Deploying the SAE J2450 Translation Quality Metric in MT Projects

Jörg Schütz
IAI Saarbrücken
Martin-Luther-Str. 14
D-66111 Saarbrücken
Germany
email: joerg@iai.uni-sb.de

Abstract

This paper provides a nutshell description of how the recently published proposal of a translation quality metric for automotive service information is applicable in an evaluation scenario that deploys multilingual human language technology (mHLT). This proposal is the result of the J2450 task force group of the Society of Automotive Engineers (SAE). The main focus of the developed metric is on the syntactic level of a translation product. Since it is our belief that any evaluation of a translation (human and machine) should also take into account the semantic level of a human language product, we have slightly reshaped the SAE J2450 metric. In addition, we have embedded the whole evaluation process into an object-oriented quality model approach to account for the established business processes in the acquisition, production, translation and dissemination of automotive service information in SGML/XML environments. This scenario then provides the solid grounding for the setup of a quality assurance process for all dimensions related to the processing (human and machine) of automotive service information. The work reported here is one part of the ongoing European Multidoc project that has brought together several European automotive companies to taming the complexity of service information products in an integrated way. Within Multidoc integration means first and foremost the coupling of advanced information technology and mHLT. These aspects will be further motivated and detailed in the context of the specification of an evaluation scenario.

1 Introduction

Automotive documentation is faced with a tremendous increase of translating technical information into multiple languages. Since most of the translation tasks are contracted out to translation companies, there are no efficient and effective measures for controlling the translation process and for benchmarking the translation quality. This is an unfortunate situation because the quality of a company's technical documentation is an important added value resource and one basis of competitive advantage, including the various legal impacts of inaccurate and inconsistent technical information.

In 1997 the Society of Automotive Engineers (SAE) has initiated a task force with the aim to establish a translation quality metric that could be used by automotive companies to compare the quality of service information translation deliverables. This task force has recently published a proposal which defines eight quality metrics for language translation of automotive service information [SAE J2450, 1999]. Currently, the proposal covers only the syntactic aspects of a translation product, and thus has to be seen as a first step towards a standard for defining and proofing translation quality of automotive service information.

Since this activity will have a broader scope and impact on the quality of technical information as an accompanying product of a technical product, it is an interesting investigation to view this work in combination with the deployment of multilingual human language technology (mHLT), in particular machine translation (MT), in the production and translation of technical information.

For this purpose we have examined the SAE proposal with the result to slightly reshape it by taking into account the requirements of the envisaged Multidoc mHLT deployment scenario ([Multidoc, 1997-1999]). Furthermore, we have defined an evaluation methodology based on a quality model to allow for effective extensibility and efficient maintainability.

In the remainder of this paper, we introduce our strategy of defining a SAE J2450 inspired quality model. For the actual evaluation process we have chosen an object-oriented representation of the service information elements which is particularly feasible in today's SGML-based automotive documentation environments. In section 2, we briefly introduce this service information object model. Section 3 deals with the established quality model and the reshaping of the J2450 quality metric. The application of the model to the evaluation of machine translation products is then presented in section 4. The paper closes with an outlook to future developments in this very important area and its impact on the introduction of controlled languages for service information in general.

2 Information Object Model

2.1 Structural Dimension

Technical information, in particular process-oriented service information, is very well suited to be embedded in an object-oriented framework. For the definition of an appropriate model and its implementation, we have to apply the following steps:

1. Identify the different information types existing in the domain; thereby take into account the reusability of the information type hierarchy for types defined at a later stage, i.e. keep the right balance between a fine-grained and a coarse-grained type hierarchy.
2. Identify the information providers (in the automotive case: design and construction departments), the information producers (technical writers and translators) and the information users (workshop technicians and mechanics).
3. Identify the non-taxonomic relationships between the information types, i.e. the definition of multidimensional relationships (type linking) including a definition for each type; this is similar to the specification of a DTD in an SGML (XML) application.
4. Specify the information space, i.e. define the information objects associated with the information types: content specification regarding the smallest information unit closed under domain-specific semantic aspects, for example, a warning, a certain repair step, and so forth.
5. Eliminate redundancies in the specification of information objects. This might cause a redefinition of the information types.
6. Define the entire information space of the domain which is sort of the basic gene program of service information.

