Supports

This workshop has been organised in part with support from the ISLE Project – International Standards for Language Engineering (IST-10647). The ISLE Project is funded by the European Union, the National Science Foundation of the USA, and the Swiss Government.

Workshop Programme

9:00 – 9:15	Introduction and welcome to the workshop – *Maghi King*
9:15 – 9:45	Introduction to the ISLE taxonomy for MT evaluation – *Maghi King* and *Andrei Popescu-Belis*
9:45 – 10:15	Overview of human-based metrics for MT evaluation – *Florence Reeder*
10:15– 10:45	The DARPA 2001 automated metric and its relation to IBM's BLEU – *George Doddington*
10:45 – 11:00	Summary of the proposed evaluation tasks – *Andrei Popescu-Belis*
11:00 – 11:40	**Coffee break** and extra evaluation time [general LREC break: 11:00-11:20]
11:40 – 12:00	Summary of the goals of our collective hands-on experiment – *Andrei Popescu-Belis*
12:00 – 13:00	Reports on individual evaluations (human vs. automatic) – *All workshop participants*
13:00 – 14:30	**Lunch break**
14:30 – 15:30	Reports on individual evaluations (human vs. automatic, *continued*) – *All workshop participants*
15:30 – 16:40	Synthesis of evaluation exercises: reliability and correlation of the metrics used in the hands-on exercises – *Workshop organisers*
16:40 – 17:00	**Coffee break** [synchronised with general LREC break]
17:00 – 18:30	Roundtable discussions of the observed results. Conclusions – *All workshop participants*

Workshop Organisers

Marianne Dabbadie	EVALING, Paris (France)
Anthony Hartley	Centre for Translation Studies, University of Leeds (UK)
Eduard Hovy	USC Information Sciences Institute, Marina del Rey (USA)
Margaret King	ISSCO/TIM/ETI, University of Geneva (Switzerland)
Bente Maegaard	Center for Sprogteknologi, Copenhagen (Denmark)
Sandra Manzi	ISSCO/TIM/ETI, University of Geneva (Switzerland)
Keith J. Miller	The MITRE Corporation (USA)
Widad Mustafa El Hadi	Université Lille III - Charles de Gaulle (France)
Andrei Popescu-Belis	ISSCO/TIM/ETI, University of Geneva (Switzerland)
Florence Reeder	The MITRE Corporation (USA)
Michelle Vanni	U.S. Department of Defense (USA)

Table of Contents

Author Index

An Introduction to MT Evaluation

Eduard Hovy*, Maghi King, Andrei Popescu-Belis****

*USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695, USA
hovy@isi.edu

**ISSCO/TIM/ETI, University of Geneva,
École de Traduction et d'Interprétation
40 Bvd. du Pont d'Arve
CH-1211 Geneva 4 – Switzerland
margaret.king@issco.unige.ch
andrei.popescu-belis@issco.unige.ch

Abstract

This section of the workbook describes the principles and mechanism of an integrative effort in machine translation (MT) evaluation. Building upon previous standardization initiatives, above all ISO/IEC 9126, 14598 and EAGLES, we attempt to classify into a coherent taxonomy most of the characteristics, attributes and metrics that have been proposed for MT evaluation. The main articulation of this flexible framework is the link between a taxonomy that helps evaluators define a context of use for the evaluated software, and a taxonomy of the quality characteristics and associated metrics. The document overviews these elements and provides a perspective on ongoing work in MT evaluation.

1. Introduction

Evaluating machine translation is important for everyone involved: researchers need to know if their theories make a difference, commercial developers want to impress customers, and users have to decide which system to employ. Given the richness of the literature, and the complexity of the enterprise, there is a need for an overall perspective, something that helps the potential evaluator approach the problem in a more informed way, and that might help pave the way toward an eventual theory of MT evaluation.

Our main effort is to build a coherent overview of the various features and metrics used in the past, to offer a common descriptive framework and vocabulary, and to unify the process of evaluation design. Therefore, we present here a parameterizable taxonomy of the various attributes of an MT system that are relevant to its utility, as well as correspondences between the intended context of use and the desired system qualities, i.e., a quality model. Our initiative builds upon previous work in the standardization of evaluation, while applying to MT the ISO/IEC standards for software evaluation.

We first review (Section 2) the main evaluation efforts in MT and in software engineering (ISO/IEC standards). Then we describe the need for two taxonomies, one relating the context of use (analyzed in Section 3) to the quality characteristics, the other relating the quality characteristics to the metrics. In Section 4 we provide a brief overview of these taxonomies, together with a view on their dissemination and use. We finally outline (Section 5) our perspectives on current and future developments.

2. Formalizing Evaluation: from MT to Software Engineering

2.1. Previous Approaches to MT Evaluation

The path to a systematic picture of MT evaluation is long and hard. While it is impossible to write a comprehensive overview of the MT evaluation literature, certain tendencies and trends should be mentioned. First, throughout the history of evaluation, two aspects – often called quality and fidelity – stand out. Particularly MT researchers often feel that if a system produces syntactically and lexically well-formed sentences (i.e., high quality output), and does not distort the meaning (semantics) of the input (i.e., high fidelity), then the evaluation is sufficient. System developers and real-world users often add evaluation measures, notably system extensibility (how easy it is for a user to add new words, grammar, and transfer rules), coverage (specialization of the system to the domains of interest), and price. In fact, as discussed in (Church and Hovy, 1993), for some real-world applications quality may take a back seat to these factors.

Various ways of measuring quality have been proposed, some focusing on specific syntactic constructions (relative clauses, number agreement, etc.) (Flanagan, 1994), others simply asking judges to rate each sentence as a whole on an N-point scale (White et al., 1992 1994; Doyon et al., 1998), and others automatically measuring the perplexity of a target text against a bigram or trigram language model of ideal translations (Papineni et al., 2001). The amount of agreement among such measures has never been studied. Fidelity requires bilingual judges, and is usually measured on an N-point scale by having judges rate how well each portion of the system's output expresses the content of an equivalent portion of one or more ideal (human) translations (White et al., 1992 1994; Doyon et al., 1998). A proposal to measure fidelity automatically by projecting both system output and a number of ideal human translations into a vector space of words, and then measuring how far the system's translation deviates from the mean of the ideal ones, is an intriguing idea whose generality still needs to be proved (Thompson, 1992). In similar vein, it may be possible to use the above mentioned perplexity measure also to evaluate fidelity (Papineni et al., 2001).

The Japanese JEIDA study of 1992 (Nomura, 1992; Nomura and Isahara, 1992), paralleling EAGLES, identified two sets of 14 parameters each: one that characterizes the desired context of use of an MT system, and the other that characterizes the MT system and its output. A mapping between these two sets of parameters allows one to determine the degree of match, and hence to predict which system would be appropriate for which user. In similar vein, various companies published large reports in which several commercial MT systems are compared thoroughly on a few dozen criteria (Mason and Rinsche, 1995; Infoshop, 1999). The OVUM report includes usability, customizability, application to total translation process, language coverage, terminology building, documentation, and others.

The variety of MT evaluations is enormous, from the influential ALPAC Report (Pierce et al., 1966) to the largest ever competitive MT evaluations, funded by the US Defense Advanced Research Projects Agency (DARPA) (White et al., 1992 1994) and beyond. Some influential contributions are (Kay, 1980; Nagao, 1989). Van Slype (1979) produced a thorough study reviewing MT evaluation at the end of the 1970s, and reviews for the 1980s can be found in (Lehrberger and Bourbeau, 1988; King and Falkedal, 1990). The pre-AMTA workshop on evaluation contains a useful set of papers (AMTA, 1992).

2.2. The EAGLES Guidelines for NLP Evaluation

The European EAGLES initiatives (1993-1996) came into being as an attempt to create standards for language engineering. It was accepted that no single evaluation scheme could be developed even for a specific application, simply because what counted as a "good" system would depend critically on the use of the system. However, it did seem possible to create a general framework for evaluation design, which could guide the creation of individual evaluations and make it easier to understand and compare the results. An important influence here was the 1993 report by Sparck-Jones and Galliers, later published in book form (1996), and the ISO/IEC 9126 (cf. next section).

These first attempts proposed the definition of a general quality model for NLP systems in terms of a hierarchically structured set of features and attributes, where the leaves of the structure were measurable attributes, with which specific metrics were associated. The specific needs of a particular user or class of users were catered for by extracting from the general model just those features relevant to that user, and by allowing the results of metrics to be combined in different ways in order to reflect differing needs. These attempts were validated by application to quite simple examples of language technology: spelling checkers, then grammar checkers (TEMAA, 1996) and translation memory systems (preliminary work), but the EAGLES methodology was also used outside the project for dialogue, speech recognition and dictation systems.

