Evaluation of machine translation systems at CLS Corporate Language Services AG

Elisabeth Maier*, Anthony Clarke[+], Hans-Udo Stadler[+]

[+] CLS Corporate Language Services AG
Elisabethenanlage 11
CH-4051 Basel
Switzerland
anthony.clarke@cls.ch, hans-udo.stadler@cls.ch
*Canoo Engineering AG
Kirschgartenstr. 7
CH-4051 Basel
Switzerland
elisabeth.maier@canoo.com

Abstract

This paper describes the evaluation of Machine Translation (MT) System for use in a large company. To take into account the specific requirements of such an environment, a pragmatic approach for the evaluation was developed. It consists of five steps ranging from a specification of the evaluation process to the integration of the chosen MT system in a given infrastructure. The process includes a specification of MT evaluation criteria relevant to systems which have to be employed for a large customer base. The paper also shows the results of such an evaluation study which was recently carried out at CLS Corporate Language Services AG, where COMPRENDIUM is in the meantime being employed as corporate MT system.

Keywords

MT evaluation method, intelligibility, correctness, linguistic and non-linguistic evaluation criteria.

Introduction

Over the past few years, a growing number of large enterprises that operate on an international level have developed a growing need for a large amount of translations, which must be delivered in a very short time. The deployment of machine translation as a service offering to company employees and customers is the consequence.
In the case of CLS Corporate Language Services AG, this need arose at its largest shareholder and customer, UBS AG, due to the multilinguality of UBS's employees all over the world and especially in Switzerland. German and English have established themselves as core languages, while other languages such as French. Italian, Spanish etc. are becoming marginalized. Thus the need has arisen – at other multilingual companies as well – to have rapid, cost-effective and good quality translations of texts in German and English, which normally are not translated.
In this context, Machine Translation is to be seen as a complement to human translation and not as a means to rationalise away human translators.

The evaluation of Machine Translation (MT) systems for the deployment in large companies has a number of consequences on the evaluation method. Among the most important are:

- The chosen system needs to comply with the technical standards of the company (e.g. system platform(s), availability, scalability);
- The system must satisfy a range of business requirements (e.g. business languages and domains, types of use);
- The evaluation process needs deliver fast results to guarantee rapid deployment.

In the following chapter, we present a pragmatic approach for the evaluation of MT systems, which takes these factors into account. We first describe the evaluation process and then discuss the various evaluation criteria, which were used.

In the subsequent chapter we show how this method was used for the evaluation of COMPRENDIUM, iTranslator, REVERSO and SYSTRAN, which was carried out in two stages at CLS Corporate Language Service AG. In conclusion, we describe the roadmap for the deployment of the system.

Evaluation Process

The evaluation process deployed for Machine Translation systems was adopted from a generic evaluation framework used for IT components. The process consists of the following four steps:

Step 1: Definition of the evaluation process, co-ordination and execution

The process of evaluation must be determined and co-ordinated by a central unit; this process consists of the following subtasks:

- A detailed specification of the evaluation method;
- A definition of the evaluation criteria;

- The naming and co-ordination of the persons involved in the evaluation (evaluators);
- The co-ordination of the individual evaluation steps;

Step 2: Defining the evaluation criteria and their prioritization

In the second step of the evaluation, the criteria have to be determined which have to be considered when the system candidates are examined. In order to gain an understanding of which criteria are important for the future users and administrators of the system, but also to gain an understanding of what business goals have to be pursued with the deployment of the technology, interviews with representative persons have to be carried out.

Step 3: Development of a test suite

In order to carry out a detailed evaluation, an appropriate test suite has to be drawn up. In the context of an evaluation of MT systems this task includes:

- The preparation of representative texts in the desired source languages and from the desired sectors; if possible these should be in the different text formats required;
- The preparation of appropriate evaluation documents (for example, Excel sheets for the evaluation of the translation quality according to the pre-determined linguistic criteria, evaluator manuals);
- Test installations of the systems in question;
- The preparation of required specialized terminology in the formats supported by the MT systems; if necessary, import and/or generation of import scripts;
- The preparation of sections of the translation memories in the formats supported by the MT systems; if necessary import and/or generation of import scripts;
- The preparation of the translations for all translation directions, without and with dictionaries;
- Preparation of the translations in the style of evaluation forms.

