phrase: ... time to hold off with the damsons and deep chocolates of last season.

Raw MT German translation: ... Zeit, um mit dem damsons und tiefen Schokoladen letzter Jahreszeit.

User's comment: 'This text has been translated fairly well. Even the genitive has been correctly employed. This is due to the shortness of the sentence. An adjective has been translated as a noun: chocolates. However, this is also unclear in English. It would be necessary to define it with the addition of a noun.'

What the user/student really should be saying is something like: '(a) the prepositional phrase (..) attached to the noun phrase (...) has been correctly translated, with the English preposition rendered by a German genitive; (b) the English noun 'chocolates' is, unusually, used as a noun referring to colour (here: shades of chocolate); this is because the text domain is fashion. The question is how the dictionary could be modified to handle this feature (which also occurs in the use of 'damsons').'

Assessing Text Suite Output

The following section contains samples of raw MT output of an extract from the HP test suite of (minimally) annotated sentences. The output of three systems (referred to S1, S2 and S3) is shown. The systems are relatively low-cost, commercially available packages for PC: S1 is the EASY TRANSLATOR system for translating on-line the contents of Web pages or the contents of a Windows clip-board; S2 is Langenscheidt's T1 system (version 3.0), and S3 is Globalink's POWER TRANSLATOR (version 2.0). The samples illustrate some of the problems of evaluating MT output, especially the relative merits of different systems, using a general purpose test suite. The language direction is English to German. In conclusion, the appropriateness of using text suite output in teaching MT is briefly discussed.

Firstly, consider the translation of examples of English restrictive relative clauses (wh-type clauses):

important, as in the German versions for 'approve of', where the choice of 'zustimmen' (among other things) makes the sense harder to understand. But it would be wrong, at least for language that is not domain-specific, to draw from isolated examples conclusions about the adequacy of the lexical choices made by a system as a whole. Such conclusions could only be drawn from test suites devised for particular domains of vocabulary. At the same time, syntax and vocabulary cannot be artificially separated. As seen in German separable verb prefixes, the lexical choice of a verb in the source language may trigger a particular syntactic feature in the target language; a single general purpose test suite may not be all that useful for evaluating MT output in different languages.

It is tempting but unwise to draw broad inferences about intelligibility from test suite output. Consider, for instance, sentence 11 above ('Abrams has an office (that) Browne showed Chiang', where the subject relative pronoun 'that' is deleted). Only the S3 reintroduces the relative in German, where is it obligatory: both S2 and S1 omit it, resulting in a significant loss of intelligibility. We could infer from this that S3's output for this feature will be consistently more intelligible. We would, however, have to be certain that the feature was translated similarly over similar constructions. Intelligibility is better assessed as a feature of texts rather than individual sentences.

In the following example, S1 is possibly less intelligible, but only because it has departed from the SL word order (often required when translating from English into German). In other structures, the tendency to stick to SL word order might hinder intelligibility. It is difficult to tell.

ENGLISH: Abrams has an office Browne showed Chiang.
S1: Abrams hat ein Büro Browne Chiang hat gezeigt.
S2: Abrams hat ein Büro Browne zeigte Chiang.

Test suites are especially good at revealing precise information about structures in syntactic combination. A simple example is the use of 'mass/mass-creating partitives' in subject NPs:

ENGLISH: Most of the staff is competent.
Most of the program works.
Almost all of the program works.

S1: Am meisten ist vom Personal fähig.
Am meisten von den Programmarbeiten.
Fast alle Programmarbeiten.

S2: Meiste vom Personal ist fähig.
Meiste vom Programm funktioniert.
Fast alle das Programm funktioniert.

S3: Der meiste des Stabes ist qualifiziert.
Die meisten der Programm-Werke.
Fast all die Programm-Werke.

The partitives are translated tolerably well. But the outputs suggest that S1 and S3 are most likely to have problems disambiguating the English plural noun and 3rd person singular present tense verb forms (here: 'works') when the partitive occurs in a subject NP. Since the suite is annotated, we know precisely what linguistic structure is being input and tested. As an MT user or developer, however, we might also like to know what output structure the MT system thinks it has produced. But only the developer might be able to tell the MT system to annotate the output; the user typically has only the output sentences to go on, from which it is often difficult to judge what the MT system thinks it has produced. Operational systems, moreover, do not necessarily operate with clear linguistic models through to the final output generation stage; there may be no abstract structure or 'linguistic annotation' at all that we can attach to the output.

Consider the following output from subject NPs containing the apostrophe marker for possessives.

ENGLISH: The project's engineers work for Abrams.
Abrams's engineers were interviewed by Browne.
Abrams' engineers were interviewed by Browne.

S1: Die Ingenieure des Projekts arbeiten für Abrams.
Ingenieure Abrams wurden von Browne interviewt.
Abrams wurden' Ingenieure von Browne interviewt.

S2: Die Ingenieure des Projekts arbeiten bei Abrams.
Daß Ingenieure von Browne geinterviewt wurden, Abrams ist.
Die Ingenieure von Abrams wurden von Browne geinterviewt.