The building of the information object model should be based on inheritance and composition. Each object in the object model is composed of other objects, knows about other objects, or works with other objects. Inheritance is used to extend the attributes that further describe an information object, methods that define the processes (activities) associated with an information object, and connections of objects that define the relationships between information objects. Composition is used to extend responsibilities by delegating work/information to other more appropriate objects. If inheritance is employed, the following criteria must be satisfied by the objects involved: any sub-object

1. must be a special kind of a super-object,
2. never needs to transmute to be an object of some other type,
3. extends rather than overrides or nullifies a super-object,
4. does not sub-type what is merely a utility type (useful functionality one would like to reuse),
5. expresses a special kind of roles, transactions, devices, and so forth.

2.2 Content Dimension

It is interesting to notice that each of the above steps for building the information object model has its equivalence on the linguistic level, namely:

1. Identify the most appropriate linguistic realization types for each of the identified information types. This concerns the morphological level including the terminological level, the syntactic level and the semantic level.
2. Identify the linguistic support utilities, i.e. mHLT, which should be employed by the information providers and the information producers for the acquisition, the production and the translation of information objects, mainly information management utilities, linguistic proofing tools and translation support utilities, and by information consumers for the fulfillment of a service task, mainly retrieval tools and translation gisting tools to support translation on demand.
3. Identify the relationships between linguistic realization types to ensure the appropriate SGML (XML) specification (cf. above).
4. Specify the linguistic information space to allow for an appropriate semantic clustering of the language objects and thus the information objects (cf. above).
5. Eliminate redundancies in the specification.
6. Define the entire linguistic space associated to each information object, which is the linguistic gene program of a service information object. At a later stage this gene program will be used for learning capabilities (cf. below).

The combination of both methods, the structuring of information according to business processes and to content semantics, permits the optimal integration of information technology (IT) and mHLT, which still is a step that (automotive) industry has to fulfill.

3 Quality Model

3.1 Quality Factors and Criteria

The definition of any quality metric should be embedded in a generic quality model to allow for easy extensibility and maintenance. As a general rule of thumb we propose to adopt the quality model specification used in software engineering. This is motivated by the fact that we want to integrate human language technology into the evaluation process, and that we are looking for how some of these processes can be automated. Just as this is the case with source code control in software engineering.

Thus, the general quality model for language translation of service information consists of:

- Set of quality factors
- Set of quality criteria assigned to each quality factor
- Set of quality metrics assigned to each quality criterion
- Set of quality measures assigned to each quality metric

To account for the different language related aspects of the quality of multilingual technical information, we define five quality factors which are concerned with

- the **naming level** of the concepts of the domain
- the **object level** of the information units of the domain

This assumes an object-oriented view of the service information elements as briefly introduced in the previous section. Information objects are organized according to information types, such as service preparation type, warning type, special tools type, service step type, and so forth. Ongoing work for identifying such information types resulted in numbers between 20 and 120 information types [Multidoc, 1997-1999]. We could also talk about the **textual level** to assure a direct applicability of our approach to existing non object-oriented SGML documentation, i.e. the PCDATA and CDATA parts (text between SGML/XML ELEMENT tags), and the text data contained in SGML/XML ENTITY definitions.

Our five main **quality factors** are:

- **Terminology** which concerns the appropriate naming of the domain concepts. It is not restricted to the domain nomenclature since it includes concepts for actions and events as well.
- **Grammar** which is concerned with the grammatical fidelity of the information objects, or the text between SGML tags.
- **Style** which is concerned with general writing guidelines for technical information and specialized corporate writing styles based on company-specific writing guidelines for technical information. This could be also a controlled language (CL) in the spirit of the air and space industry (Simplified English and AECMA) for the source language and the target language (cf. [Godden, 1998]). It also includes localization specific aspects such as the proper selection of the honorific level for Asian languages.
- **Content** which is concerned with the semantic fidelity of the information objects (or text, see above).
- **Structure** which is concerned with the SGML/XML level of the information objects and the representation of the textual units in terms of, for example, an appropriate code page selection.

For each of these quality factors we define the following five **quality criteria.** It should be noted that these quality criteria can be also applied to the source (base) language product. This then makes our quality model sort of generic or universal for human language products:

- **Accuracy:** The capability of the translation product to provide the right results or effects, i.e. process-oriented service descriptions.
- **Compliance:** The capability of the translation product to adhere to standards, conventions or regulations in laws and similar descriptions. This also includes a so-called corporate style.
- **Consistency:** The capability of the translation product to maintain a specific level of human language performance and human language competence.
- **Understandability:** The capability of the translation product to enable the user, i.e. the workshop mechanic in our domain, to understand and fulfill the described processes and procedures, i.e. the suitability regarding particular tasks and conditions of use.
- **Interpretation:** The capability of the translation product to provide the user with the right and unambiguous semantic content.

