Rationale for a multilingual corpus for machine translation evaluation

Debbie Elliott
debe@comp.leeds.ac.uk

Anthony Hartley
a.hartley@leeds.ac.uk

Eric Atwell
eric@comp.leeds.ac.uk

School of Computing and Centre for Translation Studies,
University of Leeds, Leeds LS2 9JT. England.

1. Introduction

An overview of research to date in human and automated machine translation (MT) evaluation (Elliott 2002) points to a growing interest in the investigation of new automated methods, allowing for the quick and inexpensive evaluation of MT output. It is clear, however, that corpora designed for this purpose are lacking. Our own research in automated evaluation methods will require not only a corpus of source texts with machine translations that represent actual MT use, but also the detailed scores for these translations given by human evaluators. These scores will allow us to test the reliability of new automated evaluation methods. It is our intention, therefore, to compile a multilingual corpus specifically for MT evaluation, to meet not only our own research requirements, but the needs of the MT community at large.

2. Machine translation evaluation

The evaluation of machine translation output has played a crucial role in the development of MT systems since their emergence over five decades ago. Evaluations are required both by developers, before and after system modifications, and by end-users who wish to compare different systems before making a purchase. However, evaluating the quality of any translated text is complex. Unlike the evaluation of part-of-speech taggers, parsers or speech recognisers (Atwell et al. 2000) it is not simply a matter of comparing MT output to some "gold standard" human translation, since translation is legitimately subject to stylistic and other variation. Instead, MT evaluation relies on either the objective scoring of very specific linguistic phenomena using test suites, or the somewhat subjective quality judgements made by evaluators, who are trained to score individual sentences or text segments using a chosen metric. The problem of subjectivity can, however, be reduced by obtaining scores from several evaluators for each sentence and by calculating a mean score. The reliability of results can also be increased by using a large number of texts.

Designing and conducting reliable human MT evaluations has proven to be costly and time-consuming. As a result, more recent research has involved the investigation and application of automated methods, including IBM's BLEU (BiLingual Evaluation Understudy) method (Papineni et al. 2001) and work by Rajman and Hartley (2001, 2002). Successful automated evaluation methods will allow both developers, who need to conduct frequent evaluations after system modifications, and end-users to evaluate systems more quickly and cheaply.

3. Corpora or test suites?

The evaluation of MT output involves the use of either a collection of texts, which in few cases seem large enough to be classified as corpora, or test suites. A corpus designed for this purpose has typically comprised texts in the chosen source language(s), machine translations produced by the systems for evaluation and one or more expert human translations of each text. Bilingual evaluators might then rate the fidelity (preservation of original content) of each machine-translated sentence or marked segment by comparing it to the source text and assigning a score using a particular scale. Alternatively, monolingual native speakers of the target language would perform the same kind of evaluation using the expert human translations for comparison. Scoring the fluency of each sentence, on the other hand, requires access only to the machine translations from the corpus, as no reference to the source text is needed when evaluating this attribute in isolation.

Whereas corpora are widely used for "black box" MT evaluations by end-users, test suites are more often devised and used by researchers and developers, who need to pinpoint the handling of specific linguistic phenomena to guide system modifications (a "glass box" approach). Test suites for MT evaluation typically comprise many short annotated test items in the source language, with correct target translations, which are referenced according to specific linguistic categories. They allow for the systematic

and objective evaluation of carefully selected linguistic phenomena, complete control over every test point (which may be tested in isolation or in combination with other features) and the opportunity to include negative data to determine how a system deals with input errors. However, as test suites are normally designed to evaluate the handling of grammatical phenomena, the vocabulary is intentionally limited, making them less suitable than corpora for the evaluation of MT system glossaries. Furthermore, test suites for natural language processing applications "normally list items on a par without rating them in terms of frequency or even relevance with respect to an application" (Oepen et al. 1997: 25). Corpora, on the other hand, represent naturally occurring data and can be designed to include texts that reflect user needs. This factor is particularly important for end-users who wish to select an MT system to translate specific text types. It is clear, therefore, that the use of test suites and corpora are not competing evaluation methods, but complementary, insofar as they serve different purposes. Our own research interests lie in the evaluation of MT systems for end-users. We require, therefore, a corpus that represents current user needs.

4. A need for multilingual corpora for MT evaluation

Previous research in MT evaluation has involved the use of either sentences or fairly small numbers of texts. Papineni et al. (2001), for instance, rely on a very small corpus that includes human reference translations. Other research (see Table 1) has made use of the much larger DARPA (Defense Advanced Research Projects Agency) corpus, along with results from the largest DARPA human MT evaluation, carried out in 1994. Researchers have used the DARPA corpus and evaluation results to validate (or not, as the case may be) experimental automated evaluation methods, by seeking correlations between the human DARPA scores and those from new methods. Table 1 details texts or corpora used in a sample of published MT evaluation projects, listed chronologically.