At this point, it should be stressed that the evaluation of the linguistic criteria accounts for the greatest part of the task; the size of this task, however, stands in a direct relationship to the "depth" of the evaluation, i.e. to

- The size of the test suite,
- The number of evaluators,
- The number of the linguistic phenomena examined
- The clarity of the text units examined,
- The complexity of the analysis methods and/or heuristics.

The extent of this task can vary considerably and take up to several months. We thus adopted a pragmatic approach:

- The time-frame for the linguistic evaluation (machine translation, instruction of the evaluators, evaluation of the translations by the evaluators, evaluation of the linguistic criteria) was limited to a fixed period;
- The tasks during the testing phase and the rating by the evaluators were to be run in parallel as far as possible.

Step 4: Detailed evaluation

The detailed evaluation determines the final selection of a product. The systems examined in this phase were installed on-site (evaluation license). The following tasks were carried out:

- Definition of a circle of evaluators;
- Instruction of the evaluators;
- Carry out the evaluation;
- Assessment;
- Draw up a report containing a final recommendation.

Step 5: Implementation

After selection of a system, the system has to be installed and integrated into the infrastructure of the company, in our case into the infrastructure of CLS Corporate Language Services AG.

Evaluation Criteria

A survey of the literature on MT system evaluation carried out during the initial phase of the project showed that a large proportion of papers deal with the evaluation of linguistic features of machine translation, and in particular with the correctness and appropriateness of translations (e.g. Nübel & Seewald, 1998). Non-linguistic evaluation criteria are described only very sporadically. Recently, a classification was developed which brings together a wide range of evaluation criteria (The ISLE Classification of Machine Translation Evaluations, 2000). This consolidated list can be considered the point of departure for the determination of the evaluation criteria used in our study.

The individual evaluation criteria were re-classified into five groups

1. User-oriented criteria
2. Linguistic criteria
3. Technical criteria
4. Economic criteria
5. Strategic criteria

The criteria were specified and weighted according to their importance by the management of CLS Corporate Language Services AG (1, 4, 5), by the IT support staff (3) and by language experts / translators employed at CLS Corporate Language Services AG (2).
In the following, we describe each of the five groups in turn.

User-oriented evaluation criteria

These criteria address issues that concern the user type and type and range of MT services offered to the prospective users:

- **User group**: Lay users (no or only superficial knowledge of the target language) versus expert users (very good knowledge of the target language)**;**

- **Intended use of the target text:** unedited delivery of translation to users versus post-editing of text by human translation service.
- **Language pairs:** specification of source and target languages
- **Domains**: subject areas to which most of the source texts belong.
- **Text formats**: e.g. rtf, doc, HTML, ASCII; Retention of formatting information in the target text;
- **Vocabulary**: possibility to use domain-specific vocabulary and/or customer-specific dictionaries;
- **Technical preferences and system requirements of the user**:
    - **Up-/Download:** browser-based solution, mail solution;
    - **User control:** possibility for the user to manipulate system parameters (e.g. languages, dictionaries, translation preferences)
- **Miscellaneous:** maximum length of a text; maximum response time tolerable; user identification (log-in);

Linguistic criteria

In order to determine a suitable number of linguistic criteria, a brief literature study was carried out; additionally, a criteria list developed in an evaluation carried out in 1999 at CLS Corporate Language Services AG was used (Wenk-Furter, C. & Käser, M. (2000)). The resulting list of linguistic evaluation criteria can be assigned to the following groups:

- **Intelligibility**: correct reflection of the text sense in the target text;
- **Correctness**:
    - **Grammatical correctness**: Recognition and correct translation of grammatical structures;
    - **Lexical correctness**: appropriate and context-driven translation of words and word groups; recognition and correct processing of names and idiomatic expressions.

The evaluation according to the above-mentioned linguistic criteria was carried out in a differentiated manner, depending on which supplementary sources of knowledge were included during translation:

- Use of the MT system **"as-is"**;
- Inclusion of **technical dictionaries provided by the vendor**;
- Inclusion of **terminology databases of CLS Corporate Language Services AG**;
- Inclusion of **the translation memories of CLS Corporate Language Services AG**;

Technical criteria

For the implementation of an MT system, a very heavy weight must be given to technical feasibility and the general requirements to be placed on any future provider; amongst the most important criteria are:

