Compiling and Using a Shareable Parallel Corpus for Machine Translation Evaluation

Debbie Elliott, Eric Atwell, Anthony Hartley
School of Computing and Centre for Translation Studies, University of Leeds, Leeds LS2 9JT
debe@comp.leeds.ac.uk, eric@comp.leeds.ac.uk, a.hartley@leeds.ac.uk

Abstract

TECMATE is a dynamic TEchnical Corpus for MAchine Translation Evaluation currently being compiled and used at the University of Leeds. A purpose-built corpus for machine translation (MT) evaluation differs in terms of size and content from corpora used for other kinds of linguistic analysis. For example, our research in automated MT evaluation requires source texts with human and machine translations as well as the scores for these translations given by human judges. These scores will allow us to test the reliability of experimental automated evaluation methods. Furthermore, a representative sample of machine translations annotated with fluency errors is also required to guide our research into automated error detection. In this paper, we summarise our rationale for corpus design and describe the different stages of corpus development. We provide an example of the content for one language pair and present findings from our recent evaluations of MT output using texts from the French-English sub-corpus. TECMATE will shortly be available online for research.

Introduction

Shareable corpora for MT evaluation research are lacking. The largest known freely available resource is the DARPA corpus (White & O'Connell, 1994), which has been widely used for the testing of new automated evaluation methods (eg. Rajman & Hartley, 2002; White & Forner, 2001; Reeder et al., 2001; Vanni & Miller, 2002). The fluency, adequacy and informativeness scores associated with the translations from the corpus have been used to validate or reject experimental automated evaluation methods, enabling the investigation of correlations between human and automated scores. Although a valuable resource, the DARPA corpus has its limitations; all texts are newspaper articles, representing only a small part of MT use; the 300 source texts are in only three languages (French, Spanish and Japanese) and all human and machine translations are in American English. It is our intention, therefore, to provide a shareable resource that will complement the DARPA corpus.

Rationale for Corpus Design

Corpus Size

Before text collection began, informed decisions had to be taken with respect to corpus size. A large corpus would be impractical for human MT evaluation, as the greater the number of source texts, the more expensive and time-consuming it would be to evaluate the translations. Furthermore, our own research would require expert human translations of each text for comparison against MT output, and 'reference translations' (conveying the content of the source text without stylistic flourishes) to enable monolinguals to evaluate the fidelity of both the human and machine translations. These human translations are expensive to produce.

A large number of texts is not necessary for MT system comparison if reliable evaluation results can be obtained from a smaller corpus. We carried out a statistical analysis of the DARPA scores, for all three language pairs, to determine how many texts would be required to reliably compare MT systems. Results showed that for adequacy, fluency or informativeness evaluations, ten texts (approx. 3,500 words) would be sufficient to rank MT systems, and no more than forty texts (14,000 words) would be needed to offer a clear picture of system performance (Elliott et al., 2003).

Text Types

In 2003, we conducted a worldwide survey of MT users to guide corpus design. The main purpose of the survey was to determine which text types were most frequently translated using MT systems and should, therefore, be represented in our corpus. Responses showed a great difference between the use of MT by companies/ organisations and by individuals who machine translate documents for personal use (Elliott et al., 2003). Individuals most often translated various kinds of web pages, followed by academic papers and newspaper texts. Companies, on the other hand, most frequently machine translated user manuals and technical documents on a large scale. As a result, the decision was taken to represent these texts in our corpus, along with a smaller number of legislative and medical documents. Corporate use of MT put newspaper texts in twelfth place.

Language Pairs

Texts in a number of language pairs (translations into and out of English) will be required to test the portability of automated evaluation methods. To date, the French-English and English-French sub-corpora are complete, and we are currently working on the Spanish, German and Italian into English language pairs. We hope to add further language pairs, including typologically different languages at a later stage.

Corpus Development

Text collection began with the French-English, followed by the English-French sub-corpus. Appropriate parallel texts in other language pairs were also discovered during this process. Our initial aim was to find French original texts with existing good quality human translations. Most freely available parallel corpora were unsuitable for our needs. However, extracts from technical reports were obtainable from the BAF Corpus[1]. The remaining documents were mined from the Web.

Finding good quality translations was a difficult task. Many were badly written, often by non-native speakers, and others, although of excellent quality, were localised to such an extent that they were unusable for MT evaluation. Obtaining copyright permissions was an arduous task, so methods were used to locate suitable documents that contained a permission notice to copy, distribute and modify the text and/or translations. Searches for "Guide de l'utilisateur" + "reproduction permitted" and "logiciel libre" + "copyleft" gave useful results, and many texts produced under the GNU Free (software) Documentation Licence and by the Free Software Foundation Europe were selected.

Although technical in nature, texts were chosen on the basis that they would be understandable to regular users of computer applications, enabling evaluators to confidently judge the quality of the translations.