Table 1: The use of corpora and test sentences in previous MT evaluation projects

Author(s) and/or project name	**Evaluation type**	**Attributes tested**	**No. of source items used for evaluation = N**	**No. of human translations of N**	**No. of machine translations of N**
Carroll (Pierce 1966)	Human	Intelligibility Fidelity	144 sentences Scientific Russian	3 English	3 English
Nagao et al. (1985)	Human	Intelligibility Accuracy	1,682 sentences Scientific Japanese	0	1 English
Shiwen (1993)	Human and automated	6 test points: words, idioms, morphology, elementary, moderate, advanced grammar	3,200 random sentences English	1 Chinese	1 Chinese
DARPA 1994 series (White 1997, 2001, forthcoming)	Human	Adequacy Fluency Informativeness	100 texts French 100 texts Spanish 100 texts Japanese (news articles of approx. 300-400 words or 800 Japanese characters)	2 English	5 English (Human and machine translation scores available for research)
JEIDA (Isahara 1995)	Human	Linguistic test sets	770 sentences English	1 Japanese	8 Japanese
Author(s) and/or project name	**Evaluation type**	**Attributes tested**	**No. of source items used for evaluation = N**	**No. of human translations of N**	**No. of machine translations of N**

IBM BLEU (Papineni et al. 2001)	Human and automated	Number of *n*-gram matches between MT output and human translations (with penalties)	Approx. 500 sentences Chinese (all from news articles)	Up to 4 English	3 English
White and Forner (2001)	Test: potential automated method	Noun-compound handling	33 texts French 33 texts Spanish (DARPA corpus)	0	5 English (DARPA corpus with scores)
Reeder et al. (2001)	Test: potential automated method	Named-entity handling	0	1 English of 1 Spanish text (DARPA corpus)	5 English of 1 Spanish text (DARPA corpus with scores)
Miller and Vanni (2001) Vanni and Miller (2001, 2002)	Test: potential automated methods	Coherence, clarity, syntax, morphology, dictionary update, names, terminology	0	1 English of 2 Spanish texts 1 English of 1 Japanese text (DARPA corpus)	3 English of 2 Spanish texts 3 English of 1 Japanese text (DARPA corpus)
Rajman and Hartley (2001, 2002)	Human and automated	Grammaticality, preservation of semantic content	20 French (DARPA corpus)	1 English	5 English of 100 French texts (DARPA corpus with scores) 1 English of 20 French texts (DARPA) by an additional MT system

The largest known corpus for MT evaluation, the DARPA corpus, makes available the associated evaluation scores, which has proved invaluable to the MT community. However, this corpus does have its limitations; it comprises only newspaper articles, representing only a small part of MT use, the source texts are in only three languages and all target texts are in American English. It is also clear from the above information that most projects and, therefore, corpora for MT evaluation are concerned with English as a target language.

In response to these findings, it is our intention to compile a multilingual corpus specifically for MT evaluation. This will not only be used for our own work, but will also be made available for research within the MT community. Before text collection begins however, decisions must be made regarding corpus content, size, language pairs and text types for inclusion.

5. Corpus content

We intend to provide a balanced corpus in terms of the number of words and text types for each language pair. Texts and language pairs will be selected to reflect the actual use of MT systems and our decisions will be guided by a survey of MT users. The corpus will comprise source texts with at least one human translation and a number of machine translations of each one, along with our own detailed human evaluation scores.

The corpus will be made available online, allowing users to browse the contents of each language pair, displayed in the form of a list of text types and topic areas. Users will be able to view each source text along with its human and machine translations, and analyse our human evaluation scores, which will be regularly updated as soon as they become available. The source texts will be of use to anyone wishing to evaluate their own system(s), and the human reference translations will provide material for comparison when scoring the MT output. Furthermore, our evaluation results, in addition to those from the DARPA series, will allow for the testing of experimental automated metrics.

6. Corpus size

Constraints in terms of research time and cost mean that informed decisions must be made with respect to corpus size. Using a very large corpus would be unsuitable for human MT evaluation projects for practical reasons: the greater the number of texts, the more time-consuming and expensive the evaluation. Furthermore, the provision of expert human translations of thousands of texts is costly and unnecessary if valid evaluation results can be obtained from a smaller corpus. On the other hand, proven automated evaluation methods might benefit from a larger corpus, which would allow for the generation of more scores at no greater cost than if a smaller number of texts were used.

This begs the question: at what point does a larger number of texts cease to give us more reliable evaluation results? How many texts do we need to obtain valid scores for system comparison? Our first attempt to answer this question has involved analysing DARPA scores with varying numbers of evaluated texts. We used the three scores (adequacy, fluency and informativeness) for the five machine translations and one human translation of each of the 100 French source texts (of approximately 300-400 words) to calculate a mean score for each number of texts evaluated. Figures 2, 3 and 4 show the mean scores for each of the three attributes for every number of texts evaluated (ie. from one text to one hundred texts). Figure 4 shows the overall mean scores.