3.2 Quality Metric

Now, we are in the position to define our set **of quality metrics** which will be assigned to each of the quality criteria. Most of our metrics are also present in the SAE J2450 proposal for the syntactic level which we, however, extend to the semantic level including style conventions. Style rules or writing guidelines include general guidelines for technical writing, company-specific guidelines as well as controlled languages which are located at the top of the style scale.

It is our belief that the evaluation of a translation product (human or machine generated) for service information should take into account style aspects and semantic aspects that are essential for an effective and efficient execution of service operations in automotive workshops. Such style aspects include, for example, noun phrase coordination, support verb constructions, verb ellipses, negation, cross-references, and so forth. A good example where semantics come into play is also provided by the SAE task force in their recent J2450 document ([Godden, 1999]): the translation of the English verb *replace* into French depends on the actual context, *replacer* vs. *remplacer.* Here, our proposed object-oriented approach would also help to clarify the situation on the terminological level. For example, *replace* within an information object type putBackIntoPlace will have its French equivalent *replacer,* and within an information object type useNewItem the French equivalent will be *remplacer.* The current SAE J2450 metric proposal consists of eight classes ([SAE J2450, 1999]):

1. Wrong term (WT)
2. Omission (OM)
3. Grammatical error related to word structure, agreement and part of speech (GE)
4. Wrong word order (WO)
5. Misspelling (SP)
6. Punctuation error (PE)
7. Superfluous text (SF)
8. Miscellaneous error (ME)

It is intended that a human classifies these errors into the above eight classes. In a second step she then categorizes the errors into serious (s) and minor (m), and in a third step each error is assigned a weight between one (1) and five (5). 5 corresponds to a very serious error and 1 indicates an error with a minimum of consequences for the service operation.

To reflect the above considerations we have established the following metric (first part) which is a slight redefinition of the original SAE J2450 classes:

1. Wrong or unapproved term, abbreviation and acronym. In contrast to the J2450 classification, we restrict this class entirely to the terminological level in its genuine sense, i.e. we do not include function words. In addition to genuine terminography we include terms denoting actions and events. Because of this terminology orientation this class also covers semantic errors on the conceptual level as it is also intended in the SAE classification. This class is denoted WT.
2. Omission of text and of graphics with text elements remains as defined in the J2450 class OM.
3. Superfluous text remains as defined in the J2450 class SF.
4. Morphological error regarding word structure, orthography, etc. This class combines the first part of the J2450 class GE (word structure) and the J2450 class SP (spelling errors) in one class MO.
5. Grammatical error regarding word order, agreement, punctuation, etc. This class combines the second pan of the J2450 class GE (agreement), the J2450 class WO (word order) and the J2450 class PE (punctuation) in one class GE.

In addition, the following new classes are defined (second part):

6. Style violation of a specific set of writing rules including controlled language use, honorifics and localization issues (writing system or code page). This class is denoted SV.
7. SGML structure error which could be a wrong SGML structure, the omission of an SGML structure, or a superfluous SGML structure. This class is denoted SS.

Last but not least, to cover errors that cannot clearly classified into the above classes we have

8. Miscellaneous error which is denoted ME.

3.3 Applying the Quality Metric

For measuring the metrics we currently use the same procedure as in the SAE J2450 proposal (cf. above). The actual rating in terms of statistics and radar plots is described in the next section.

As promised, we have only slightly reshaped the original classification with the effect that our metric better accounts for a linguistically motivated classification, and that the metric also takes into account the style level of service information (cf above). In addition, this reshaping is also more suitable for the employment of computational proofing tools as will be also shown in the next section.

It should be noted that translation companies use similar metrics for the implementation of quality assurance processes for their products. However, the number of defined classes varies between 8 as in the SAE case and 21 and even more for particular translation tasks. Our research has shown that in most cases the 8 metrics we have specified is sufficient.

In our application scenario, the actual benchmarking process of a given language translation product should employ the power of existing and emerging human language technology. This gives us the possibility of automating the benchmarking process, and thus the entire evaluation process. In addition, we extent this scenario to the use of HLT for proofing the quality of the base language before benchmarking the translation result. This allows us to establish several feedback flows: From

- base language proofing to the production of information objects which has an impact on the terminological, stylistic and structural level of the information objects.
- translation to base language proofing which has an impact on possible multilingual resources, as well as on certain stylistic and structural aspects of the base language.
- target language proofing to translation which may also trigger additional feedback to the base language production.

The motivation for this approach is to minimize the time and the costs for language proofing including quality validation, and to provide evidence to decision makers that the setup of a quality model combined with the deployment of mHLT for the production, translation and processing (human and machine) of service information results in a ROI on several business dimensions (cf. [Schütz, 1998]).