When the ISLE project (International Standards for Language Engineering) was proposed in 1999, the American partners had also been working along the lines of taxonomies of features (Hovy, 1999), focusing explicitly on MT and developing in the same formalism a taxonomization of user needs, along the lines suggested by the JEIDA study (Nomura, 1992). The evaluation working group of the ISLE project (one of the three ISLE working groups) therefore decided to concentrate on MT systems.

2.3. The ISO/IEC Standards for Software Evaluation

2.3.1. A Growing Set of Standards

The International Organization for Standardization (ISO) together with the International Electrotechnical Commission (IEC) have initiated in the past decade an important effort towards the standardization of software evaluation. In 1991 appeared the ISO/IEC 9126 standard (ISO/IEC-9126, 1991), a milestone that proposed a definition of the concept of quality, and decomposed software quality into six generic quality characteristics. Evaluation is the measure of the quality of a system in a given context, as stated by the definition of quality as "the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" (ISO/IEC9126, 1991, p. 2).

Subsequent efforts led to a set of standards, some still in draft versions today. It appeared that a new series was necessary for the evaluation process, of which the first in the series (ISO/IEC-14598, 1998 2001, Part 1) provides an overview. The new version of the ISO/IEC 9126 standard will finally comprise four inter-related standards: standards for software quality models (ISO/IEC-9126-1, 2001), for external, internal and quality in use metrics (ISO/IEC 9126- 2 to 4, unpublished). Regarding the 14598 series (ISO/IEC14598, 1998 2001), now completely published, volumes subsequent to ISO/IEC 14598-1 focus on the planning and management (14598-2) and documentation (14598-6) of the evaluation process, and apply the generic organization framework to developers (14598-3), acquirers (14598-4) and evaluators (14598-5).

2.3.2. The Definition of a Quality Model

This subsection situates our proposal for MT evaluation within the ISO/IEC framework. According to ISO/IEC 14598-1 (1998 2001, Part 1, p. 12, fig. 4), the software life-cycle starts with an analysis of user needs that will be answered by the software, which determine in their turn a set of specifications. From the point of view of quality, these are the external quality requirements. Then, the software is built during the design and development phase, when quality becomes an internal matter related to the characteristics of the system itself. Once a product is obtained, it is possible to assess its internal quality, then the external quality, i.e., the extent to which it satisfies the specified requirements. Finally, turning back to the user needs that were at the origin of the software, quality in use is the extent to which the software really helps users fulfill their tasks (ISO/IEC-9126-1, 2001, p. 11).

Quality in use does not follow automatically from external quality since it is not possible to predict all the results of using the software before it is completely operational. In addition, for MT software, there seems to be no straightforward link, in the conception phase, from the external quality requirements to the internal structure of a system. Therefore, the relation between external and internal qualities is quite loose.

Following mainly (ISO/IEC-9126-1, 2001), software quality results from six quality characteristics:

- functionality
- reliability
- usability
- efficiency
- maintainability
- portability

These characteristics have been refined into software sub-characteristics that are still domain-independent (ISO/IEC 9126-1). These form a loose hierarchy (some overlapping is possible), but the terminal entries are always measurable features of the software, that is, attributes. Following (ISO/IEC-14598, 1998-2001, Part 1), "a measurement is the use of a metric to assign a value (i.e., a measure, be it a number or a category) from a scale to an attribute of an entity".

The six top level quality characteristics are the same for external as well as for internal quality. The hierarchy of sub-characteristics may be different, whereas the attributes are certainly different, since external quality is measured through external attributes (related to the behavior of a system) while internal quality is measured through internal attributes (related to intrinsic features of the system).

Finally, quality in use results from four characteristics: effectiveness, productivity, safety, and satisfaction. These can only be measured in the operating environment of the software, thus seeming less prone to standardization (see however (Daly-Jones et al., 1999) and ISO/IEC 9126-4).

2.3.3. Stages in the Evaluation Process

The five consecutive phases of the evaluation process according to (ISO/IEC-9126, 1991, p. 6) and (ISO/IEC-14598, 1998 2001, Part 5, p. 7) are:

- establish the quality requirements (the list of required quality characteristics);
- specify the evaluation (specify measurements and map them to requirements);
- design the evaluation, producing the evaluation plan that documents the procedures used to perform measurements);
- execute the evaluation, producing a draft evaluation report;
- conclude the evaluation.

During specification of the measurements, each required quality characteristic must be decomposed into the relevant sub-characteristics, and metrics must be specified for each of the attributes arrived at in this process. More precisely, three elements must be distinguished in the specification and design processes; these correspond to the following stages in execution:

- application of a metric (*a*);
- rating of the measured value (*b*);
- integration (assessment) of the various ratings (*c*).

It must be noted that (a) and (b) may be merged in the concept of ‘measure’, as in ISO/IEC 14598-1, and that integration (c) is optional. Still, at the level of concrete evaluations of systems, the above distinction, advocated also by EAGLES (EAGLES-Evaluation-Workgroup, 1996), seems particularly useful: to evaluate a system, a metric is applied for each of the selected attributes, yielding as a score a raw or intrinsic score; these scores are then transformed into marks or rating levels on a given scale; finally, during assessment, rating levels are combined if a single result must be provided for a system.

A single final rating is often less informative, but more adapted to comparative evaluation. However, an expandable rating, in which a single value can be decomposed on demand into several components, is made possible when the relative strengths of the component metrics are understood. Conversely, the EAGLES methodology (EAGLES-Evaluation-Workgroup, 1996, p. 15) considers the set of ratings to be the final result of the evaluation.

3. Relation between the Context of Use, Quality Characteristics, and Metrics

Just as one cannot determine "what is the best house?", one cannot expect to determine the best MT system without further specifications. Just like a house, an MT system is intended for certain users, located in specific circumstances, and required for specific functions. Which parameters to pay attention to, and how much weight to assign each one, remains the prerogative of the user/evaluator. The importance of the context for effective system deployment and use has been long understood, and has been a focus of study for MT specifically in the JEIDA report (Nomura, 1992).

3.1. The Context of Use in the ISO/IEC Standards

While a good definition of the context of use is essential for accurate evaluation, in ISO/IEC the context of use plays a somewhat lesser role. The context of use is considered at the beginning of the software's life-cycle (ISO/IEC-14598, 1998 2001, Part 1), and appears in the definition of quality in use. No obvious connection between quality in use metrics and internal or external ones is provided. There is thus no overall indication how to take into account the context of use in evaluating a product.

There are however two interesting mentions of the context of use in ISO/IEC. First, the ISO/IEC standard for acquirers (ISO/IEC-14598, 1998 2001, Part 4, Annex B, pp. 21-22) exemplifies the link between the desired integrity of the evaluated software (integrity pertains to the risk of using the software) and the evaluation activities, in particular the choice of a quality model: for higher integrity, more evaluation procedures have to be fulfilled. The six ISO/IEC 9126 characteristics are also ordered differently according to the required integrity. Second, (ISO/IEC-14598, 1998 2001, Part 5, Annex B, pp. 22-25) gives another relation between "evaluation techniques" and the acceptable risk level. These proposals attempt thus to fill the gap between concrete contexts of use and generic quality models.

3.2. Relating the Context of Use to the Quality Model

When specifying an evaluation, the external evaluator – a person or a group in charge of estimating the quality of MT software – must mainly provide a quality model based on the expected context of use of the software. Guidelines for MT evaluation must therefore contain the following elements:

1. A classification of the main features defining a context of use: the user of the MT system, the task, and the nature of the input to the system.
2. A classification of the MT software quality characteristics, detailed into hierarchies of sub-characteristics and attributes, with internal and/or external attributes (i.e., metrics) at the bottom level. The upper levels coincide with the ISO/IEC 9126 characteristics.
3. A mapping from the first classification to the second, which defines (or at least suggests) the characteristics, sub-characteristics and attributes or metrics that are the most relevant for each context of use.

This broad view of evaluation is still, by comparison to ISO/IEC, focused on the technical aspect of evaluation. Despite the proximity between the taxonomy of contexts of use and quality in use, we do not extend our guidelines to quality in use, since this must be measured fully in context, using metrics that have less to do with MT evaluation than with ergonomics and productivity measures. Therefore, we have proposed elsewhere (Hovy, King and Popescu-Belis, 2002) a formal model of the mapping at point (3) above.