Figure 2: Comparison of adequacy scores: DARPA 1994 (French-English)

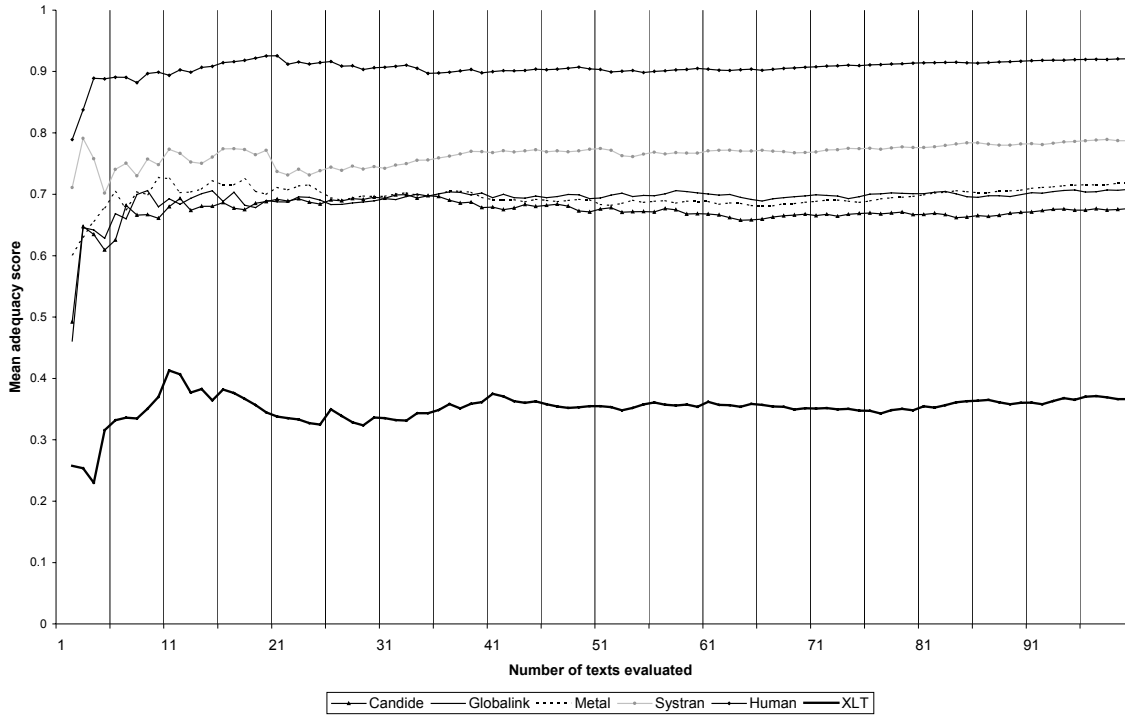


Figure 3: Comparison of fluency scores: DARPA 1994 (French-English)