In addition, this approach permits the automatic generation of base language memories as well as translation memories, as it will be outlined in the next section. In the case of MT deployment, it fosters the necessary communication between MT providers/developers and MT users (cf. [Schütz & Nübel, 1998]).

4 Applying the Quality Model to MT

4.1 Setting up of Process

To ensure the proper measuring of the quality of a translation deliverable we have set up the following evaluation process steps:

- Language proofing of the base language product regarding orthography, grammar, style and terminography. For this we use the Multidoc proofing box ([Multidoc, 1997-1999]). This step is essential for the validation of the translation since only correct information objects are delivered to the translation process.
- Translation of a correct base language information object according to 1. This can be a human translation or a machine translation. However, in the following we restrict the application to MT.
- Language proofing of the produced target language information object regarding orthography, grammar, style and terminography with the Multidoc proofing box, and its validation according to our reshaped SAE J2450 metric. This is an automatic analysis of the proofing results: each error is assigned an appropriate metric measure according to the J2450 proposal, i.e. m(inor), s(erious), and a number between 1 and 5 (cf. below).
- Generating a proofing result in form of a statistic (list form presentation) and a radar plot (graphical representation).

It should be noted that the J2450 measures could be also assigned to the base language product for classifying the quality of a given information object. This then establishes a complete quality assurance process chain for the production and the translation of service information.

4.2 Assignment and Measuring

The crucial point in the evaluation process described above is the actual assignment of the different metric values according to our reshaped error classes. For example, each spelling error in the translation product will get the value s-5 in the case of the employment of an MT system because this points to a generation problem or a lexicon problem and therefore needs the attention of either the lexicon developer or the MT developer.

The same holds for terminological errors: if the wrong term is selected or if the terminological error is based on a writing variation, for example, a term variant with a hyphen.

Grammar errors and style errors will get metric values according to a succinct sub-classification of the implemented grammar rules and style rules of the proofing box. Each subclass is assigned a specific value according to the impact the error may have on the described service operation. For example, the violation of the rule stating "*Follow the logical order of actions and events, take care that preconditions and their instructions are placed before an action or an event.*" would be marked as a serious error. This since the missing preciseness and conciseness of the description may possibly result in dangerous or hazardous consequences of a certain service operation. A rule stating "*Place abbreviations in brackets*.", however, will be marked as minor because there should be no direct effect on a particular service operation.

Other rules with a direct influence on the overall service efficiency are, for example, "*Give only one instruction per sentence*" which, however, could be avoided by checking the base language information object first, and "*Avoid the conditional.*" which again concerns the preciseness and conciseness of the information content.

If the translation result is based on an unapproved base language information object then each error must be traced back to its actual source (at least those errors that are marked serious). Otherwise the MT developer could be again responsible for a revision of her MT product. Nevertheless, the employed proofing utility should be checked in addition. So each error marked as serious with a high number value has to be checked for its actual origin which might be the information object classification, the base language of the information object, or the employed proofing utility (including the formalization of the different rules) and the translation utilities. These checks can be seen as the quality assurance mechanism of the employed automation process.

The overall consistency of an information object is indicated by the number of existing inconsistencies and measured with a predefined threshold which depends on the size of a particular information object. So the value 3 may be acceptable for an information object of the size of 2 or 3 print pages but not for an information object consisting of a 3-line service description.

The generated statistics list gives detailed information about the detected errors, whereas the graphical radar plot presentation is a condensed snapshot of the overall evaluation result. The figure below depicts the error distribution of a sample information object showing a high terminological weakness (WT), the other errors are within a defined acceptable threshold.

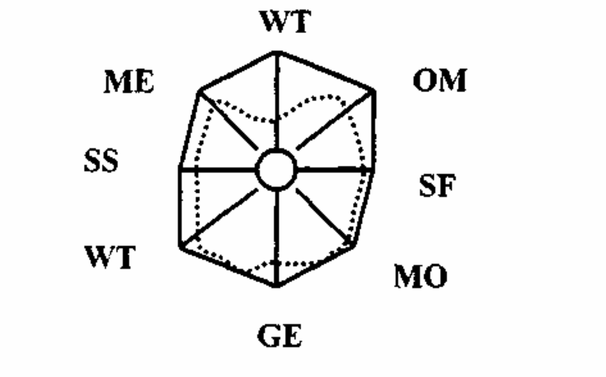

Fig. 1: *Evaluation Radar Plot of an Information Object*

Based on the result of such a translation quality evaluation, we then can automatically generate translation memory entries which are reusable in subsequent translation projects. Similarly, it is also possible to generate base language memories which can be deployed in new compilations of service information objects. The use of base language memories is currently under investigation at several (automotive) industrial sites.