- **Openness** in the context of inclusion of lexicons, translation memories, e-commerce and/or B2B software via documented interfaces; accounting software;
- **Scalability** in the context of the number of users, and/or number of the access hits, inclusion of new language pairs / glossaries /translation memory data, etc.;
- Availability of **strong security concepts**;
- **Configurability** of the system to comply with the needs of the users and of the IT administrators;
- Specification of the **system platforms** which are supported by the system (e.g. hardware, operating system, mail system, security framework);
- **Architecture** of the MT system (e.g. openness, use of standards, Web and/or thin-client solution);
- **Time required for installation** at the customer's premises;
- **Time spent on maintenance and operations**;
- Information provided by **customers already using the system;**
- **Efficiency / response times**, i.e. throughput of the MT system;
- **Stability / robustness** of the system, number of crashes during the evaluation phase, automatic recovery, means for automatic logging and monitoring.
- **Vendor support** (all levels) during the installation and evaluation period.

Economic criteria

- **Licence costs** for server and if necessary for client licences or general network licences;
- **Costs for maintenance and upgrades;**
- **Service costs;**
- **Follow-up costs** for the implementation of the necessary **infrastructure** (hardware and software, for example dedicated MT server, databases, etc.).

Strategic criteria

On the part of CLS Corporate Language Services AG:

- **Future expansion of the MT system** with regard to the languages used and the inclusion of other modules developed or hosted in-house (term bank, translation memory, if necessary both customer-specific);
- **Future application of the system**: use of the source text also as a starting point for post-editing by CLS Corporate Language Services AG's translators.

On the part of the vendor of a MT system:

- Expansion to include further **languages**;
- Integration in soft and hardware **platforms**;
- Integration of other controlled language and/or translation tools;
- Integration into **B2B** and/or **e-commerce** platforms;
- Standard interfaces.

Evaluation Data

Language Pairs

The language pairs examined were German-French, German-English and English-French.

Systems

Five different MT systems were evaluated in two project phases: in the first phase iTranslator from Lernout&Hauspie (all required language pairs), SYSTRAN (language pairs: English-German, English-French) and REVERSO from Softissimo (language pair: German-French) were examined. In the second phase COMPRENDIUM from Sail Labs (all language pairs) and REVERSO (all language pairs) were studied.

In all cases, the goal was to evaluate enterprise systems to gain a realistic picture of the system in its future system environment. In some cases (REVERSO), only standalone versions were available as an evaluation copy, so that the evaluation of most of the technical and some user-oriented criteria could not be carried out.

In all cases the vendors were asked to supply us with the evaluation copy of their choice, which in the first project phase was installed by the IT support staff of CLS Corporate Language Services AG and in the second project phase, due to severe time restrictions, by the system vendors themselves.

Test Suite

At the beginning of the first project phase, translators of CLS Corporate Language Services AG selected a range of texts which they considered typical for their customers. From the resulting set, a test suite was created by additionally choosing texts that fulfilled the following criteria:

- Coverage of different text types (letters, memos, market reports);
- Occurrence of different text characteristics (tables, header/footer information, figures and/or diagrams, weakly and strongly formatted texts).

The texts were taken from three different domains:

- Finance industry / insurance (6 texts),
- Computer science (4 texts),
- Human resources, logistics, general (4 texts).

The texts were translated with all systems in all language directions available; where possible, the texts were also translated using dictionaries (vendor dictionaries and dictionaries produced by CLS Corporate Language Services AG) and translation memories (also produced by CLS Corporate Language Services AG).

The translations were then prepared for evaluation in the following way:

- **Evaluation of intelligibility**: The intelligibility of translations was measured by evaluating sentences, sections and the texts as a whole on a trivalent scale as either "intelligible", “intelligible but not certain" or "unintelligible". Additionally, the subjective intelligibility measures produced by the evaluators were examined by means of comprehensibility questions on the text.
- **Evaluation of correctness**: for every text, a style sheet was prepared which could be used to annotate the translations with any of the translation errors that were determined.
- **Evaluation** of translation improvements by use **of vendor dictionaries**: for every text where the translation done with dictionaries differed from translation produced by the raw system, all the translation changes were marked in the text. In such cases evaluation sheets were prepared where the evaluators could rate the individual changes according to one of the values “better translation”, “worse translation” “same translation quality”.