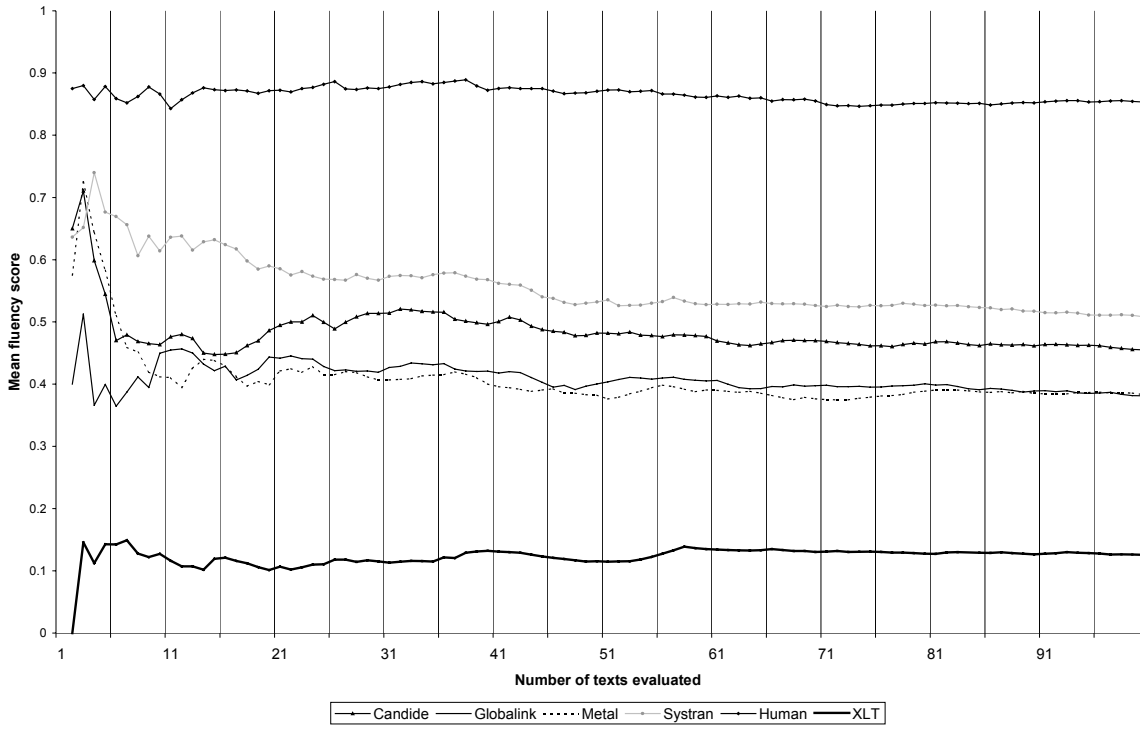


Figure 4: Comparison of informativeness scores: DARPA 1994 (French-English)

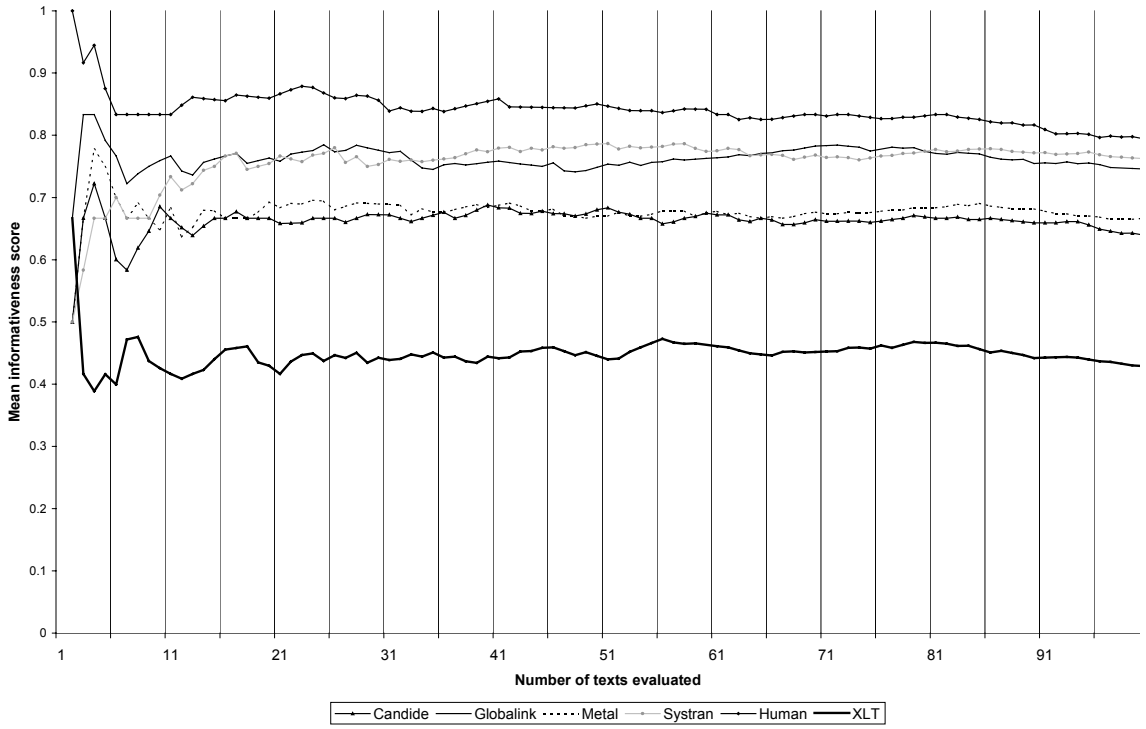
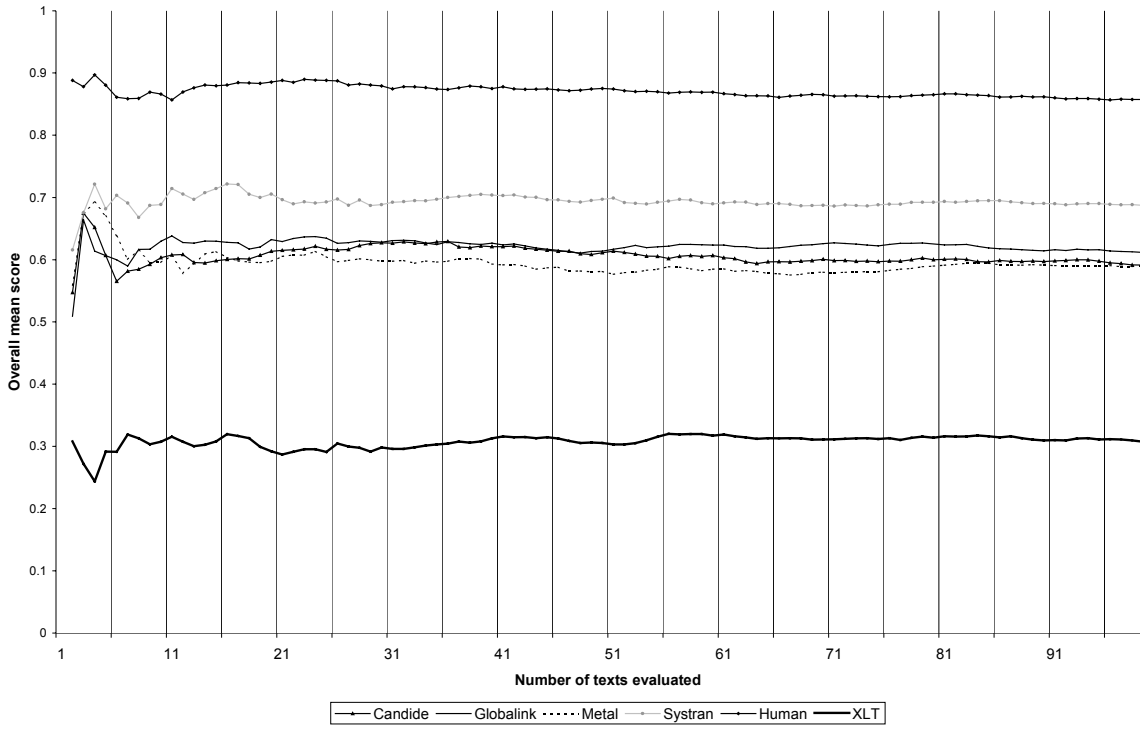