All selected source texts and translations were checked for errors and translation correspondence. A number of corrections were made, as only perfect input and 'gold standard' translations would enable us to reliably evaluate the quality of the MT output. An English reference translation was then produced for each text. Machine translations of all source texts were generated from three commercial systems (Systran, Reverso Promt and Comprendium) and one online system (SDL's FreeTranslation).

Corpus Content

Each language pair comprises forty source texts of approximately 400 words (equal to the longer texts in the DARPA corpus), and the same categories of text types:

- 10 software user manuals (extracts)
- 10 technical press releases
- 5 technical FAQs (Frequently Asked Questions)
- 5 technical reports (extracts)
- 5 legislative documents (extracts)
- 5 medical documents (extracts)

(The press releases were included at a later stage to represent a greater variety of verb tenses, as the documents initially collected were found to contain mostly imperative and present tense verbs.)

Each source text has an expert human translation, a reference translation, and currently four machine translations. The size of each sub-corpus is approximately 110,000 words. Expert human translations and machine translations will have three human evaluation scores per segment (usually a sentence or heading) for both fluency and adequacy; due to the subjective nature of translation evaluation, one score per segment is insufficient. In addition to these scores, the machine translations of twelve of the source texts (around 20,000 words in total) have been annotated with errors using the Systemic Coder[2] and our new fluency error categorisation scheme.

Evaluation of MT Output

Texts and Evaluators

In our first evaluation, the five translations of a sample of twelve source texts from the French-English sub-corpus were evaluated by thirty monolingual native speakers of English (mostly postgraduate students at the University of Leeds) who had little or no knowledge of French. The intention was to prevent untranslated words in the machine translations from being understood, therefore influencing evaluator judgements.

Design of the Experiment

To provide detailed scores for comparison with results from our new automated evaluation methods, we required translations to be judged at segment level. Each evaluator rated one translation of each source text; judging six translations for fluency and six for adequacy. Both evaluations were based on the DARPA methods. To avoid the "training effect" no evaluator saw more than one translation of the same text.

Thirty evaluator packs were compiled, each comprising translations from different systems in different orders. As every translation would be judged for each attribute by three different evaluators, the same translation would appear in a different position in each pack, preventing the text order from affecting judgements. In half of the packs, the six fluency evaluations appeared first; the other half began with the adequacy evaluations. Judges were not told that the texts were translations. Scores were entered electronically to facilitate their collation and avoid transcription errors.

Fluency

With access only to the translation, evaluators rated each "candidate segment" (most often a sentence or heading) using the Fluency Metric (Figure 1). To simplify the metric, judges were not provided with definitions for scores 2, 3 and 4. For both evaluations, they were asked

[1] http://www-rali.iro.umontreal.ca/arc-a2/BAF/Description.html

[2] http://www.wagsoft.com/Coder/index.html

not to go back to a segment once a judgement had been made.

Fluency
Look carefully at each segment and give each one a score according to how much you think the text reads like fluent English written by a native speaker. Give each segment of text a score of 1, 2, 3, 4, or 5 where:
5 = All of the segment reads like fluent English written by a native speaker
1 = None of the segment reads like fluent English written by a native speaker

Figure 1: Fluency Metric

Adequacy

Judges compared the “candidate text” segments with the aligned "reference text” (reference translations) and used the Adequacy metric (Figure 2) to score each segment.

Adequacy
For each segment, read carefully the reference text on the left. Then judge how much of the same *content* you can find in the candidate text, *regardless of grammatical errors, spelling errors, inelegant style or the use of* synonyms. Give each segment of text a score of 1, 2, 3, 4, or 5 where:
5 = All of the content in the reference text is present in the candidate text
1 = None of the content is present (OR the text completely contradicts the information given on the left hand side).

Figure 2: Adequacy Metric

Results

Three scores were obtained for each segment for each of the two evaluations. A mean score was the calculated per segment of each translation. These scores were used to generate a mean score per text and subsequently per system. Figure 3 and Figure 4 summarise the human evaluation results for both fluency and adequacy.

System	Fluency Score	Adequacy Score
FreeTranslation	2.827	3.644
Comprendium	3.221	4.013
Reverso	3.466	4.142
Systran	3.519	4.136
Human	4.893	4.826

