Evaluating Translation Quality as Input to Product Development

Niamh Bohan*, Elisabeth Breidt*, **Martin Volk**†

*Sail Labs GmbH
Balanstr. 57, D-81541 München
elisabeth.breidt@sail-labs.de
nbohan@ireland.com

†Department of Computer Science,
University of Zürich,
Winterthurerstr. 190, CH-8057 Zürich
volk@ifi.unizh.ch

Abstract

In this paper we present a corpus-based method to evaluate the translation quality of machine translation (MT) systems. We start with a shallow analysis of a large corpus and gradually focus the attention on the translation problems. The method constitutes an efficient way to identify the most important grammatical and lexical weaknesses of an MT system and to guide development towards improved translation quality. The evaluation described in the paper was carried out as a cooperation between an MT technology developer, Sail Labs, and the Computational Linguistics group at the University of Zürich.

1. Different types of evaluation for different purposes

Sail Labs does various types of translation quality (TQ) evaluations (absolute, comparative, text and sentence-based) and uses different methods (glass-box, black-box evaluations, pre- and postrelease, using linguistic test suites and real text corpora). Most of the evaluations are from a developer's rather than a user's point of view. Please note that these evaluations were carried out with earlier product versions and the results were used in the development of Sail Labs' current MT technology. For this reason, the concrete results included in this paper (statistics and phenomena) do not reflect the current status of Sail Labs' technology.

Before designing an evaluation method it is crucial to answer the following questions (King, 1997):

- What is the purpose of the evaluation?
- What exactly is being evaluated?

In this paper we focus on a TQ evaluation to answer the question: In which linguistic areas does the evaluated MT system have the most problems? Thus, the purpose of our evaluation is to identify the most costly grammatical and lexical weaknesses so that by concentrating development on these areas, we can most effectively improve the TQ of our systems. We did not want to evaluate the overall TQ of our systems, but rather the problems encountered by the worst translations.

2. Our Evaluation method 'Survival of the Weakest'

We chose a corpus-based approach as we wanted to measure the performance of the MT system with minimal user involvement (e.g., no prior adaptation of bad texts nor lexical coding of unknown words). This means that we checked the 'performance' rather than the 'competence' of the system (Falkedahl, 1998).

We were not merely interested in determining which linguistic problems the system could handle and to which degree, but rather in which areas the system encounters the most severe problems when translating real texts. To achieve a realistic distribution of linguistic phenomena, it is best to use a collection of test sentences covering various linguistic phenomena proportional to their frequency of occurrence in corpus texts. However, this is very difficult if not impossible to obtain. Constructed test suites for linguistic phenomena for which the real occurrence frequency is unknown would also be of little use to us. For these reasons, we selected texts from the Internet and from a corpus CD and considered all phenomena occurring in the test corpus. To this end, the corpus must be big enough to yield representative frequencies of linguistic phenomena. Another advantage of real texts is that they also contain interactions between various linguistic phenomena, which is another important aspect in evaluating the performance of a system.

The evaluation method described here adopts, on the whole, a black-box approach. The advantage is that the evaluation can be outsourced to an institution not involved with the system development. This ensures a more objective evaluation. After the final step of the black-box evaluation, the external evaluators from the University of Zürich passed the results to Sail Labs system developers who carried out the more time-consuming glass-box evaluation using standard methods (e.g., by isolating the suspected phenomenon, tracing the grammar rules etc.).

In the black-box evaluation we applied a 4-step filtering mechanism, where each step involved narrowing down the set of sentences for the next step according to certain criteria. This allowed us to start the evaluation with an extensive data set while continually reducing the data set for the more costly subsequent steps.

Each metric and its rating scale was defined in written form, where possible also with reference to quantitative assignment criteria (e.g., the sentence is bad if more than half is not understandable). From our experience with other evaluation projects and as reported by Sparck-Jones and Galliers (1995), it is crucial to define the evaluation criteria and the values for the text and sentence ratings in as much detail as possible. Among the evaluators, cross-checking and regular discussions helped to ensure that the metrics were applied consistently and subjectivity of ratings was kept to a minimum.