To summarize, building upon the definitions in Section 2.3.3., we consider the set of all possible attributes for MT software $\{A_1, A_2, \ldots, A_n\}$, and the process of evaluation is defined using three stages and the corresponding mappings: m_{Ai} (application of metrics), r_{Ai} (rating of measured value), and α (assessment of ratings).

From this point of view, the correspondence described at point (3) above holds between a context of use and the assessment or averaging function α. Point (3) is thus addressed by providing, for each context of use, the corresponding assessment function, i.e. the function that assigns a greater weight to the attributes relevant to that particular context. In the formal model, α is simplified by choosing a linear selection function.

4. The Contents of the Two Taxonomies

The schema below gives a general view of the contents of the two taxonomies. The first one enumerates non exclusive characteristics of the context of use grouped in three complementary parts (task, user, input). The second one develops the quality model, and its starting point is the six ISO/IEC quality characteristics. The reader will notice that our efforts towards a synthesis have not yet succeeded in unifying internal and external attributes under these six characteristics. As mentioned in Section 2.3.2., the link between internal features and external performance is not yet completely clear for MT systems. So, the internal attributes are structured here in a branch separate from the six ISO/IEC characteristics, which are measured by external metrics.

For lack of space, the hierarchies below represent a brief snapshot of the actual state of our proposal, which may be revised under feedback from the community. The full version available over the Internet (`http://www.issco.unige.ch/projects/isle/taxonomy2`) has about 30 pages, and expands each taxon with the corresponding metrics extracted from the literature. The website provides an interactive version and a printable version of the taxonomy.

- Specifying the context of use
    - Characteristics of the translation task
        - Assimilation
        - Dissemination
        - Communication
    - Characteristics of the user of the MT system
        - Linguistic education
        - Language proficiency in source language
        - Language proficiency in target language
        - Present translation needs
    - Input characteristics (author and text)
        - Document / text type
        - Author characteristics
        - Sources of error in the input
            - Intentional error sources
            - Medium-related error sources
            - Performance-related errors
- Quality characteristics, sub-characteristics and attributes
    - System internal characteristics
        - MT system-specific characteristics (translation process)
        - Model of translation process (rule-based / example-based / statistical / translation memory)
        - Linguistic resources and utilities
        - Characteristics related to the intended mode of use
            - Post-editing or post-translation capacities
            - Pre-editing or pre-translation capacities
            - Vocabulary search
            - User performed dictionary updating
            - Automatic dictionary updating
    - System external characteristics
        - Functionality
            - Suitability (coverage – readability – fluency / style – clarity – terminology)
            - Accuracy (text as a whole – individual sentence level – types of errors)
            - Interoperability
            - Compliance
            - Security
        - Reliability
        - Usability
        - Efficiency
            - Time behavior (production time / speed of translation – reading time – revision and post-editing / correction time)
            - Resource behavior
        - Maintainability
        - Portability
        - Cost

Practical work using the present taxonomy was the object of a series of workshops organized by the

Evaluation Work Group of the ISLE Project. There has been considerable continuity between workshops, with the result that the most recent in the series offered a number of interesting examples of using the taxonomy in practice. A very wide range of topics was covered, including the development of new metrics, investigations into possible correlation between metrics, ways to take into account different user needs, novel scenarios both for the evaluation and for the ultimate use of an MT system and ways to automate MT evaluation. The four workshops took place in October 2000 (at AMTA 2000), April 2001 (stand-alone hands-on workshop at ISSCO, Geneva), June 2001 (at NAACL 2001) and September 2001 (at MT Summit VIII).

Among the first conclusions drawn from the workshops is the fact that evaluators tend to favor some parts of the second taxonomy – especially attributes related to the quality of the output text – and to neglect some others – for instance the definition of a user profile. It appears that the sub-hierarchy related to the "hard problem", i.e. the quality of output text, should be better developed. Sub-characteristics such as the translation quality for noun phrases (which is further on split into several attributes) attracted steady interest.

The proposed taxonomies can be accessed and browsed through a computer interface. The mechanism that supports this function also ensures that the various nodes and leaves of the categories are stored in a common format (based on XML), and simplifies considerably the periodic update of the classifications (Popescu-Belis et al., 2001). A first version of our taxonomies is visible at `http://www.isi.edu/natural-language/mteval` and the second one at `http://www.issco.unige.ch/projects/isle/taxonomy2` – the two sites will soon mirror a third, updated version.

5. Towards the Refinement of the Taxonomies

The taxonomies form but the first step in a larger program – listing the essential parameters of importance to MT evaluation. But for a comprehensive and systematic understanding of the problem, one also has to analyze the nature and results of the actual evaluation measures used. In our current work, a primary focus is the analysis of the measures and metrics: their variation, correlation, expected deviation, reliability, cost to perform, etc. This section outlines first a theoretical framework featuring coherence criteria for the metrics, then lists the (unfortunately very few) examples from previous research.

5.1. Coherence Criteria for Evaluation Metrics

We have defined coherence criteria for NLP evaluation metrics in an EAGLES-based framework (Popescu-Belis, 1999). The following criteria, applied to a case where there is no golden standard to compare a system's response to, enable evaluators to choose the most suitable metric for a given attribute and help them interpret the measures.

A metric m_{Ai} for a given attribute A_i is a function from an abstract 'quality space' onto a numeric interval, say [0,1] or [0%, 100%]. With respect to definition (a) in Section 2.3.3., each system occupies a place in the quality space of A_i, quantified by that metric. Since the goal of evaluators is to quantify the quality level using a metric, they must poll the experts to get an idea of what the best and the worst quality levels are for A_i.

It is often easy to find the best quality of a response, but there are at least two kinds of very poor quality levels: (a) the worst imaginable ones (which a system may rarely actually descend to) and (b) the levels attained by simplistic or baseline systems. For instance, for the capacity to translate polysemous words, a system that always outputs the most frequent sense of source words does far better than the worst possible system (the one that always gets it wrong) or than a random system. Once these limits are identified, the following coherence criteria should be tested for:

- **UL – upper limit**: A metric for an attribute A_i must reach 1 for best quality of a system, and (reciprocally) only reach 1 when the quality is perfect;

- **LL – lower limit:** A metric for an attribute A_i must reach 0 for the worst possible quality of a system, and only reach 0 when the quality is extremely low. Since it is not easy to identify the set of lowest quality cases, one can alternatively check that:
    - receiving a 0 score corresponds to low quality;
    - all the worst quality responses receive a 0 score;
    - the lowest theoretical scores are close or equal to 0 (a necessary condition for the previous requirement).

- **M – monotonicity**: A metric must be monotonic, that is, if the quality of system *A* is higher than that of system *B*, then the score of *A* must be higher than the score of *B*.

One should note that it is difficult to prove that a metric does satisfy these coherence criteria, and much easier to use counter-examples to criticize a measure on the basis of these criteria. Finally, one can also compare two metrics, stating that m_1 is more severe than m_2 if it yields lower scores for each possible quality level.

5.2. Analyzing the Behavior of Measures

Since our taxonomy gathers numerous quality attributes and metrics, there are basic aspects of MT that may be rated through several attributes, and each attribute may be scored using several metrics. This uncomfortable state of affairs calls for investigation. If it should turn out, for a given characteristic, that one specific attribute correlates perfectly with human judgments, subsumes most or all of the other proposed measures, can be expressed easily through one or more metrics, and is cheap to apply, we should have no reason to look further: that aspect of the taxonomy would be settled.

The full list of desiderata for a measure is not immediately clear, but there are some obvious ones. The measure:

- must be easy to define, clear and intuitive;
- must correlate well with human judgments under all conditions, genres, domains, etc.;

- must be `tight', exhibiting as little variance as possible across evaluators, or for equivalent inputs;
- must be cheap to prepare (i.e., not require a great deal of human effort for training data or ideal examples);
- must be cheap to apply;
- should be automated if possible.

Unexpectedly, the literature contains rather few methodological studies of this kind. Few evaluators have bothered to try someone else's measures too, and correlate the results. However, there are some advances. In recent promising work using the DARPA 1994 evaluation results (White et al., 1992 1994), White and Forner have studied the correlation between intelligibility (syntactic fluency) and fidelity (White, 2001) and between fidelity and noun compound translation (Forner and White, 2001). As one would expect with measures focusing on aspects as different as syntax and semantics, some correlation was found, but not a clear one. Papineni et al. (2001) compared the scores given by BLEU, an algorithm mentioned above, with human judgments of the fluency and fidelity of translations. They found a very high level of agreement, with correlation coefficients of 0.99 (with monolingual judges) and 0.96 (bilingual ones).