Figure 3: Mean Segment Scores by System

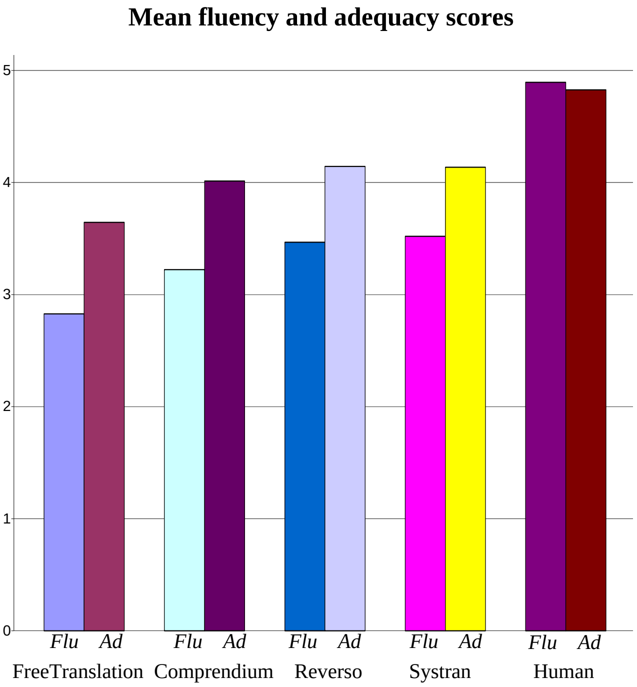

Figure 4: Comparison between MT and Human Translation scores

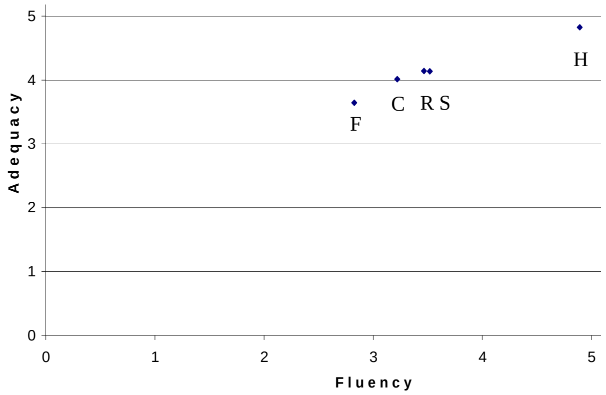

Figure 5: Association between fluency and adequacy values for each system

Systran was the highest scoring MT system for fluency and Reverso for adequacy, by a very small margin. FreeTranslation was the lowest scoring system for both attributes. The machine translations scored consistently more highly for adequacy, indicating that despite a lower level of fluency, the content of raw MT output can be useful. Conversely, there was little difference between the fluency and adequacy scores for the human translations. For all five ‘systems’, a high degree of association was found between values for the two attributes, as shown in Figure 5. Pearson’s correlation coefficient was used to test this hypothesis: using the mean system scores for fluency and adequacy in Figure 3, the value of *r* = 0.98803, showing a very strong correlation between the two variables. This correlation indicates that evaluating either fluency or adequacy would be sufficient to predict values for the other attribute. This supports earlier findings (eg. White, 2001).

Evaluation Time Required

Each evaluator judged 327 segments, rating approximately half for adequacy and half for fluency. The average time taken to complete the fluency evaluation was 33 minutes. The adequacy evaluation contained more reading material and took 48 minutes on average to complete. Without including an introduction to the task, time needed to read instructions, and at least one break, 30 evaluators each required 81 minutes to complete the evaluations. Therefore, the total time needed to evaluate five translations of twelve texts amounted to 40.5 hours.

Conclusions and Further Work

As our experiment shows, machine translation evaluation by humans is expensive and time-consuming. Not only does it involve the careful selection of source texts, often accompanied by good quality human translations, it also requires the preparation of materials (here, segmented aligned texts and metrics) and a sufficient number of human judges. However, these evaluations are necessary to create shareable corpora, with the added value of human scores, to allow for the testing of results from experimental automated evaluation methods.

In terms of corpus development, our next stage will involve the completion of existing language pairs and obtaining human judgements for a greater number of texts. We also plan to investigate correlations between human scores from our recent evaluation and the ranking of the same translations at text level (a cheaper way to evaluate).

We are currently fine-tuning our fluency error classification scheme for French-English machine translations. The annotated texts will be available as a component of the corpus at a later stage. Furthermore, we intend to extend the scheme to additional language pairs, to compare translation errors in English output from different source languages. Statistics resulting from the annotated texts will guide our selection of errors for automated detection. Finally, we will seek to validate our automated methods by using our corpus to find a correlation between human judgements on fluency and adequacy and automated scores.

Each sub-corpus of TECMATE will be made available online when completed. It is hoped that the texts will be of use for research in MT evaluation and other areas of translation studies.

References

Elliott, D., Hartley, A & Atwell, E. (2003). Rationale for a multilingual corpus for machine translation evaluation. In Proceedings of CL2003: International Conference on Corpus Linguistics (pp. 191-200). Lancaster University, UK.

Rajman, M. & Hartley, A. (2002). Automatic Ranking of MT Systems. In Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain.

Reeder, F., Miller, K., Doyon, K. & White, J. (2001). The Naming of Things and the Confusion of Tongues. In Proceedings of the 4th ISLE Evaluation Workshop, MT 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 3rd International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain.

White, J. (2001). Predicting Intelligibility from Fidelity in MT Evaluation. In Proceedings of the 4th ISLE Evaluation Workshop, MT Summit VIII. Santiago de Compostela, Spain.

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

White, J. & O'Connell, T. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In Proceedings of the 1994 Conference, Association for Machine Translation in the Americas. Columbia, MD.