S3: Des Projektes Ingenieure arbeiten für Abrams.
Abrams's Ingenieure wurden von Browne interviewt.
Abrams' Ingenieure wurden von Browne interviewt.

S2 has succeeded in transposing the Nounl's + Noun2 construction into Noun2 + von + Nounl, which indicates a superior transfer capability. At the same time, where Nounl ends in 's' (as in 'Abrams's), S2's output is catastrophically garbled. This 3-sentence sample exemplifies an important dilemma in MT evaluation: how do we rate overall different systems in which 2/3 of the output of one system is very good and the remaining 1/3 very poor, while 3/3 of another system are neither very good nor very poor?

Finally, even relatively short test sentences can be syntactically complex and produce compounded errors which are difficult to measure except in general terms of intelligibility

and comprehensibility. In the following sample it is hard to explain why S2 has gone wrong, but first impressions about the sentence are clear and immediate. S1 has added a relative 'dass' which aids clarity; S2 misparses the subject - verb construction in the main clause and misconstructs a past tense form (so does S1 incidentally, but not so badly); S2 compounds the syntactic failure with poor lexical choice (mieten).

ENGLISH: Abrams hired a woman who Browne knew Chiang interviewed.
S1: Abrams hat eine Frau angestellt, die Browne gewußt hat, daß Chiang interviewt hat.
S2: Abrams mieten eine Frau die Browne kannte Chiang geinterviewte.

In conclusion it would appear that MT evaluation has made serious efforts to become a science, with different modes of evaluation for different purposes and the prospect of clearer benchmark criteria for users. On the other hand, the limited resources available in higher education for teaching MT suggest that, while evaluation should be an essential component of MT-based courses, students are more likely to benefit from a discovery-based approach using less rigorous techniques based on relatively small volumes of text and sections of text suites, as illustrated above. The approach may be well correspond to how low cost MT systems are likely to be evaluated by users in the market.

REFERENCES

1. Arnold, D., et al., 1994, Machine Translation. An Introductory Guide. NCC Blackwell, Oxford.

2. Balkan, L., 1994, "Test suites: Some issues in their use and design", Machine Translation Ten Years On, Conference at the University of Cranfield, 12-14 November 1994, 26-1.

3. Balkan, L., et al., 1994, "TSNLP: Test Suites for Natural Language Processing" in Proceedings of Language Engineering Convention, CNIT - La Défense, Paris, France: 17-22.

4. Clark, R., 1994, "Computer-assisted translation: the state of the art" in Dollerup, 301-308.

5. Dollerup, C., and Lindegaard, A., 1994, Teaching Translation and Interpreting 2. Papers from the Second Language International Conference, Elsinore, 4-6 June 1993, Amsterdam: John Benjamins.

6. EAGLES (Expert Advisory Group on Language Engineering Standards) Evaluation of Natural Language Processing Systems. Draft - Work in Progress (1994). EAGLES Document EAG-EWG-IR.2, Consiglio Nationale delle Richerche, Instituo di Linguistica Computazionale, Pisa.

7. Flickinger D., et al., 1987, Toward Evaluation of NLP Systems. Hewlett Packard Laboratories. Palo Alto.

8. Hutchins, W. J., and Somers, H. L., 1992, An Introduction to Machine Translation. Academic Press, London.

9. Jordan, P.A.W., et al., 1993, "A First-Pass Method for Evaluating Machine Translation Systems" in Machine Translation 8, Nos 1-2, Special Issue on Evaluation of MT Systems, 49-58.

10. Krauwer, S., 1993, "Evaluation of MT Systems: A Programmatic View" in Machine Translation 8, Nos 1-2, Special Issue on Evaluation of MT Systems, 59-66.

11. Lehrberger J., and Bourbeau, L., 1987, Machine Translation. Linguistic Characteristics of MT Systems and General Methodology of Evaluation, Amsterdam: John Benjamins.

12. Minnes, S. 1993, "Constructive Machine Translation Evaluation" in Machine Translation 8. Nos 1-2, Special Issue on Evaluation of MT Systems, 67-7 5.

13. Nerbonne, J., et al., 1993, "A Diagnostic Tool for German Syntax" in Machine Translation 8. Nos 1-2, Special Issue on Evaluation of MT Systems, 85-107.

14. Newton, J., 1992, Computers in Translation. A Practical Appraisal, Routledge, London.

15. Roudaud, B., et al., 1993, "A Procedure for the Evaluation and Improvement of an MT System by the End-User", in Machine Translation 8. Nos 1-2, Special Issue on Evaluation of MT Systems, 109-116.

16. Sparck Jones, K., and Galliers, J., 1996, Evaluating Natural Language Processing Systems. An Analysis and Review. Springer Verlag, Heidelberg and Berlin.

17. Way, A., 1991, "A Practical Developer-Oriented Evaluation of Two MT Systems", Working Papers in Language Processing 26.

18. Wilks, Y., 1992, "SYSTRAN: it obviously works but how much can it be improved?" in Newton, 166-188.