4.3 Impact on MT development

Does our approach has any influence on the development and the evaluation of MT systems in general? First, the presented approach is purpose-driven which means that we are able to define the expected language competence based on our object-oriented approach. For each information object we have an associated predefined linguistic realization class which delimits the number of possible MT results, so that an MT vendor should be able to estimate the fidelity of her MT product. This, however, has to be seen independently of an MT system's terminological competence which is a matter of how a system is able to interact with existing resources, i.e. what exchange formats are supported and what kind of minimal linguistic information is needed to deploy external resources.

Second, our approach is technology-driven which means that the MT system has to provide appropriate open interfaces at the software level (APIs or SDKs) which allow for an effective integration into an existing IT environment.

For both aspects the performance is of crucial importance. A system that performs well on the purpose level but poor on the technology level would certainly not fit our application scenario. As such our approach is suitable for the setup of a general evaluation methodology and strategy which should be purpose-driven as well as technology-driven.

The development of an MT product may distinguish between an OEM/VAR business track for the deployment of special purpose incarnations of the product together with a customer-specific service track (system integration), and a product business track for general purpose developments of the product.

5 Summary and Outlook

In this paper we have presented some insights into our ongoing evaluation work and integration work within the European Multidoc project. It is our opinion that this research and development can be also applied in technical information environments of other industrial branches, such as software localization, telecommunications and even call center applications where we recently started a feasibility study on the deployment of mHLT with the prime aspect being the human language quality analysis procedure for the specification of a problem solving support warehouse (case base).

Our evaluation approach employs several techniques ranging from an object-oriented classification of service information units to the setup of a human language quality model in the automotive application field where at all stages the integration of advanced IT and mHLT is the leading trigger.

For the first implementation of a possible quality assurance process for human language service products we have investigated the SAE J2450 translation quality metric. This metric was slightly reshaped according to our specific needs within our deployment scenario, and integrated into the already existing language proofing tools. The application has now be proven for its suitability on a broader scope, and will also provide feedback to the SAE task force for future extension and amendments of the J2450 work.

Employing the presented object model approach and the associated evaluation strategy also contributes to further study the feasibility and suitability of the introduction of a restricted controlled language for (automotive) service information. This has to be seen in combination with the representation of the service information content in a kind of human-language-neutral meta-language which could be embedded in a specialized set of XML markups (DTD for semantic tagging).

The idea behind such an information markup language is the exploitation of multilingual generation based on CLs, that is the CL defines the capabilities (competence) of the generation module. This area is the envisaged subject of a Multidoc follow-up project. On the research agenda we have put the exploration of further automating the different processes including the evaluation process through the deployment of machine intelligence. At this point, we think of machine learning capabilities based on neuronal networks for the implementation of so-called production and evaluation softbots (cf. [Schütz, 1997]).

6 References

Kurt Godden (1998). “Machine Translation in Context”. In: Proceeding of AMTA 98, Langhorne, PA, USA, pp. 158-163.

Kurt Godden (Chair of J2450 Task Force, 1999). “J2450 Category Definitions”. Update of [SAE J2450, 1999] of April 15, 1999. Personal communication. Multidoc (1997-1999) EU Project Multidoc. http://www.iai.uni-sb.de/MULTIDOC (general project information); see also Schiitz, 1998

SAE J2450 Society of Automotive Engineers Task Force on Translation Quality Metric (1999). World Wide Web documentation available at http://www.sae.org/TECHCMTE/j2450pl .htm (overview) and http://www.sae.org/TECHCMTE/j2450p2.htm (metric categories).

Jorg Schütz (1997). “Utilizing Evaluation in Networked Machine Translation”. In: Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI), Santa Fe, NM, USA, pp. 208-215.

Jorg Schütz (1998). “Multilingual Human Language Technology in Automotive Documentation Workflows”. In: Proceedings of the 20th ASLIB Conference Translating and the Computer, London, Great Britain.

Jorg Schütz and Rita Nübel (1998). “Evaluating Language Technologies: The Multidoc approach to Taming the Knowledge Soup”. In: Proceeding of AMTA 98, Langhorne, PA, USA, pp. 236-249.

Acknowledgements

The work reported in this paper is partly funded by the European Commission under Contract LE4-8323 within the Language Engineering Action Line of the Fourth Framework Programme of the European Union. The views and conclusions expressed in this paper are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the Multidoc consortium and the European Commission.

Thanks for valuable discussions on our approach are due to the automotive members of the Multidoc consortium and to Kurt Godden of the SAE J2450 Task Force Group.