Evaluation as a Language Technology Deployment Trigger

Rita Nübel and Jörg Schütz
IAI Saarbrücken
{rita,joerg}@iai.uni-sb.de

Abstract

Language technology (LT) plays an increasingly important role in the field of multilingual information content production and knowledge management. Companies are more and more faced with the problem of the multilingual bottleneck, and are looking for new technologies which bring in productivity gains in the localisation and globalisation process. On the other hand, the increasing number of LT products, especially machine translation systems (MT) calls for useful and practical assessment procedures for potential users in order to prove if a given system fits a user's needs. *Static* evaluations which test a system's fitness as-is, may be good for coarse-grained system comparisons, e.g. for best buy reports, or to make a system rating as the first step of more detailed system tests. *Dynamic* evaluations which consider also a system's adaptability and extensibility in a certain use case are more time-consuming, but they are the most realistic and constructive approaches to evaluation: they ideally combine both the assessment of the state-of-art of a system with the assessment of its extensibility and adaptability potential. The paper investigates the requirements for a useful evaluation of MT and presents practical guidelines to trigger the successful deployment of LT in an industrial context. The scenario is multilingual technical documentation in the automotive industry.

1 Introduction

Deploying MT in a multilingual content production and management scenario is challenging because of the many well known problems of MT. On the other hand, jumping to "no go" conclusions when looking at bad MT output means ignoring the potential of MT. Today, MT deployment and LT deployment in general is complex not only because of the growing market of LT products which are still (or once again) offered as problem solvers -- though with a little less enthusiasm than in their first declared prime decades ago. Also, the situatedness of these technologies is becoming more and more complex. In the scenario of multilingual document production, MT has to be considered as part of a complex workflow and thus has to "communicate" or to interface with a couple of other processes. Hence, deploying language technology (LT) in a multilingual information production chain within an industrial setting requires a careful analysis of work-flows, including the anticipation of changes induced by LT technology, as well as detailed requirements analyses in various areas (business, user, quality). Furthermore, a thorough evaluation of other existing LT products is essential in order to determine the range of products which shall support the multilingual document production.

Taking these rather complex tasks into account, this paper presents some practical guidelines for a successful deployment of LT focusing on integrating MT into the multilingual document production process. The chosen scenario is the production process of technical documentation in the automotive industry. Section 2 starts from the definition of evaluation of LT / MT, and then identifies possible evaluation addressees, their reasons and some relevant criteria for individual translation tasks. Section 3 describes a possible evaluation scenario in a concrete setting, i.e. multilingual documentation in the automotive industry, and section 4 concludes

with discussing future developments in document production, based on the findings of the previous sections.

2 LT / MT Evaluation: Who, Why, What

2.1 The Definition of LT/MT Evaluation

Definition 1: Evaluation is a decision about the significance, value, or quality of something based on a careful study of its good and bad features ([Collins94]).

Definition 2: Evaluation is the determination of the worth of something to someone. ([EAGLES99]).

These two definitions of the notion of evaluation focus on the result of evaluation activities which concern the evaluation of *objects* like e.g. a software product. We believe that it is necessary to extend these definitions for LT since is bears much on the specific circumstances and related processes in which it is to be deployed. Therefore, our definition of evaluation extends to the following:

Extended Definition : Evaluation is a *decision process* about the significance, value, or quality of something based on a careful *assessment of a given situation and its related processes.*

Given this extended definition it is obvious that setting up an evaluation for LT -- especially for the use of MT is complex task and does not restrict itself to determining features or characteristics of a given product. On the contrary, this rather static view on the evaluation object alone runs the risk of neglecting the effects that the deployment of LT products within a complex working environment may have. This concept of the "situatedness" of LT, especially MT is nothing new ([White00]) but it heavily affects the setting up of appropriate evaluation procedures. And here, potential users of LT / MT are faced with two complex tasks, instead of just one, namely the evaluation project. As a prerequisite, they also have to analyse their own situation, i.e. the working environment within which LT shall be deployed. This analysis in turn determines the evaluation procedure, criteria and metrics to measure them.

2.2 Who wants LT / MT evaluations

Different user needs require different evaluations and associated measures and criteria. Developers in research and product engineering as well as the language industry and investors in both sectors look at LT / MT quality characteristics and potentials from different perspectives. In this paper, we take the perspective of a potential industrial LT / MT user who needs to know if and under which conditions the deployment of the technology will be most profitable in his specific situation.

Industrial MT users may belong to one of the following groups:

- Translators (in-house or outsourced)
- Translation editors
- Monolingual information consumers
- Decision makers.

Our scenario focuses on the translators which have to produce publishable multilingual technical documentation.

2.3 Why to evaluate LT / MT products

Evaluating LT / MT does not only have to answer questions about which system would best fit into a user's workflow. This could already be done on the basis of informal system demos, translation examples of text fragments, best-buy comparative overviews, and other communication channels. This phase can be considered as a pre-evaluation which ideally results in keeping just one or two candidate systems for a more detailed evaluation process.