Another important matter is inter-evaluator agreement, reported on by most careful evaluations. Although the way one formulates instructions has a major effect on subjects' behavior, we still lack guidelines for formulating the instructions for evaluators, and no idea how variations would affect systems' scores. Similarly, we do not know whether a 3-point scale is more effective than a 5- or 7-point. Experiments are needed to determine the optimal point between inter-evaluator consistency (higher on a shorter scale) and evaluation informativeness (higher on a longer scale). Still another important issue is the number of measure points required by each metric before the evaluation can be trusted, a figure that can be inferred from the confidence levels of past evaluation studies.

In the ISLE research we are now embarking on the design of a program that will help address these questions. Our very ambitious goal is to know, for each taxon in the taxonomy, which measure(s) are most appropriate, which metric(s) to use for them, how much work and cost is involved in applying each measure, and what final level of score should be considered acceptable (or not). Armed with this knowledge, a would-be evaluator would be able to make a much more informed selection of what to evaluate and how to go about it.

5.3. A View to the Future

It can be appreciated that building a taxonomy of features is an arduous task, made more difficult by the fact that few external criteria for correctness exist. It is easy to think of features and to create taxonomies; we therefore have several suggestions for taxonomy structure, and it is unfortunately very difficult to argue for the correctness of one against another. We therefore explicitly do not claim in this work that the present taxonomy is correct, complete, or not subject to change. We expect it to grow, to become more refined, and to be the subject of discussion and disagreement – that is the only way in which it will show its relevance. Nonetheless, while it is possible to continue refining the taxonomy, collecting additional references, and classifying additional measures, we feel that the most pressing work is only now being started. The taxonomy is but the first step toward a more comprehensive and systematic understanding of MT evaluation in all its complexity, including a dedicated program of systematic comparison between metrics.

The dream of a magic test that makes everything easy – preferably an automated process – always remains. A recent candidate, proposed by (Papineni et al., 2001), has these desirable characteristics. Should it be true that the method correlates very highly with human judgments, and that it really requires only a handful of expert translations, then we will be spared much work. But we will not be done. For although the existence of a quick and cheap evaluation measure is enough for many people, it still does not cover more than a small portion of the taxonomy; all the other aspects of MT that people have wished to measure in the past remain to be measured.

A general theme running throughout this document is that MT evaluation is simply a special, although rather complex, case of software evaluation in general. An obvious question then is whether the work described here can be extended to other fields. Some previous experience has shown that it applies relatively straightforwardly to some domains, for example, dialogue systems in a specific context of use. However, as the systems to be evaluated grow more complex, the contexts of use become potentially almost infinite. Trying to imagine them all and to draw up a descriptive scheme as we are doing for MT systems becomes a challenging problem, that must be addressed in the future. It is nevertheless our belief that the basic ISO notion of building a quality model and associating appropriate metrics to it should carry over to almost any application.

6. References

AMTA. 1992. MT evaluation: Basis for future directions (Proceedings of a workshop held in San Diego, CA). Technical report, Association for Machine Translation in the Americas (AMTA).

K. W. Church and E. H. Hovy. 1993. Good applications for crummy MT. *Machine Translation*, 8:239-258.

O. Daly-Jones, N. Bevan, and C. Thomas, editors. 1999. *Handbook of User-Centred Design*: INUSE 6.2. http://www.ejeisa.com/nectar/inuse.

J. Doyon, K. Taylor, and J.S. White. 1998. The DARPA MT evaluation methodology: Past and present. In *Proceedings of the AMTA Conference*, Philadelphia, PA.

EAGLES-Evaluation-Workgroup. 1996. EAGLES evaluation of natural language processing systems. Final report, Center for Sprogteknologi, Denmark, October 1996.

M. Flanagan. 1994. Error classification for MT evaluation. In *Proceedings of the AMTA Conference*, Columbia, Maryland.

M. Forner and J.S. White. 2001. Predicting MT fidelity from noun-compound handling. In *Workshop on MT Evaluation "Who did what to whom?" at Mt Summit VIII*, Santiago de Compostela, Spain.

E.H. Hovy. 1999. Toward finely differentiated evaluation metrics for MT. In *EAGLES Workshop on Standards and Evaluation*, Pisa, Italy.

Infoshop. 1999. Language translations: World market overview, current developments and competitive assessment. Technical report, Infoshop Japan, Global Information Inc., Kawasaki, Japan.

ISO/IEC-14598. 1998-2001. *ISO/IEC 14598 – Information technology – Software product evaluation – Part 1: General overview (1999), Part 2: Planning and management (2000), Part 3: Process for developers (2000), Part 4: Process for acquirers (1999), Part 5: Process for evaluators (1998), Part 6: Documentation of evaluation modules (2001)*. ISO/IEC, Geneva.

ISO/IEC-9126-1. 2001. *ISO/IEC 9126-1:2001 (E) – Software engineering – Product quality – Part 1: Quality model*. ISO/IEC, Geneva, June.

ISO/IEC-9126. 1991. *ISO/IEC 9126:1991 (E) – Information Technology – Software Product Evaluation – Quality Characteristics and Guidelines for Their Use*. ISO/IEC, Geneva.

M. Kay. 1980. The proper place of men and machines in language translation. Research Report CSL-80-11, XEROX PARC.

M. King and K. Falkedal. 1990. Using test suites in evaluation of MT systems. In *18th Coling Conference*, volume 2, Helsinki, Finland.

J. Lehrberger and L. Bourbeau. 1988. *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation.* Lingvisticae Investigationes Suppl. 15. John Benjamins, Amsterdam.

J. Mason and A. Rinsche. 1995. Translation technology products. Report, OVUM Ltd.

M. Nagao. 1989. A Japanese view on MT in light of the considerations and recommendations reported by ALPAC, USA. Technical report, Japan Electronic Industry Development Association (JEIDA).

H. Nomura and J. Isahara. 1992. The JEIDA report on MT. In *Workshop on MT Evaluation: Basis for Future Directions*, San Diego, CA. Association for Machine Translation in the Americas (AMTA).

H. Nomura. 1992. JEIDA methodology and criteria on MTevaluation. Technical report, Japan Electronic Industry Development Association (JEIDA).

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. BLEU: a method for automatic evaluation of MT. Research Report, Computer Science RC22176 (W0109- 022), IBM Research Division, T.J.Watson Research Center, 17 September 2001.

J.R. Pierce, J.B. Carroll, E.P. Hamp, D.G. Hays, C.F. Hockett, A.G. Oettinger, and A. Perlis. 1966. Computers in translation and linguistics (ALPAC report). report 1416, National Academy of Sciences / National Research Council, 1966.

A. Popescu-Belis, S. Manzi, and M. King. 2001. Towards a two-stage taxonomy for MT evaluation. In *Workshop on MT Evaluation "Who did what to whom?" at MT Summit VIII*, pages 1-8, Santiago de Compostela, Spain.

A. Popescu-Belis. 1999. Evaluation of natural language processing systems: a model for coherence verification of quality measures. In Marc Blasband and Patrick Paroubek, editors, *A Blueprint for a General Infrastructure for Natural Language Processing Systems Evaluation Using Semi-Automatic Quantitative Black Box Approach in a Multilingual Environment.* ELSE Project LE4-8340 (Evaluation in Language and Speech Engineering).

K. Sparck-Jones and J.R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Lecture Notes in Artificial Intelligence 1083. Springer-Verlag, Berlin / New York.

TEMAA. 1996. TEMAA final report. Technical Report LRE-62-070 (March 1996), Center fo Sprogteknologi, Copenhagen, Danemark, http://www.cst.ku.dk/projects/temaa/D16/d16exp.html.

H. S. Thompson, editor. 1992. *The Strategic Role of Evaluation in Natural Language Processing and Speech Technology* (Record of a workshop sponsored by DANDI, ELSNET and HCRC). University of Edinburgh (Technical Report, May 1992).

G. Van Slype. 1979. Critical study of methods for evaluating the quality of MT. Technical Report BR 19142, European Commission / Directorate for General Scientific and Technical Information Management (DG XIII).

J.S. White et al. 1992-1994. ARPA workshops on MT (series of four workshops on comparative evaluation). Technical report, PRC Inc., McLean, Virginia.