Figure 5: Comparison of overall scores: DARPA 94 (French-English)



Results show that scores from a very small number of texts (perhaps a sample of ten, amounting to around 3,500 words) can allow us to determine the highest and lowest ranking systems, in terms of individual attributes and overall scores. However, the highest scoring "system" here was the human, whom we would normally expect to perform better than the MT systems. It must also be noted that some MT evaluation projects do not involve the evaluation of human translations, but focus on the comparison of MT systems alone. Even then, we are able to determine that Systran performs better than the other MT systems by using scores from as few as ten texts. The only anomaly here is the informativeness score, where Systran and Globalink compete.

A clearer picture of how all five MT systems compare can be obtained after the evaluation of approximately 40 texts (around 14,000 words) for each attribute, and further sampling serves only to confirm this. After around 30 samples, we see that scores begin to remain consistent within a relatively small variance fluctuation, although we do find instances of pairs of systems constantly switching position as more texts are evaluated (Systran/Globalink for informativeness, Globalink/Metal for fluency and adequacy and Metal/Candide for the overall score). In these cases, any number of samples may never see the situation resolved, and the systems that continue to compete according to the number of texts evaluated, can be considered "equal" in terms of particular attributes. It would then be up to the potential user to decide which attribute was more important for their translation needs. For example, a high adequacy score and low fluency score would be more acceptable to someone wishing to use an MT system for gisting or information extraction.

Having obtained these results, our second step was to conduct the same statistical analysis using the Spanish-English and Japanese-English DARPA scores. Results for both language pairs confirmed that reliable scores can be obtained from the evaluation of around 40 texts. We now intend to use texts from our new corpus to conduct human evaluations and to carry out the same analysis. Our initial sample will comprise 35-40,000 words, equal in size to one language pair in the DARPA corpus. This will allow us to compare the number of words required for valid results when evaluating both newspaper articles and different text types, which better represent MT user needs. Our findings will then guide us in terms of the initial number of words required per language pair.