The final result of the evaluation is a list of grammatical and lexical errors with their respective frequencies within the set of worst translations. This list

documents the causes of the most frequent and severe translation problems with the corpus of real texts.

2.1. Selection of test material

For each language direction, we selected between 100 and 140 texts totaling approximately 5500 to 6000 sentences (translation units), mainly from the Internet, some from the ACL/ECI Multilingual Corpus CD1. We chose texts from various subject areas but with little specialised terminology, a) to ensure a good general understanding of the topics by the evaluators, and b) because we develop general-purpose MT technology. The texts were short in order to get a broad variety for a given corpus size and contained sentences of varying linguistic style (simple and complex, short and long sentences, listings and other non-sentence structures). Texts were taken from different domains to suit the purposes of the particular evaluation. We used general texts as well as texts from data processing, car industry, economics, medicine, biology, geography & geology, recreation & sports, linguistics and art & literature. Where available, the texts were translated using the systems' relevant terminology lexica. In order to capture different linguistic styles pertinent to particular domains we selected texts that served various functions (newspaper, manual, internet, dialog). Due to the fact that we used the texts as we found them and no pre-evaluation changes were made, we decided to exclude texts with severe and multiple spelling errors or slang as we were not interested in evaluating the robustness of the MT system when facing bad input, but rather its performance with relatively well-formed texts.

2.2. Step 1: Evaluate TL texts after translation with the MT system

After translation with our MT system, the TL[1] texts were evaluated to identify bad translations. The SL sentences were not taken into consideration in this step as we wanted an evaluation of the generated TL as a stand-alone text. The texts were evaluated according to the following three parameters:

- understandability (the amount of information that is understood by the reader).
- grammaticality (syntactically ill-formed sentences and incorrect morphology).
- lexical correctness (number of unknown, i.e. untranslated words and suitability of chosen words in the given context, not with regard to the SL sentence).

These three parameters were chosen to capture the various purposes a machine translation may serve (information translation or input for postediting). Each criterion was rated on a 3-point value scale:

1. Bad
2. Neither bad nor good
3. Good

The rating for the three parameters was done paragraph-wise to ensure each paragraph contributed equally to the overall score. The average was then computed for the whole text and this introduced decimal scores. Each text was evaluated by three persons to reduce subjectivity. All texts evaluated as generally not good (average point value below 2) progressed to evaluation step 2.

The results were documented extensively including valuable additional information on the overall quality of the translated texts in various subject areas. We documented the grades for the 3 criteria for each text and computed the average across subject areas.

Presuming that even texts that are translated well will contain their share of badly translated sentences, by excluding these texts, we are decreasing our set of badly translated sentences for subsequent steps. For more accurate data on the frequency of problematic phenomenon, we could have skipped this step and evaluated all sentences immediately. However we designed this step to exclude understandable texts with many well translated sentences in order to maximise the relevance of problems contained in the remaining sentences. This also has the advantage of excluding texts from certain genres that are generally translated well.

2.3. Step 2: Evaluate individual sentences

In step 2, the goal was to identify within the 'bad' texts those sentences that are translated the worst. This time, the SL sentences were taken into account for the assessment of the TL sentences to enable a more informed evaluation.

This step was carried out for approx. 3500 – 4000 translation units per language direction. We used two metrics which were rated on a 10 pt scale[2]

- **Preservation of meaning:** Is the meaning of the TL sentence the same as the meaning of the SL sentence?

  7 – 10 points (Good): meaning of SL and TL sentence is about the same. Almost no post-editing with respect to meaning is necessary.
  Example:
  SL: *C'est sur le terrain social que le changement est le plus spectaculaire.*
  TL: *It is on the ground social that the change is the most spectacular.*

  4 – 6 points (Understandable): meaning of SL and TL sentence are not exactly the same, but the sentence can be understood. The sentence may have to be retranslated during postediting.
  Example:
  SL: *"Je doute que les administrateurs d'Alcatel aient eu droit au même traitement", raconte un administrateur, sous le charme du président.*
  TL: *"I doubt that the administrators of Alcatel have had right to the same one treatment", tell one administrator, under the charm of the president.*

  0 – 3 points (Bad): sentence cannot be understood at all or has a completely different meaning to the SL text. Retranslation is definitely necessary.
  Example:
  SL: *A bord, l'aspect détente et loisirs est de rigueur.*
  TL: *To edge, the appearance relaxation and leisure is of rigour.*

- **Grammatical correctness**: Compared to the SL sentence, is the TL sentence syntactically well formed and does it include correct morphology?