For the evaluation of the linguistic criteria the following evaluator groups were employed for the different evaluation types:

- **Evaluation of intelligibility**: in order to estimate the intelligibility of the texts and therefore the usefulness of the translations, evaluators were engaged who were considered typical users of such a system in an enterprise environment; the evaluators came from the most diverse backgrounds, such as IT, banking, administration. Additionally, the (much smaller group of) language experts rated the intelligibility of the translations.
- **Evaluation of correctness**: since the assessment of the translation correctness requires deep linguistic skills, language experts working at CLS Corporate Language Services AG carried out this part of the evaluation.
- **Evaluation** of translation improvements through the use of **vendor dictionaries**: expert and lay evaluators rated the changes of translation quality through the use of subject area dictionaries.

Evaluations

In the second phase, only the translation correctness was not studied since this process was shown to be very time consuming while the available time budget of the expert evaluators and of the project was very limited.

First project phase:

- Intelligibility
  - 117 texts were evaluated by 16 lay evaluators;
  - 26 texts were evaluated by 10 language experts;
- Correctness
  - 26 texts were evaluated by 10 language experts;
- Dictionary Improvements
  - 9 texts were examined by 5 language experts

Second project phase:

- Intelligibility
  - 144 texts were evaluated by 18 lay evaluators;
- Dictionary Improvements
  - 41 texts were evaluated by 18 lay evaluators;

Evaluation Results

After an overall comparison of the individual systems, the COMPRENDIUM system of Sail Labs was chosen. In this chapter we will give a brief overview of the results with concerning the various criteria types. The detailed results are described in (Maier, E et al., 2001) and (Maier & Hengartner, 2001).

It has to be noted that in the first project phase an integration of SYSTRAN with REVERSO 4.0 (language pair: German-French) via a uniform user interface was considered, in order to cover all necessary language pairs, which at that time were not available in the enterprise version of SYSTRAN. In the second phase, instead, REVERSO 5.0 was examined as a full system covering all necessary language pairs.

Table 1 gives an overview of the rankings of the individual systems. Economic criteria are not included in this table since at the time of compiling the final report the costs of the individual systems were still to be negotiated:

Criteria	iTranslator	SYSTRAN/REVERSO	COMPRENDIUM	REVERSO
User-oriented	3	2	1	4
Linguistic	4	2	1	3
Technical	2	3 (SYSTRAN only)	1	Not fully evaluated
Strategic	4	2	1	2
Overall	4	2	1	3

Table 1: Rankings of the individual MT systems for the various types of evaluation criteria.

In the following we discuss the results in detail.

User criteria

Concerning the user-related criteria, COMPRENDIUM fared much better than the competing systems. This is mostly due to the fact that all necessary languages were offered, that TRADOS translation memories could be integrated without problems, that the editing of lexicons is supported with comfortable tools and that only very minor formatting problems occurred through translations. Although the other systems officially also offer similar functionalities, problems were encountered during the evaluation: the integration of Translation Memories, for example, could either not be brought into operation during the evaluation phase, or was not part of the evaluation license.

Taking all these factors into account, COMPRENDIUM's evaluation of user-related features resulted in a total of 28 points while the second-ranking system only received 15 points.

Linguistic Criteria

Intelligibility

In terms of the intelligibility of sentences, COMPRENDIUM turned out to be the system with the highest number of comprehensible sentences. iTranslator, on the contrary, had the highest number of incomprehensible sentences, while at the same time never faring as best system in terms of comprehensibility. The numbers for paragraphs and texts as a whole were analogous.

Correctness

As mentioned above, the correctness evaluations were only carried out in the first phase of the project. i.e. with REVERSO 5.0 and COMPRENDIUM no comparison can be made. Comparing the results of the intelligibility and correctness tests of the first project phase, the results diverge slightly: in contrast to the intelligibility results, the correctness of the translations shows results which are slightly in favour of a combination of SYSTRAN/ REVERSO 4.0. Nevertheless we consider the results of the intelligibility tests trustworthier for the following reasons:

- A significantly larger data set was examined as part of the intelligibility tests, so that the data can be considered more reliable. In some cases, only one correctness test could be carried out for any given translation direction. The annotation of more texts, on the other hand, was beyond the project's time budget.
- The various translation error types examined showed a high degree of interdependence, so that the figures for the individual errors are not fully reliable.

For these reasons we attributed more importance to the results of the intelligibility evaluation.