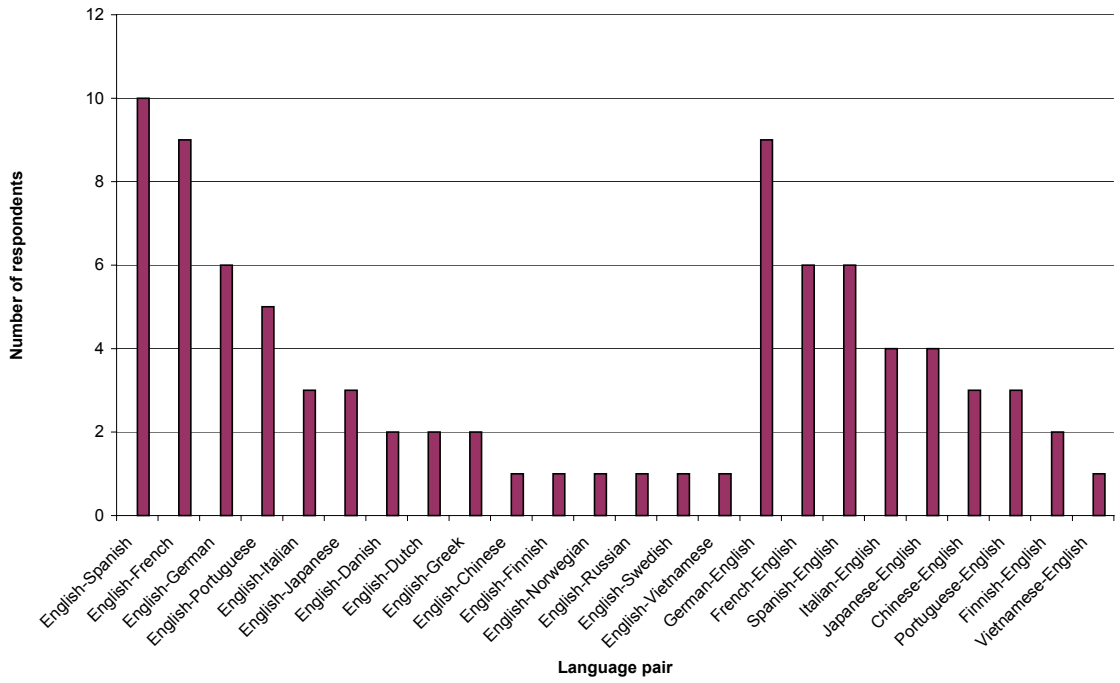
7. Language pairs

In January 2003 we carried out a survey of MT users in order to obtain guidelines for corpus content. In the survey, sent as an email to a number of MT and translation-related mailing lists, we asked which language pairs and text types users regularly translate with the aid of fully automatic MT systems. The 25 replies received to date (16 from large translation providers or international corporations/organisations, 9 from single users) have provided valuable information on both issues. Of the 25 responses, 21 were used for this research, as 4 reported only on their use of translation memory tools. The survey is ongoing and will shortly be available online.

Texts in a number of different language pairs will be needed for our own research, when we investigate new approaches to automated MT evaluation. Furthermore, the availability of texts and translations in several languages will make the corpus more useful for other research projects. It is important to evaluate texts translated from and into more than one language, including languages that are typologically different from one another, to explore the portability of new evaluation methods. Additionally, translation providers often use MT to translate more than one language pair and may need to test systems for several languages.

Figure 6 shows the language pairs (in which the source or target language is English) translated by respondents using MT systems. A very small number of respondents also use systems to translate language pairs that do not involve English.

Figure 6: Language pairs translated by MT users



The number of language pairs that MT systems are now able to handle is constantly increasing. The IAMT (International Association for Machine Translation) Compendium of Translation Software (Hutchins and Hartman 2002) lists an enormous number of MT systems translating many more languages than those shown above. As a starting point, therefore, we plan to collect source texts (with human and machine translations in English) in French, German, Spanish and Italian, along with texts in some typologically different languages, such as Chinese and Japanese to begin with. These will allow us to carry out our initial evaluations of systems translating into English. However, in a second phase we will add translations out of English, which will allow us to test how well existing MT evaluation methods transfer to other language pairs and to develop new machine learnt metrics, which generalise across languages. The target languages for inclusion will be the subject of further research.

8. Text types

Since expectations of MT systems have become more realistic, a greater number of uses have been found for imperfect raw MT output. Consequently, a variety of text types, genres and subject matter are now machine-translated for different text-handling tasks, including filtering, gisting, categorising, information gathering and post-editing (White 2000). It is crucial, therefore, to represent this variety of texts, ranging from emails to technical reports, in our corpus, allowing for the evaluation of texts that represent real MT use.

The main purpose of our survey was to gather information on the kinds of texts and topics most frequently translated using MT systems. Information obtained from this part of the survey is providing useful guidelines on the types of texts to include in our corpus, but there are several problems involving the analysis of data. Firstly, results are based on respondents' own interpretations of the "text types" suggested in the survey and these inevitably overlap in terms of content and grammatical structures. For example, technical material can be found in several separate categories in our questionnaire: internal company documents, technical documents, user manuals, instruction booklets, academic papers and web pages. This must be taken into account when we select our texts. Secondly, some respondents did not specify the subject matter of the material they machine translate, and many were unable to provide details on the number of texts. Finally, it is difficult to equate the comparatively small number of words translated by single users with the millions of words translated by international companies every year. In response to this last problem, we present two sets of results at this stage. Figure 7 shows the number of companies and Figure 8, the number of single users who use MT to translate particular text types.

Responses to date show that single users and companies use MT systems to translate different types of documents. Five of the international companies/organisations who responded did give information about the number of texts they translate. Of these five respondents, all use MT systems to obtain a first draft of either user manuals, instruction booklets, technical documents or internal company documents, or a combination of these. Their total monthly word count is estimated at 3.5 million words. It is crucial, therefore, to represent these documents in our corpus. However, the single user market is likely to grow, as systems become cheaper, so it is important to reflect the needs of such users also.