J.S. White. 2001. Predicting intelligibility from fidelity in MT evaluation. In *Workshop on MT Evaluation "Who did what to whom?" at MT Summit VIII*, Santiago de Compostela, Spain.

A Hands-On Study of the Reliability and Coherence of Evaluation Metrics

**Marianne Dabbadie[1], Anthony Hartley[2], Margaret King[3], Keith J. Miller[4],
Widad Mustafa El Hadi[5], Andrei Popescu-Belis[3], Florence Reeder[4], Michelle Vanni[6]**

[1] EVALING, Paris (France)
[2] Centre for Translation Studies, University of Leeds (UK)
[3] ISSCO/TIM/ETI, University of Geneva (Switzerland)
[4] The MITRE Corporation (USA)
[5] Université Lille III - Charles de Gaulle (France)
[6] U.S. Department of Defense (USA)

Abstract

This section of the workbook provides the description of the MT evaluation exercise that is proposed to the workshop participants, including the specification of the metrics for MT evaluation that the participants are suggested to use at the workshop.

1. A Collective Hands-on Exercise

1.1. Motivation

The motivations behind the LREC 2002 MT Evaluation workshop are grounded in previous work in the field, described at length in the previous section. The workshop is the sixth in a series of hands-on workshops on MT Evaluation, organized in the framework of the ISLE Project.

The goal of these hands-on evaluation workshops is to carry on a collective effort towards the standardization of MT evaluation. The ISLE taxonomy has been designed for standardization, but it would have not reached the present state without feedback from the participants at the workshops. Conversely, the participants have broadened their view of MT Evaluation, through the concrete use of the ISLE taxonomy for the design of toy evaluations, but also through extensive discussions with the organizers and other participants.

Some of the workshops have focused more on the setup of an evaluation depending on the desired context of use, others on metrics, others on reporting results obtained in this framework. As pointed out in the previous section, the need for a clear view of the performances of various metrics has prompted the organization of the present workshop, "Machine Translation Evaluation: Human Evaluators Meet Automated Metrics". Through hands-on application of selected metrics from the present workbook, the participants will be able to familiarize themselves with the current problems of MT Evaluation, to get a first-hand experience with recent metrics and to contribute to research in this field by their own observations of the metrics' behaviors.

1.2. Description of the exercise

The participants to the workshop are suggested to register with the organizers well before the day the workshop will take place (May 27, 2002). Thus, both organizers and participants will be able to prepare in advance an evaluation exercise (requiring several hours of work), so that the workshop itself can be devoted to the exploitation of those results.

The evaluation study that all participants are kindly required to carry on can be summarized as follows:

1. Select two evaluation metrics among those described below, preferably one "human-based" and one "automated" (more than two is welcome!).

2. Optionally, add one of the metrics that you have used before in MT evaluation, or any personal suggestion for a metric.

3. Using the test data provided by the organizers, apply the selected metrics and compute the scores of each translation, on a 0%–100% scale.
The test data is described in the next document of the workbook and can be downloaded from *http://www.issco.unige.ch/projects/isle/mteval-may02/*. It consists in two source texts in French, each with a reference translation and about a dozen translations to be evaluated, from various systems and humans.

4. Send the results by email to the organizers (e.g., `Andrei.Popescu-Belis@issco.unige.ch`), together with any comments you believe useful.

5. Prepare a brief account of the evaluation (about 10–15 minute talk) to be presented at the workshop, for instance by first answering the question "what are the strongest and the weakest points in the measures that you used?"

1.3. Exploitation of the Results

The results of these evaluations will be discussed and highlighted at the workshop from the perspective of present research goals. Regarding individual metrics, the scores obtained by different evaluators using the same metric will inform the community about the reliability of that metric (cf. preceding document, 5.2), by computing standard deviation and inter-annotator agreement.

The other important result of the pre-workshop evaluations will be data on cross-metric correlation, i.e. the agreement between pairs of metrics. This is important both for metrics based on human judges (it illustrates how well the specifications are defined or how coherent the judges are) and for automated metrics (for which agreement with a reliable human judgement is almost the only proof of coherence). These meta-evaluation considerations will be analyzed at the workshop by the organizers, based on the results sent to them by the participants. These considerations will constitute the basis for discussion and conclusions of the workshop.

2. Specifications of the Metrics

2.1. Preamble

The metrics that are proposed in this application illustrate a broad spectrum of those that were synthesized for the ISLE MT evaluation framework. The two categories identified below parallel of course the title of the workshop, "Human Evaluators Meet Automated Metrics". In the history of MT evaluation, given the difficulty of the task, most of the quality judgments, and later 'metrics', we carried on by humans. However, as explained in the previous chapter, the utility of automatic measures has always been clear: they provide cheap, quick, repeatable and objective evaluation. 'Objective' means here that the same translation will always receive the same score, as opposed to human judges that may have fluctuating opinions. However, since human judges are the final reference in MT evaluation, the results of automated metrics must correlate well with (some aspect of) human-based metrics.

The metrics specified below must of course be integrated in a broader view of evaluation, since none of them is sufficient to determine the overall quality of a system. As stated in the ISLE taxonomy, it is the desired context of use of the evaluated system that determines a 'quality model', namely a set of useful features, to which several metrics are associated. It is only the combination of these scores that provides a good view of the quality of the system in the given context.

Documentation about the metrics below (apart from the references quoted) can be found in several papers available over the Internet. The ISLE evaluation workgroup has a webpage at `http://www.issco.unige.ch/projects/isle/ewg.html`, with links to previous workshop material for MT Evaluation, and to electronic versions of Van Slype's (1979) report and of the MT Evaluation workshop held at the MT Summit VIII conference. The ISLE taxonomy can be found at `http://www.issco.unige.ch/projects/isle/taxonomy2/`.

Below is a synopsis of the metrics that will be described in the remaining part of this document.

(A1)	IBM's BLEU and the NIST version
(A2)	EvalTrans
(A3)	Named entity translation
(A4a)	Syntactic correctness
(A4b)	X-Score / parsability
(A5a)	Dictionary update / number of untranslated words
(A5b)	Translation of domain terminology
(A6)	Evaluating syntactic correctness from the implementation of transfer rules
(H1)	Reading time
(H2)	Correction / post-editing time
(H3)	Cloze test
(H4a)	Intelligibility / fluency
(H4b)	Clarity
(H5)	Correctness / adequacy / fidelity
(H6)	Informativeness: comprehension task

2.2. Automated/automatable metrics

2.2.1. IBM's BLEU and the NIST version (A1)

We mention first the most recent proposal of an automated metric for MT Evaluation, namely the BLEU algorithm proposed by a team from IBM (Papineni et al., 2001; Papineni, 2002). The principle of this metric, which was fully implemented, is to compute a distance between the candidate translation and a corpus of human "reference" translations of the source text. The distance is computed averaging *n*-gram similitude between texts, for $n = 1, 2, 3$ (higher values do not seem relevant). That is, if the words of the candidate translation, the bi-grams (couples of consecutive words) and tri-grams are close to one or more of those in the reference translations, then the candidate scores high on the BLEU metric.

Apart from intuitive arguments, the method to find out whether this metric really reflects translation quality is to compare its results with human judgements, on the same texts. In-house data (Papineni et al., 2001), as well as the DARPA 1994 data (Papineni et al., 2002), were used to test the coherence between human scores and BLEU scores, and this was found acceptable.

The metric was also adapted for the recent NIST MT Evaluation campaign (Doddington, 2001). The main changes were: text preprocessing, a differentiated weight associated to N-grams based on their frequency, and the use of tri-grams only. These modifications must still be discussed by the community, but the NIST provides yet the scripts implementing the BLEU metric as well as its adaptation, at: `http://www.nist.gov/speech/tests/mt/mt2001/resource/`.

We do not describe further this metric, but would like to refer the participants to the documentation quoted above, which provides enough resources to apply it.

2.2.2. EvalTrans (A2)

Automatic corpus evaluation extrapolation using EvalTrans (Niessen et al., 2000) gives statistics, such as the average Levenshtein distance standardized to the length of the target sentence. The tool can be downloaded at `http://www-i6.informatik.rwth-aachen.de/HTML/Forschung/Uebersetzung/Evaluation/`.

The first step is to load and save the human translations. For the present workshop, the reference translation as well as the other human translations of the same source text will constitute the "reference set". When the system is set up to work automatically, it will search this reference database for sentences which are most similar to the machine translated sentence that must be scored.

However, in order for the extrapolation to be performed, the Levenshtein distance algorithm needs to be seeded with scores for some (at least one) manually evaluated sentence. For this, a baseline machine translation (for instance) needs to be loaded and some sentence pairs need to be evaluated.