[1] TL = target language; SL = source language

[2] It proved that a 3-point scale with the possibility to assign + and – to each value as intermediate ratings would have been sufficient and indeed more intuitive.

7 – 10 points (Good): sentence is grammatically correct. Post-editing would only entail simple style corrections.
Example:
SL: *Le style de l'Audi S3 est certes sportif mais sans tapage, sans arrogance.*
TL: *The style of the Audi S3 is indeed sportif but without uproar, without arrogance.*

4 – 6 points (Understandable): despite grammatical errors, the sentence can be understood. Post-editing would include grammatical corrections of the sentence.
Example:
SL: *Les pères de famille qui lorgnait sur l'A3 peuvent aujourd'hui franchir le pas.*
TL: *The fathers of family that peered on l'A3 today can overcome the step.*

0 – 3 points (Bad): sentence contains massive grammatical errors and can hardly be understood. Post-editing would entail completely rewriting/retranslating the sentence.
Example:
SL: *Sur circuit, ce bolide ne met que 15,8 secondes, départ arrêté, pour atteindre les 200 km/h.*
TL: *On circuit, this bolide does not put that 15,8 second, stopped departure, to affect the 200 km/hr.*

All sentences that received a score of 4 or less were taken as input for Step 3. In our case this consisted of around 2000 translation units per language pair. This constituted 50% of sentences from the 'bad' texts, but just 30% of the original set.

The results again included valuable additional information on the overall quality of the individual sentence translations in various subject areas, which gave us information on the effect of lexical coverage on translation quality. Thus we found for the language pair German to English that the scores were worst for the art texts (grammar average: 4.5; meaning average 3.9) and best for nature texts (grammar average: 6.3; meaning average: 5.5). Averages varied considerably between language pairs reflecting differing degrees of translation quality.

2.4. Step 3: Retranslate sentences with comparison system and evaluate the results

The goal of this step was to further refine the set of phenomena by distinguishing between phenomena generally (too) difficult for MT systems and those which other systems can handle. We compared the translation by the MT system under evaluation with the translation of the same source sentence by a second MT system. This additional test served to isolate the translations that can be handled by other MT systems and was not a general comparison with the second MT system. As evaluation metric we used the better/worse criterion, this time with a 5-point scale to allow distinction between differences of varying gravity. For documentation purposes the MT system under evaluation is referred to here as MT1 and the second MT system as MT2.

++ MT1 TL sentence is much better
\+ MT1 TL sentence is better
= the quality of both TL sentences is similar
\- MT1 TL sentence is worse
-- MT1 TL sentence is much worse

It turned out that the distinction between 'much better' and 'better' and on the other end between 'worse' and 'much worse' was too fine grained. In most cases our judgement was between 'better', 'similar' and 'worse'. Let us first give an example, where MT1 fared better than MT2.

SL: *Y entre todos, nadie disfruta menos del fútbol que Guerrero, metido en una dinámica terrible para sus condiciones.*
MT1: *And among all of them, nobody enjoys the soccer less than Warrior, involved in a terrible dynamics for his conditions does.*
MT2: *And among all, nobody enjoys less than the soccer that Warrior, put in a terrible dynamics for their conditions*.

Please note that this does not mean that MT1 provides a good translation. We must consider that our input in this step consists only of the sentences that MT1 translated badly. With this in mind the ratings take on a slightly altered meaning. For example, '+' means 'MT1 is bad, but MT2 is even worse'.

Given this preselection there were of course numerous cases where the comparison system MT2 fared better than MT1.