Findings so far tell us that we must represent all of the above text types to reflect MT use. Documents in our corpus will be categorised, enabling anyone wishing to compare MT systems to easily select source texts for evaluation according to their own needs. The subject matter of these texts will inevitably overlap, as it does in the real world. We are still receiving replies to our survey and updated results will shortly be available online.

Figure 7: Number of companies who use MT to translate particular text types

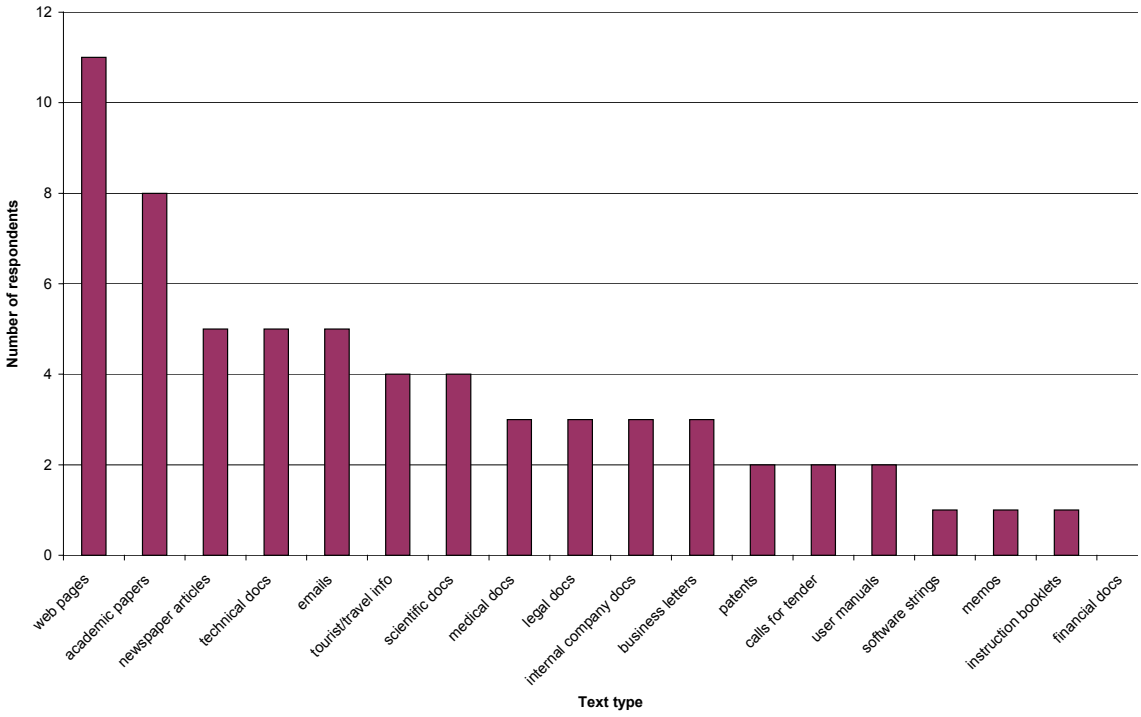
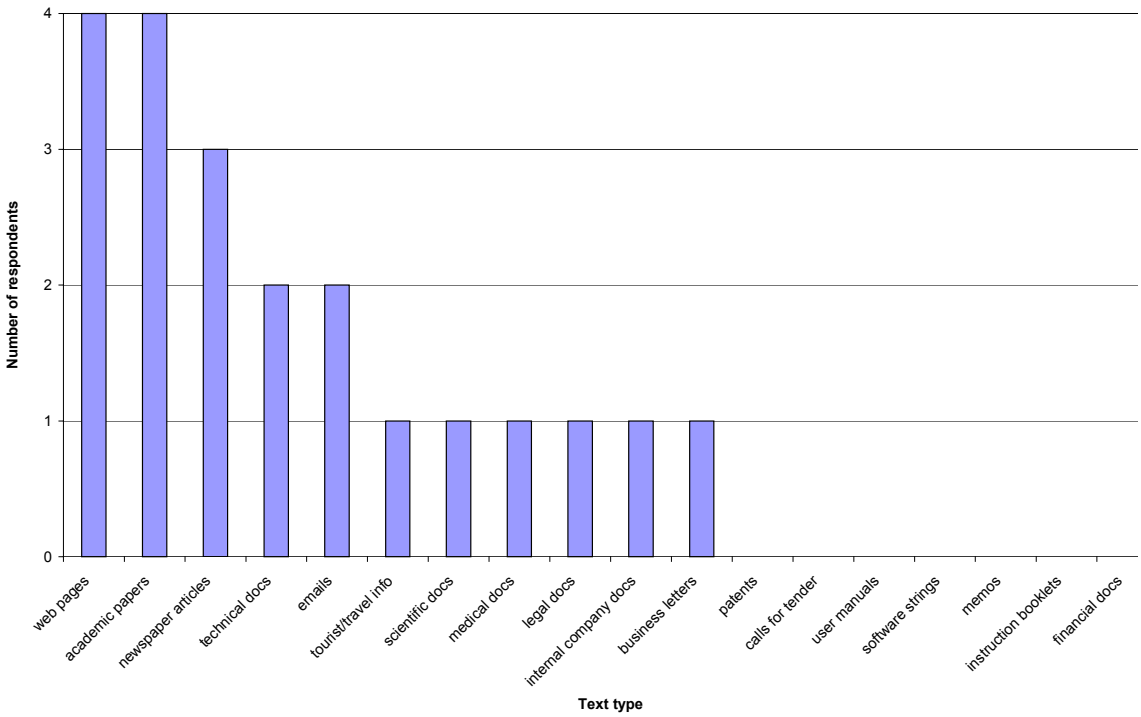