The most important questions to be tackled then are whether the chosen candidate system or systems is apt to a maximum support of human productivity *and* whether the MT system's efficiency can be increased by introducing other LT components. This evaluation phase will thus trigger the deployment of LT as satellite components of the MT system.

2.4 What to evaluate: Different Criteria for Different Tasks

2.4.1 Tasks

Translation Purpose. Translation tools like MT may be used for different tasks or *purposes*, the most prominent of which are the following[1]:

- Information assimilation
- Information dissemination
- Information communication

In our scenario, the dissemination of multilingual information is the focused purpose of MT, and the evaluation procedure has to be specified accordingly. The quality of the output is a central criterion here, since information dissemination requires high-quality publishable translations. Translation quality is hard to assess since it is a highly subjective task. In-house quality standards could help to reduce the impressionistic touch of the classic criteria such as intelligibility, fluency or readability of the translations.

Translation Process / Workflow. Additional factors which have to be accounted for in the evaluation relate to the translation *process*:

- Changes within the translation workflow
- Integration of LT / MT with existing processes and technologies
- Extra work needed for
  - Pre-editing (formats, grammar, layout, style)
  - Post-editing
  - External resources (import, export)
- Changes of time and cost schemes

Once the translation purpose and the workflow implications are identified, it is necessary to make an analysis of *actual translation costs* to be related to the expected pricing of the possible new translation scenario. Actual translation costs include also hidden costs such as rework and repairing errors. Apart from hardware and infrastructure, f*uture translation costs* include reorganisation, training and maintenance as well as the expected reduction of errors.

---

[1] see also [Hovy98].

2.4.2 Criteria

In a multilingual environment, the most important criteria for the cost-effective introduction of LT tools relate to usability, functionality, and the quality of output.

Usability. According to ISO 9126 software standards ([EAGLES96]) usability is a quality characteristic that is composed of three subcategories:

- understandability
- learnability
- operability

Understandability is a requirement for a software product to allow for a degree of transparency that enables the user to understand the behaviour of a given system when confronted with the problems it shall solve. For LT / MT this means that especially inaccurate behaviour which results in *flawed* output should be more or less understood, because this could help to differentiate between

- lexicon gaps or lexicon errors
- true system weaknesses (e.g. grammatical coverage)
- defective input

and then make it possible for the user to take appropriate measures.

Learnability is an important factor for the estimation of the productivity curve, which usually drops during learning phases but should augment again as fast as possible.

Operability refers to the ease of use of the software. E.g. how simple are import and export functions, how user-friendly are the facilities for lexicon extension, and so forth.

Functionality. According to ISO 9126, functionality is composed of a number of sub-characteristics like suitability, accuracy, interoperability, compliance, and security. Most interesting from a user's point of view here are suitability and accuracy of the software: does the system solve the specific problem at all, and if so, does it exactly do what it is designed for. The functionality characteristic naturally overlaps with questions of the output quality (see below), but on a more general level.

Quality of output. For MT we determine quality of output in terms of the linguistic quality. This includes not only the classical cognitive criteria such as readability, fluency, understandability, and so forth, but takes also into account more objectively measurable criteria like style conformity (pre-defined writing standards, e.g. in technical documentation), or the consistent and correct use of terminology. For other LT products like e.g. language checkers, quality of output refers to quantitative measures like precision and recall, i.e. does the system retrieve all and only those "errors" which it is expected to.

3 Setting up an Evaluation: A Virtual Evaluation Project

3.1 The User Profile

In our scenario, the company interested in LT / MT deployment is the service information department of a global car manufacturer. Among the measurable facts they have an enormous

translation volume covering at least twelve target languages. They have a number of different document types to be edited and translated, e.g.

- manuals
- bulletins
- diagnostics
- hotline information

The LT already in use ranges from TM technology including terminology databases to simple spell checking which is not extended to the specialised terminology.

3.2 The User Requirements

A problem analysis ("increasing translation costs") derives user requirements which can be defined on several levels:

- General:
  - supporting human translators (increase actual throughput):
    - ~ Fully automatic with / without post-editing
    - ~ Raw translation or publishable output
- Technical:
  - Hardware and software platforms
  - Input formats such as Word, SGML/XML, etc.
  - Scalability (import of existing resources, lexicon update)
  - Exchange formats for lexical and terminological resources such as ISO 12620, OLIF, and others
- Linguistic:
  - Grammar and style (source and target language)
    - ~ Compliance with information types
    - ~ Writing rules
  - Comprehensibility and accuracy
    - ~ Compliance with existing in-house standards

3.2.1 Technical Requirements

Any introduction of a new technology has to fit with a given technical environment, and the existing processes and operations. In our scenario, this means at least the new technology must provide appropriate interfaces to ensure a seamless integration. This could be an API on the programmatic level, or a separate exchange/interchange tier that accounts for the ability to deal with different input formats (converter interface) for the proper processing and the sharing of information resources. This is best accomplished through existing standards as, for example, provided by ISO, or recommendations in the sense of de-facto standards such as those provided by the World Wide Web Consortium, the LISA/OSCAR group, OASIS, and others.