Next, the "test corpus" sentences need to be loaded. These are the machine translations for each source text. For each set of "test corpus" sentences, which comprise each machine translation of a source text, subjective sentence error rate (SSER) and multi-reference word error rate (mWER) will be calculated by the automatic metric.

- Several statistics of interest will be produced:
- Average number of "perfect" (scored 10) reference sentences per evaluation sentence pair (to indicate how reliable the mWER is).
- (average-score) / (value of all (evaluated/ extrapolated) sentence pairs)
- Standard deviation of the score
- Subjective sentence error rate (i.e., 100% * (1 – average-score)). An average score of 0.0 results in a SSER of 100%, an average score of 10.0 in a SSER of 0%.
- Subjective sentence error rate weighted by the length of the target sentences
- Average extrapolation distance: average Levenshtein distance (per target word) of all extrapolated sentences

The SSER indexes each sentence, then uses the mWER, the number of perfect reference sentences, the absolute Levenshtein distance to each sentence, and the Levenshtein distance to that sentence v. the length of current sentence.

The mWER is the word error rate against the most similar reference sentence which has been evaluated as "perfect" (i.e., has been assigned a score of ten). It is calculated as Levenshtein operations per reference word (and can thus exceed 100%). Average mWER for an evaluation corpus is calculated word-wise, not sentence-wise.

Another measure, the information item error rate, is not included because it relies heavily on manual scores, use of which would defeat the purpose of the automated metric.

2.2.3. Named entity translation (A3)

The NEE metric (Named Entity Evaluation) is described for instance in (Reeder et al., 2001). Since automated software to support this metric is available, it has been considered here an automated metric. Participants to the workshop may of course apply it manually, given the small amount of test data.

The process for utilizing this metric is relatively straightforward: a) identify the named entities within a given test corpus; b) pull unique entities from the document; c) find the entities in the system output text; and d) compare entities in the output text with those identified in the reference text (see Figure 1 below). Identifying the named entities in the reference translation requires human annotation, and is the only stage of the process to do so.

In a concrete example of this metric, to prepare the corpora for evaluation, two expert annotators used the Alembic Workbench (Day et al., 1997; see also `http://www.mitre.org/technology/alembic-workbench/`) annotation tool to tag occurrences of named entities according to the MUC annotation guidelines. After the named entities are tagged in the reference translation (designated here by ANNO), the metric can be applied.

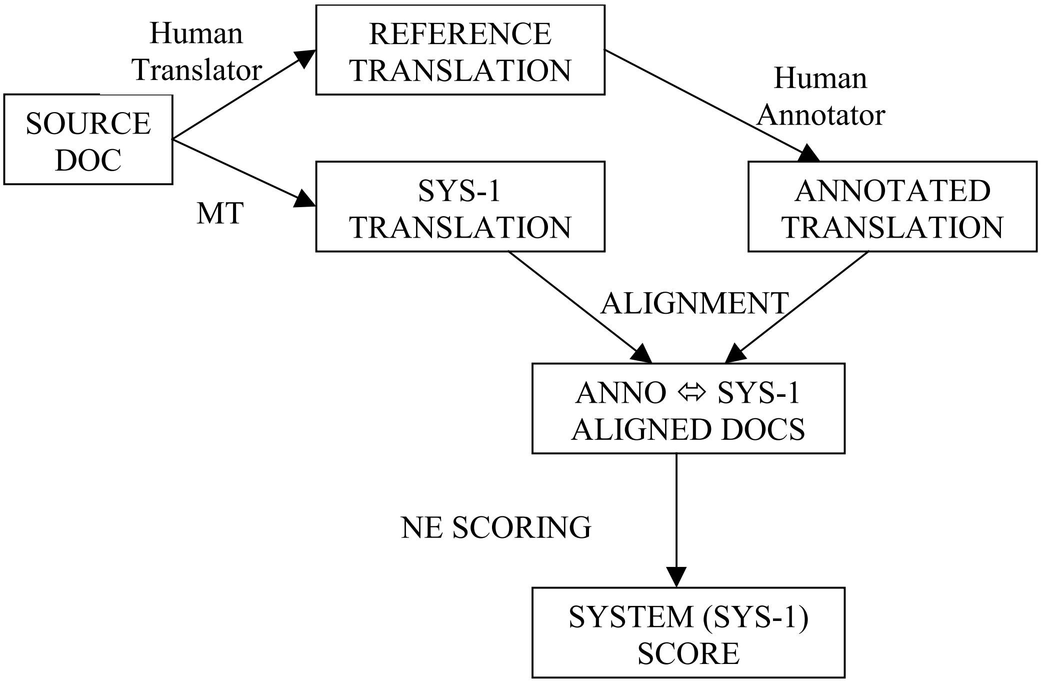

Figure 1. Scoring technique for the NEE metric.

The next stage is to align the ANNO translation text with the evaluation text (the output of the system SYS-1 for this example). To score the translation, for each article in the aligned pair, the tagged named entities are pulled from the ANNO and a list of unique names for the comparison unit (paragraph or article) is prepared. This is followed by normalization. At this time, the normalization steps applied are: (a) substitution of non-diacritic marked letters for the equivalent diacritic mark character for Romance languages (for instance **ã** becomes **a**); (b) down-casing; (c) the normalization of numeric quantities (particularly for numbers under 100) and (d) the removal of possessives. Other normalization steps may be needed, as well as the incorporation of partial match scoring (see Reeder et al., 2001). Once the named entity list and the SYS-1 tokens have been normalized, the search for named entities in the token lists is straightforward. Only exact matches given the normalization steps described are considered at this time and all results here reflect this.

2.2.4. Syntactic correctness (A4a)

The following describes a syntax metric based on the minimal number of corrections necessary to render an MT output sentence grammatical. Each evaluator must transform each sentence in the MT output into a grammatical sentence by making the minimum number of replacements, corrections, rearrangements, deletions, or additions possible. The syntax score for each sentence is then defined as the ratio of the number of changes for each sentence to the number of tokens in the sentence. For the purposes of this test, a token is defined as a whitespace-delimited string of letters or numbers. Additionally, individual punctuation marks, since they are subject to correction, are also counted as separate tokens. Each item of punctuation that occurs in pairs (e.g. brackets, braces, quotation marks, parenthesis) is counted as a separate token. Thus, in the following sentence, there are 24 tokens:

- *Mary, who had gone to see the fountain (in the center of town), said that it was turned off.*

It is important to remember that the final edited sentence need only be syntactically correct. That is, the final result may be semantically anomalous. Raters should endeavor to produce a syntactically correct sentence by making as few changes possible to the original MT output. Deletions, substitutions, additions, and rearrangements are counted by totaling the number of words deleted, substituted, added, or moved. In the event that there are combined operations, for example, moving a phrase consisting of four words, of which one has been deleted, the move is computed *after* the deletion is counted, thus the above-mentioned operation would result in one deletion and 3 moves. Finally, errors in inflectional morphology are not counted in the syntax metric. In applying this metric to test data, it was found that even when evaluators arrive at the same score for a given sentence (that is, they have the same total number of changes), they often choose a different combination of the four operations to arrive at their final grammatical sentence. The metric as it stands has not been automated, and would indeed be very difficult to automate; however, partial automation, such as automatic tracking and counting of necessary edit operations, would greatly assist in applying this metric in an efficient manner.

2.2.5. Automatic Ranking of MT Systems by X-Score (A4b)

Background: The X-Score metric aims to rank MT systems in the same order as would be given by a human evaluation of the Fluency of their outputs (Hartley & Rajman, 2001; Rajman & Hartley, 2002). The metric is especially adapted to rank machine translations relative to one another, rather than comparing human and machine translations. This metric was derived from experiments conducted on the French-English segment of the corpus used in the 1994 DARPA MT evaluation exercise. In that exercise, human evaluators scored translations of 100 source texts by 5 MT systems for their Fluency (among other attributes). To establish the present metric, the F-scores (Fluency scores) for individual texts were converted into rankings of systems using the aggregation technique of ranking by average ranks (average rank ranking or ARR). Using the same ARR technique, rankings were computed on the basis of the X-score for each document. The X-scores were found to represent a very good predictor of the ranking derived from the human evaluations (H-rankings). The distance between the H-ranking and the X-ranking is 1, corresponding to a similarity of 93.3%, a precision of 93,3% and a recall of 93.3%. If restricted to the most complete partial ranking, these values improve to a distance of 0.5, a similarity of 96.7%, a precision of 100% and a recall of 93.3%.