Figure 8: Number of single users who use MT to translate particular text types



9. Conclusion

Our findings to date have provided valuable guidelines for the size and content of our corpus. Analysis of the existing DARPA scores indicates that a small sample of texts (amounting to around 14,000 words) is sufficient to rank a range of MT systems in terms of individual attributes and overall scores. However, our user survey indicates that we need to cover a much wider range of genres, beyond newspaper articles, so there is still a need for a larger corpus. We intend to compile a dynamic corpus, which will be updated to reflect changing trends in the MT user market. New source texts and translations will be added to reflect language change and the introduction of new terminology, and additional MT systems will be added to our evaluations over time. The key feature of our corpus, however, will be the detailed scores from our human evaluations, which will be made available to aid research in automated MT evaluation.

Acknowledgements

We wish to thank everyone who has responded to our MT user survey.

References

Atwell E, Demetriou G, Hughes J, Schiffrin A, Souter C, Wilcock S 2000 A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal* 24: 7-23.

Elliott, D 2002 *Machine Translation Evaluation: Past, Present and Future.* Unpublished MA dissertation, University of Leeds.

Hutchins J, Hartmann W 2002. *IAMT Compendium of Translation Software* 1.5. http://www.eamt.org/compendium.html

Isahara H 1995 JEIDA's Test-Sets for Quality Evaluation of MT Systems – Technical Evaluation from the Developer's Point of View. In *Proceedings of Machine Translation Summit V,* Luxembourg.

Miller K, Vanni M 2001 Scaling the ISLE Taxonomy: Development of Metrics for the Multi-Dimensional Characterisation of Machine Translation Quality. In *Proceedings of Machine Translation Summit VIII,* Santiago de Compostela, Spain.

Nagao M, Tsujii J, Nakamura J 1985 The Japanese government project for machine translation. *Computational Linguistics* 11: 91-109.

Oepen S, Netter K, Klein J 1997 TSNLP – Test Suites for Natural Language Processing. *Linguistic Databases. CSLI Lecture Notes,* CSLI Stanford.

Papineni K, Roukos S, Ward T, Zhu W 2001 BLEU: a Method for Automatic Evaluation of Machine Translation. *IBM Research Report* RC22176. Yorktown Heights, NY:IBM.

Pierce J (Chair) 1966 Language and Machines: computers in Translation and Linguistics. *Report by the Automatic Language Processing Advisory Committee (ALPAC).* Publication 1416. National Academy of Sciences National Research Council.

Rajman M, Hartley A 2001 Automatically predicting MT systems rankings compatible with Fluency, Adequacy and Informativeness scores. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, Machine Translation Summit VIII,* Santiago de Compostela, Spain.

Rajman M, Hartley A 2002. Automatic ranking of MT systems. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, pp 1247-1253.

Reeder F, Miller K, Doyon J, White J 2001 The Naming of Things and the Confusion of Tongues. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, Machine Translation Summit VIII,* Santiago de Compostela, Spain.

Shiwen Y 1993 Automatic evaluation of output quality for machine translation systems. *Machine Translation* 8: 117-126.

Vanni M, Miller K 2001 Scaling the ISLE Framework: Validating Tests of Machine Translation Quality for Multi-Dimensional Measurement. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, Machine Translation Summit VIII,* Santiago de Compostela, Spain.

Vanni M, Miller K 2002 Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Evaluation Metrics across Languages. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain.

White J 1997 MT Evaluation: Old, New and Recycled Methods. *Tutorial slides, Machine Translation Summit VI,* San Diego.

White J 2000 Toward an Automated, Task-Based MT Evaluation Strategy. In *Proceedings of the Workshop on the Evaluation of Machine Translation, Third International Conference on Language Resources and Evaluation*, Athens, Greece.

White J, Forner M 2001 Predicting MT fidelity from noun-compound handling. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, Machine Translation Summit VIII*, Santiago de Compostela, Spain.

White J (Forthcoming) How to evaluate Machine Translation. In Somers H (ed), *Machine translation: a handbook for translators.* Amsterdam, Benjamins.