A certain attention must be given to scalability to avoid island solutions, which fit only a given deployment scenario but do not scale with process re-organisation or changes in requirements.

3.2.2 Linguistic Quality Requirements

Pre-editing requirements. If MT shall be a realistic candidate tool in order to reduce translation costs in a specific area like technical documentation, it will be necessary to introduce additional LT components in the document production process. These tools shall support time-consuming human pre-editing in terms of language checking on various levels. However, also language checking tools such as spell checkers, grammar and style checkers have to meet the user's requirements and should easily be extensible to the specialised terminology, the text structure properties and the in-house writing conventions. The LT components help to standardise the source language documents on the linguistic level in order to reduce errors like spelling mistakes but also to eliminate variation on the structure and lexical level. This pre-editing process does not only improve the quality of the source language document, but also helps to optimise the results of subsequent LT components such as MT or translation memory (TM) technology.

Post-editing requirements. On the output side, expected post-editing work on MT output should quantitatively and qualitatively be evaluated and measured. Quantitative measured may include error counting or time-measures for the whole post-editing process. However, this process should be made more transparent by the classifying and / or weighting of errors. As a starting point, the SAE J2450[2] quality metric can be used. This de-facto standard has been developed for the evaluation of the quality of translation products of automotive service information. It includes a number of error classes and a weighting scheme. In our virtual evaluation, this quality metric could easily be extended to other MT - relevant error classes with corresponding weightings. E.g. the J2450 metric classifies errors according to their seriousness. This metric could be used for measuring also the extensibility of the MT system. Here, a relevant criterion would be if an error can easily be repaired e.g. by extending the system's linguistic resources, or if a complete revision of the translation is necessary.

3.3 The Evaluation Procedure

To account for the dynamic character of the evaluation, the test procedures themselves are applied in two successive phases. The first phase is intended to test an MT component on selected text samples that would usually go to the translation services for further processing. As a prerequisite, the appropriate terminology is imported to update the system's lexicon resources appropriately.

The MT output of this first testing round will then be analysed and assessed using the proposed methods, i.e. frequency of error types and qualitative analysis of errors. A this stage, impressionistic values such as "acceptable translation" or "non-acceptable translation" are of no use since they do not give any clues of whether and how a defective translation could be improved. Impressionistic evaluations is useful in the pre-evaluation phase where a system as is compares to another system.

In the second round, the same text samples are modified according to assumed linguistic standards, ranging from simple spell checking, terminology checking through to grammar and style checking. Additionally, the MT system's resources are adapted if necessary in order to reduce errors classified as "repairable" (by means of lexicon extension / modification / correction) in the first testing round. Then the texts are once again run through the MT system,

[2] For more information see http://www.sae.org/technicalcommittees/j2450p1.htm

and the same assessment routines apply to the MT output. Post-editing is then performed on the MT output and the throughput time is measured.

3.4 Assessments of the Results

For the assessment of the evaluation results all determining factors have to be taken into account, like e.g.

- Usability (learning curve) of MT system
- Usability of supporting LT tools
- System tuning (terminology update as evolutionary process)
- Pre-editing time (language checking as iterative process)
- Post-editing time (translation as iterative process)
- Product pricing (software, hardware, updates, maintenance)

In case the assessment predicts the desired productivity gains, a deployment plan is developed where both LT and MT components are integrated in the document production workflow.

4 Towards Language Standards

As has been described in the previous sections, some of the prerequisites for MT deployment refer to the introduction of additional LT components which support a linguistic standardisation process on the source language side in order to optimise results on the target language side. The (automatic) translation of technical documentation is then of course also a candidate for standardising efforts. This process will prepare and support the development and use of full-fledged controlled languages both for source language editing and translation in specific technical areas like service information. Once a sufficiently structured *content* standard of document production in the domain of technical documentation is reached, also other authoring approaches like iconic authoring or a completely different document production paradigm like multilingual generation will replace (thus revolutionise) the traditional document production procedures. This implies, however, that the deployment of LT / MT is just an intermediate step where the technology, especially MT itself, being far away from perfection, is only a medium or instrumentality to pave the way for much more sophisticated technologies which do not try to *solve* the not solvable problems of machine translation, but bypass them on various levels of abstraction.

5 References

[Cobuild94] Collins Cobuild English Language Dictionary, HarperCollins, London 1994.

[Eagles99] EAGLES Evaluation Working Group, Final Report, Draft March 1999.

[Hovy98] Hovy, E. *Useful Metrics for MT Evaluation*. Presentation at LREC 1998, Granada.

[White00] White J. / Doyon, J., Talbott, S.: Determining the Tolerance of Text-handling Tasks for MT output. In: *Proceedings of LREC 2000, Athens.*