Comprehensible Sentences	SYSTRAN	ITRANSLATOR	COMPRENDIUM	REVERSO 5.0
D-E	56.03	63.29	56.60	72.49
D-F		32.57	62.57	60.87
E-D	52.62	49.08	52.42	41.32
E-F	58.66	60.46	71.78	58.22
F-D		44.87	54.49	41.91
F-E	77.33	54.63	77.26	86.46

Incomprehensible Sentences	SYSTRAN	ITRANSLATOR	COMPRENDIUM	REVERSO 5.0
D-E	18.62	17.24	21.41	9.81
D-F		28.39	18.42	7.43
E-D	10.98	16.56	15.12	15.85
E-F	13.86	19.04	3.35	12.48
F-D		18.38	18.37	15.98
F-E	4.57	15.97	4.21	4.37

Table 2: Results of intelligibility evaluation.

Technical Criteria

Of the systems tested, COMPRENDIUM was found to be the system which fulfilled the technical criteria described above to the greatest degree (a score of 23 compared to 8 for iTranslator and 6 for SYSTRAN). The good results were due to good support and robustness and convincing scalability concepts. Details of these results are contained in the internal reports by Maier E., Stadler H.U. and Hengartner U. (2001) and Maier E. and Hengartner U. (2001).

Economic Criteria

Since the purchase of an MT system had already been approved and budgeted for at CLS Corporate Language Services AG, the various systems were required to meet the economic criteria described above. All vendors were asked to provide quotes for an enterprise system and all the quotes were within the budgeted framework. The detailed specifications were also only discussed with the vendor once the decision to purchase had been made.

Strategic Criteria

The COMPRENDIUM system was found to have the best tools for lexicon expansion and maintenance, the support from the vendor was better than with the other systems and it was also felt that the vendor's strategic positioning was altogether more in line with that of CLS Corporate Language Services AG.

Introduction Roadmap

The introduction of an MT system in a large enterprise ideally takes place in various steps: it is recommended to provide different incremental releases, so that in each increment

- Greater circles of users can be included,
- The IT support gains experience with the maintenance of the system,
- Those responsible for the MT system can build up a team for the support and the maintenance of the linguistic knowledge sources (Lexicons, Translation Memories)

The following tasks have to be performed:

- Installation of the selected system on the infrastructure of the company, i.e. the CLS Corporate Language Services AG; if required, procurement and installation of necessary hardware and software infrastructure beforehand;
- Integration into the security infrastructure; implementation of authentication / authorization modules;
- Integration into the existing Intranet, Extranet and Internet environments;
- If necessary: Implementation of APIs or import and export functionality for in-house knowledge sources
- In parallel: Carry out function, integration and production tests.
- Education of support and administration staff (both linguistic and technical)

At the time of the publication of this paper the introduction of MT, i.e. of COMPRENDIUM, within the CLS Corporate Language Services AG is at the end of the first phase. The system is running as a pilot system at the site of one of the largest customers of CLS Corporate Language Services AG, with a group of ca. 20 users.

Conclusions

Under the circumstances described, the evaluation process led to the choice of COMPRENDIUM as an MT system for use at CLS Corporate Language Services AG. However, other evaluation criteria might well have led to a different result. We realise that MT systems of other vendors are also being constantly developed and improved and CLS Corporate Language Services AG will continue to monitor these developments and evaluate new findings on the MT front.

Acknowledgements

The authors would like to thank all the people who contributed to this study, in particular Urs Hengartner, April Mackison, Monika Käser, Markus Haas, Herbert Spettel and all the evaluators.

References

Maier, E., Stadler, H.U. and Hengartner, U. (2001). Abschlussbericht. Evaluation von Maschinellen Übersetzungssystemen für die Corporate Language Services AG – Phase 2. CLS Corporate Language Services AG, Internal Report. In German.

Maier, E. & Hengartner, U. (2001). Abschlussbericht. Evaluation von Maschinellen Übersetzungssystemen für die Corporate Language Services AG, Basel, Januar 2001. CLS Corporate Language Services AG, Internal Report. In German.

Nübel, R. & Seewald-Heeg, U. (eds.) (1998): Evaluation of the Linguistic Performance of Machine Translation Systems. Proceedings of the Workshop at the KONVENS-98. Bonn.

The ISLE Classification of Machine Translation Evaluations International Standards for Language Engineering (ISLE), Draft 1, October 2000, see http://www.isi.edu/natural-language/mteval/cover.html.

Wenk-Furter, C. & Käser, M. (2000). Evaluation: Maschinelle Übersetzung D-E; Systran/T1, CLS Corporate Language Services AG, internal report. (In German).