Computing the X-Score: The X-score is taken to measure the grammaticality of the translations. For any given document, the X-score is obtained as follows. First, the document is analyzed by the Xerox shallow parser XELDA in order to produce the syntactic dependencies for each sentence constituent. For example, for the sentence The Ministry of Foreign Affairs echoed this view, the following syntactic dependencies are produced: SUBJ (Ministry, echoed); DOBJ (echoed, view); NN (Foreign, Affairs); NNPREP (Ministry, of, Affairs).

On the corpus used in (Hartley & Rajman, 2001), XELDA produced 22 different syntactic dependencies, among which:

- RELSUBJ: for example, RELSUBJ(hearing, lasted) in “a hearing that lasted more than two hours”;
- RELSUBJPASS: for example, RELSUBJPASS(program, agreed) in “a public program that has already been agreed on ...”;
- PADJ: for example, PADJ(effects, possible) in “to examine the effects as possible”;
- ADVADJ: for example, ADVADJ(brightly, colored) in “brightly colored doors”.

After each document has been parsed, we compute its dependency profile (i.e. the number of occurrences of each of the 22 dependencies in the document). This profile is then used to derive the X-score using the following formula:

- *X-score = (#RELSUBJ + #RELSUBJPASS – #PADJ – #ADVADJ)*

Note that several formulae would have been possible for computing the X-scores. The above-mentioned one was selected in such a way that, if applied to the average dependency profile, it correctly predicted the average rank ranking (ARR) derived from the F-scores. In this sense, one can say that the computation of the X-score was specifically tuned to the test data and so it was considered quite ad hoc in (Hartley & Rajman, 2001). However, this is not true of (Rajman & Hartley, 2002). This second experiment retained exactly the same formula for the X-scores, while completely changing the human evaluations – evaluators directly assigned rankings to series of translations instead of assigning individual scores to each of the translations. Moreover, a new MT system was added, not present at all in the data that was used for the tuning. Thus, there is no reason to believe the X-scores to be ad hoc, which strongly increases their chances of being highly portable to other experimental data.

Computing the Rankings: For each of the documents, the scores of the systems are first transformed into ranks and the average ranks obtained by the systems over all the documents are then used to produce the final ranking.

2.2.6. Dictionary update (A5a) and domain terminology (A5b)

Dictionary update (also known as *non-translated or untranslated words*) and *domain terminology* are two potentially automatable metrics. Although related, these two metrics are not identical, as can be seen from their descriptions below. There are many ways in which a dictionary update measure could be calculated, but it seems obvious to use two objective and easy to observe features of MT output:

- the number of words not translated;
- the number of domain-specific words that are correctly translated.

It is these two features that have been described in previous related work, including (Vanni & Miller, 2002), and that will be specified below.

2.2.7. Number of untranslated words (A5a)

This metric makes use only of the target text. It is based on the intuition that translation quality is linked to size of vocabulary. In its simplest form, the number of words left untranslated is counted. By untranslated, we mean simply that a word which should be translated is not, and is simply copied over untouched into the target text. (This reflects the behavior of many machine translation systems). There are, of course, words which should not be translated (most proper names are a good example): not translating these items is not counted as an error. A score is obtained by the following calculation:

- (number-of-untranslated-words) / (total-number-of-words-in-text) x 100 = percentage-of- untranslated-words… *high is bad*

One possible way to automate this metric would be to run a spelling checker over the target text and count the number of mistakes found. This would, of course, pick up any spelling mistakes in translated words which might exist, as well as finding words which were not legal words of the target language; however, this amount is probably low for translations programs, which generate words based on valid dictionaries. On the whole, this automatic measure might not invalidate the metric as an indicator of overall translation quality.

In discussing the automation of this measure, it is worth noting that some MT systems provide as ancillary output statistics concerning the numbers of untranslated words in the output. However, this is not the case for all systems. In these cases, other automated means must be developed for computing this measure. In cases of languages using a non-Roman script or containing characters outside the standard lower-ASCII range found in typical English text, one possible way of counting non-translated words (for systems that simply pass untranslated words through in the translation) would be to locate and count tokens containing these characters that do not appear in English text. However, even in the case of the Japanese-English systems, some systems did produce a romanization of the untranslated words, and did not leave them in the native script. The romanizations contained only characters found in the lower portion of ASCII.

Given that this metric is intended to compute the number of words that the MT system was unable to translate, another possibility would be to use a tool such as *ispell* in order to identify non-English strings within the output translation. Counting these strings and comparing with the output of a utility such as *wc* (Unix word count) could provide a ratio of untranslated words in the output text.

Two potential problems with this last approach could both lead to undercounting the number of untranslated words in a text. First, included in the untranslated word count for Japanese – English translation were Japanese particles and other bits of non-English material, which may or may not have been the result of romanization of text found in the source. Examples of this include *na*, *re*, *X*, and *inu*. Another Japanese particle, *no*, did not appear in this context in the translation, but had we relied on an automated spelling-based identification of untranslated words, words like *no*, which also happen to be valid English strings (although with a different meaning) would be left uncounted. Secondly, untranslated word scores would likewise be affected for languages that share a high number of cognates with English. For these languages, the string in the source and target language may be identical, and thus not counted as an untranslated word, regardless of whether the system actually translated the word or simply passed it through.

The application of this metric to translations produced by human translators is somewhat doubtful: human translators when faced by a gap in their lexical knowledge try to work round the problem, and do not, normally, simply transcribe the problematic word or leave a gap. It is possible though that the spelling mistake variation might be informative.

It is also worth noting that while untranslated words certainly have an impact on the usability of MT output, such output often contains sentences that are completely unintelligible, but in no way due to untranslated words. Thus, this test should clearly not be used in isolation to provide a picture of overall MT quality, whether quality is defined along the lines of clarity, fluency, adequacy, or coherence.

2.2.8. Translation of Domain Terminology (A5b)

The domain terminology score is calculated as the percentage of correctly translated pre-identified domain terms. The procedure for this test is as follows: First, a list of key term translations is extracted from the human translation. To accomplish this, raters individually select key terms from the human translation, and then the separate key term lists are reconciled before application of the test to the MT systems' output. This step is amenable to automation, but has not as yet been automated. During the test application, systems receive a point for each term for which the translation matches the human translation exactly, and no point otherwise. The final score is the percentage of exactly-matched translations of key terms.

There are two divergent directions in which this test could be developed in the future. First, it could be made more sensitive to acceptable variation in translation of key terms by application of the ACME Cloze test methodology as described for instance in Miller (2000). This methodology simulates basing lexical tests on multiple human translation, while sufficiently constraining the structure of the translation to enable automated comparison.

2.2.9. Evaluating syntactic correctness from the implementation of transfer rules (A6)

This metric proposal is the result of two previous studies. In the first former study, the authors chose to count the number of NPs (noun phrases) and VPs (verb phrases) in source text and target texts, a first indication being given by non parallel data (Mustafa El Hadi, Timimi, Dabbadie, 2001). Another study presented the results on the same corpus after terminological enrichment (Mustafa El Hadi, Timimi, Dabbadie, 2002).

Nevertheless, the use of finer grained criteria such as adjectives or prepositional phrases count could also be envisaged. Any overlap of this threshold might then be considered as an indication that MT system may have failed to analyze source syntactic structure and that therefore, the initial figures require further analysis. But this methodology is still imprecise and limited to a first indication of MT system's analysis failure, when a gap is observed on non parallel data. The use of this methodology also implies that the test is carried out on relatively syntactically isomorphic languages such as French and English. A methodology including a test tool that would implement source and target transfer rules might probably prove more accurate and also apply to non isomorphic languages.

We propose here the following steps for the application of the metrics:

1. Deduce a set of French / English transfer rules from the source text and the reference translation (this part involves manual processing).
2. Write a script (e.g., in Java or Perl) to implement these rules (if not, go to point n. 3)
3. Check that these rules apply through the various candidate translations from the test data (automatically with the script or manually).
4. Generate an output failure file (or else carry out a manual check) and work out syntactic correctness.

2.3. Human-based measures

2.3.1. Reading time (H1)

Reading time can be defined in one of two ways: oral reading time or closed reading time.

Oral reading time (Van Slype, 1979) tends to measure more closely with intelligibility and also tends to be more relevant to higher quality translations. Therefore, for each document, the evaluators should read out loud the first paragraph and time the length of time that it takes to read each sample. The number of words then can be used to calculate a words per minute (WPM) rate:

- *WPM = number-of-words / reading-time*

The closer the WPM rate is to the WPM of natural language (depending on the evaluator), the higher is the quality of the translation (on a scale to be defined by each participant).