SL: *¿De dónde vino este pueblo?*
MT1: *Of where these [people|villages] came?*
MT2: *From where did this town come?*

We used these sentences to identify shortcomings in our system that are not attributable to general MT difficulty. The filtering mechanism 'survival of the weakest' is used to reduce the effort while optimising the results. Thus, in the next step we concentrate on the sentences where MT1 is worse than MT2. Approximately 50% of the sentences compared in step 3 proceed to step 4 which means that the remaining 50% of MT1's worst translations were translated just as badly or worse by MT2. This result allows us to further filter our set of sentences for the glass-box evaluation in a meaningful way. The results again included valuable additional information on the comparison scores across various subject areas.

2.5. Step 4: Identify the phenomena causing the bad translation quality

In this final step of the black-box evaluation, all sentences translated worse by MT1 in step 3 are included. To allow representative conclusions, approx. 1000 sentences per language direction should be used. If less sentences survive, we recommend applying the backtracking mechanism described below.

The goal of this step is to identify the phenomena that failed in these sentences. To this end, we worked out a detailed table to structure the wide range of grammatical and lexical phenomena, which has proved a valuable resource for translation quality evaluations.

Phenomena	Example/Explanation
Complex lexical units	
lexical compounds	*eau de mer => water of sea*
compounds on X-bar level; also compounds connected by dashes	*book store; high frequency*
Lexical ambiguity/homography	
function words	*Comme => as of, like*
lexical choice in content words (same part-of-speech)	*se montre => to prove, to appear*
lexical choice in content words (different part-of-speech)	*Pouvoir => being able, power*
idioms vs. compositional analysis	*he's made the whole thing up =>* *das ist von A bis Z erfunden*
Unknown or untranslated word	
Grammatical phenomena	
Relative clauses	e.g. incorrect relative pronoun
Anaphora resolution	Incorrect choice of personal pronoun *je => I, me*
Adverbs	Incorrect choice of adverb or incorrect adverb position
Coordination	*Après avoir examiné ensemble les données nationales et internationales, …* *=> After having examined the national datum together and internationals, …*
Word order	*Le meilleur c'est bien sûr le moteur 1,8 l suralimenté, ...=> The good one it is of course the motor 1,8 the overfed one, …*

Table 1: Examples of grammatical and lexical phenomena

Table 1 contains an extract of the complete table, which consists of 151 entries structured in a hierarchy 3 levels deep.

This evaluation step resulted in a table with the problematic phenomena sorted by their frequency. The following table documents the eight most frequent problematic phenomena established for one language pair:

Phenomenon
Unknown or untranslated word
Incorrect lexical choice
Incorrect word order
Misplaced or incorrect function word
Incorrect negation
Not recognized fixed multiword expression
Incorrect determiner
Incorrect anaphora resolution

Table 2: Eight most frequent problematic phenomena

2.6. Usage of the Black-Box Results

In order to exactly determine the cause of a translation error and to distinguish between analysis, transfer and generation problems a glass-box evaluation is necessary: Every erroneously translated sentence needs to be debugged, e.g., by isolating the suspected phenomenon, tracing the grammar rules etc. However, as this requires detailed knowledge of the system and is a costly task, we split this step into two parts. First, in step 4 the external evaluators assigned keywords by looking at what went wrong at the surface level in the translated sentences. The second part was done by Sail Labs system developers and consisted of a full-scale glass-box evaluation yielding the final set of grammatical and lexical keywords and their frequencies. This was then used as input to develop improved versions of our MT technology.

2.7. Backtracking Mechanism

With the described evaluation strategy 'survival of the weakest' it is difficult to predict how many sentences will be translated 'badly' enough to progress to the final glass-box evaluation. Our experience indicates that a starting point of 6000 translation units per language pair, combined with the cut-off level defined for each step, is generally sufficient for a representative result.

However should the number of translation units progressing through the steps fall far below our estimates, we recommend defining a backtracking mechanism to ensure a representative set of problematic sentences for the next evaluation step.

Our mechanism involved backtracking to the previous step and redefining the cut-off point to include more input into the current step. However it is important that the cut-off point is not compromised as this would mean that sentences of higher translation standard would progress too far, defeating the purpose of the 'survival of the weakest' filter strategy. In this case, we recommend going back to the first step and broadening the text base for the evaluation.

3. Conclusion

We have shown that a large scale corpus-based evaluation of MT systems is feasible if the effort is structured so that the amount of evaluation material is systematically reduced. We propose to start from a shallow translation quality judgement of complete texts

and work towards a detailed analysis of the most problematic sentences. Clear metrics and evaluation criteria are necessary to achieve this kind of filtering.

With this in mind, it is important to state clearly how the results of the evaluation are to be interpreted. Using this evaluation method, one can identify the most problematic translation issues for your MT system. This method was **not** developed

- to document the translation quality of an MT system.
- to provide reliable statistical data on the distribution of 'good' or 'bad' translations.
- to compare one MT system with another.

The subjectivity of ratings remains a constant challenge. It can be tackled by frequent cross checks and discussions between the evaluators. It is important that the evaluators are language experts in both SL and TL and that they have at least a basic understanding of how MT systems work. This makes it much easier for them to assign the appropriate error labels.

For most of the steps a 3-point scale with the possibility to assign + and – to each value as intermediate ratings proved to be most suitable. A 10-point scale as used in step 2 is definitely too fine grained.

The nature of this evaluation method entailed detailed documentation of the results for each step which yielded reusable resources for other types of evaluation:

- a corpus of specifically selected texts (100 to 140 per source language)
- a breakdown of the results of each step according to subject area. This gave the developer important additional feedback on the coverage of each subject area (for example data processing had fewer lexical choice problems than medicine, etc.)
- annotated sentence triples with a judgement of translation quality, comparative translation quality of two machine translation systems and suspected translation problems. See Figure 1 for the format of these triples.

```
<Sentence-ID Num=1>
<MT1 Grammar=7 Meaning=3
Comparison=-  Problem-phenomena="lexical-choice">
SL:  Une école au tableau noir
MT1: A school at the black picture
MT2: A school to the blackboard

<Sentence-ID Num=2>
<MT1 Grammar=7 Meaning=3
Comparison=-  Problem-phenomena="lexical-choice, untranslated word">
SL:  leurs parents les imitent, pour dissimuler leur honte ou par attachement à
     un maître connu, issu de la région.
MT1: their [parents|relatives] imitate them, to conceal their shame or by
     attachment to a known master, issu of the region.
MT2: their parents imitate them, to conceal their shame or by attachment to a
     known master, descended of the region.
```

Figure 1: Annotated sentence triples

Of course there are many ways to modify and deepen the described method. For instance, during step 1, one could check the contents of the TL texts against the SL texts to ensure a) that what was 'understood' in the TL text actually corresponded to the information in the SL text, and b) that the evaluator could understand the SL text. During the sentence-level comparison in Step 3 one could compare MT1 with more than one other MT system. This would ensure that the identified phenomena are more representative of genuinely difficult areas in MT and would lower the risk that a single comparison system could be bad at translating banal phenomena with which other MT systems have little difficulty. Another interesting aspect of the evaluation would be to classify the SL corpus texts not only by subject area but also according to style, understandability, and length of sentences and to consider these classes in the rating of 'bad texts'. Furthermore, one could produce statistics with respect to sentence length (short, middle, long sentences) as this parameter often strongly interacts with the treatment of linguistic phenomena.

After all is said and done, a large scale evaluation as sketched above remains a management challenge. The evaluators face a hard and sometimes monotonous task. It is important that the evaluation schedule leaves them enough room for exploration and discovery of amusing and puzzling translations. Otherwise there is the "risk of testing the knowledge or intelligence of the test persons, or his/her motivation to understand the text, rather than to measure the quality of the translations" (Falkedahl, 1998).

4. References

Falkedahl, Kirsten (1998). Evaluation Problems from a Developer's Point of View. In: Nübel, Seewald-Heeg (eds.). 137-150.

Nübel, R., U. Seewald-Heeg (eds.), 1998. Evaluation of the Linguistic Performance of Machine Translation Systems. *Proceedings of the Konvens-98 Workshop in Bonn*. Gardez! Verlag, St. Augustin, Germany.

Sparck-Jones, K., Galliers, J.R., 1995. Evaluating Natural Language Processing Systems. An Analysis and Review. *Lecture Notes in Artificial Intelligence 1083*. Springer Verlag, Berlin.

King, Margaret, 1997. Evaluation Design: The EAGLES framework, *Konvens 97 Proceedings.*