Closed reading time relates to the amount of time that a user needs to read a document to a "sufficient" level of understanding. The sufficient level is often paired with other measurements such as comprehension score on a test. Still, the instructions can be given that the readers measure the amount of time necessary to arrive at an understanding they consider to be sufficient to answer basic questions about the text. Words-per-minute rate can be calculated in the same way.

2.3.2. Correction / post-editing time (H2)

This metric is based on the intuition that the time required to produce an acceptable translation from a raw translation (whether produced by a human or by a machine) is inversely proportional to the overall quality of the raw translation.

It can be measured fairly easily by noting when the person responsible for the revision/post-editing starts their task and when they finish it, normalizing the result by taking into account the size of the text measured in words, then multiplying by a fixed factor in order to obtain a number on a wider scale. For this exercise, the following calculation is suggested:

- (number-of-minutes-spent-in-correction) / (total-number-of-words-in-text) x 10 = correction-time… *high is bad*

Note that this metric can only sensibly be applied to a whole text: timing correction to smaller text elements is both annoying for the person doing the timing and difficult to do reliably.

A variation on this metric is to count not the overall time but the number of key strokes made by the corrector.

It should be noted that this metric is somewhat problematic both with respect to validity and reliability for a number of reasons:

- The amount of correction needed depends in part on the ultimate use to which the translation will be put: a text destined for publication will probably be treated with more care than a text intended for information assimilation, for example
- The errors corrected differ in their nature. There will be straightforward grammatical or lexical errors, as well as more complicated stylistic errors. This will affect the amount of time needed to carry out the correction. This would not matter

so much if those doing the correction always agreed on what corrections are needed. But, inevitably, where matters of style are concerned, no such agreement exists.

- There is considerable variety amongst correctors and the way they work. Some work quickly and decisively, others are more hesitant and sometimes change their minds.
- Correctors may be influenced by knowing whether they are dealing with a human produced translation or a machine produced translation. One anecdote tells of correctors correcting far more on machine produced translation but spending comparatively less time in doing so because they felt no need to take into account the computer's feelings.

Participants who choose to work with this metric are invited to reflect on these issues and on possible improvements to the simple metric defined here.

2.3.3. Cloze test (H3)

This metric is reported by Van Slype (1979) as a test of readability. It may however also be thought of as a test of fidelity or of intelligibility, since it is based on the ability of a reader to supply a missing word correctly, which intuitively relates both to readability and intelligibility when the target text alone is considered and to fidelity when the source text is taken into account.

The method is simple. Every *n*-th word in the translation is deleted (in the Van Slype Report (1979), $n = 8$, but other values appear also in the literature). The translation is then given to a group of readers, who are asked to supply the missing words. Two scores are normally computed, one based on the number of answers which comprise exactly the suppressed original word, the other based on the number of answers with a word close in meaning to the original word. The second score has to be interpreted partly in the light of the first score

- (number-of-exact-answers) / (number-of-deleted-items) x 100 = percentage-of-exact-items-supplied… *high is good*

- (number-of-close-answers) / (number-of-deleted-items – number-of-exact-items-supplied) x 100 = percentage-of-close-items-supplied… *high is good*

A possible weakness of this metric is that it potentially also tests the intelligence and wealth of vocabulary of the reader supplying the missing words. This weakness can be mitigated by controlling the size and type of the group of readers.

A second possible weakness appears if the translated text is technical in nature: the readers have to have sufficient knowledge of the subject matter to make it plausible that they should be able to supply the missing items.

Van Slype (1979) also points out that some texts are more redundant than others in the way they carry information, and that if translations of several texts are to be compared, it is important to take this factor into account. He suggests that this can be done by carrying out a Cloze test also on the original text.

2.3.4. Intelligibility / fluency (H4a)

Intelligibility is one of the most frequently used metrics of the quality of output. Numerous definitions (or protocols for measuring it) have been proposed for it, for instance in Van Slype's report or in the DARPA 1994 evaluations. We outline here the definition proposed by T.C. Halliday in (Van Slype, 1979, p. 70), which measures intelligibility on a 4-point scale (0 to 3).

Intelligibility or comprehensibility expresses how intelligible is the output of a translation device under different conditions (for instance, when the sentence fragments are translated while being entered, or after each sentence). Comprehensibility reflects the degree to which a complete translation can be understood. Intelligibility can be based on the general clarity of translation, or the output can be considered in its entirety or by segments out of context.

The following scale of intelligibility has been proposed, from 3 to 0, 3 being the most intelligible:

- 3 – Very intelligible: all the content of the message is comprehensible, even if there are errors of style and/or of spelling, and if certain words are missing, or are badly translated, but close to the target language.
- 2 – Fairly intelligible: the major part of the message passes.
- 1 – Barely intelligible: a part only of the content is understandable, representing less than 50% of the message.
- 0 – Unintelligible: nothing or almost nothing of the message is comprehensible

To apply the metric, the following steps are suggested:

1. Take the reference translation of a text (or the source if you are proficient in that language).
2. Separate and number the sentences.
3. Take a candidate translation and do the operation (2) on it. Match sentences with those in the reference/source translation.
4. Rate sentences from the candidate translation using the 0 to 3 scale described above.
5. Optional: to normalize scores, calculate intelligibility on a 0% to 100% scale, by averaging sentence ratings over the whole text.
6. Produce a final score for each translation

2.3.5. Clarity (H4b)

In work described in (Vanni & Miller, 2002) a metric called *clarity* is proposed that merges the ISLE categories of comprehensibility, readability, style, and clarity into a single evaluation feature. This measure ranges between 0 and 3. Raters are tasked with assigning a *clarity* score to each sentence according to the following criteria:

Score	Criterion
3	meaning of sentence is perfectly clear on first reading
2	meaning of sentence is clear only after some reflection
1	some, although not all, meaning is able to be gleaned from the sentence with some

effort

0 Meaning of sentence is not apparent, even after some reflection

Since the feature of interest is clarity and not fidelity, it is sufficient that some clear meaning is expressed by the sentence and not that that meaning reflect the meaning of the input text. Thus, no reference to the source text or reference translation is permitted. Likewise, for this measure, the sentence need neither make sense in the context of the rest of the text nor be grammatically well-formed, since these features of the text would be measured by tests proposed elsewhere, namely the *coherence* and *syntax* tests, respectively. Thus, the clarity score for a sentence is basically a snap judgement of the degree to which some discernible meaning is conveyed by that sentence.

2.3.6. Correctness / adequacy / fidelity (H5)

This evaluation metric reprises the DARPA 1994 *adequacy* test (Doyon, Taylor, and White, 1996). As with that test, the reference translation or "authority version" is placed next to each of the translations of the source text, to be used as a comparison against each one, human or machine. Before the test is performed, both the "authority version" as well as each of translations should be segmented, with each text separated into sentence fragments to appear next to the corresponding fragment in the translation.

Once each translation is lined up with its equivalent, evaluators grade each unit on a scale of one to five, where five represents a paragraph containing all of the meaning expressed in the corresponding text. The *Adequacy* scale is as follows:

- 5 – All meaning expressed in the source fragment appears in the translation fragment
- 4 – Most of the source fragment meaning is expressed in the translation fragment
- 3 – Much of the source fragment meaning is expressed in the translation fragment
- 2 – Little of the source fragment meaning is expressed in the translation fragment
- 1 – None of the meaning expressed in the source fragment is expressed in the translation fragment

2.3.7. Informativeness: comprehension task (H6)

There are two methods for testing comprehension. The most common of these is the reading comprehension exam (e.g., Somers & Prieto-Alvarez, 2000; DARPA-94; Tomita 1992). In this case, the evaluators design a set of questions, usually under 10, for the given texts. Sometimes, as in the case of Tomita, these tests are structured first and then applied to the translations. Tomita began with the Test of English as a Foreign Language (TOEFL) examinations which he then translated to Japanese and had students take. The theory being that the better scores on the exam will have resulted from the better translations. The big difficulty (Somers & Prieto-Alvarez, 2000) is that it is difficult to test only the reading without bringing a large amount of pre-existing world knowledge to the table. In addition, the design and structuring of such examinations is an art in and of itself.

The second method for a comprehension test takes instead the task of figuring out the kinds of questions that one might want to be able to answer from a translation and determining whether the translation can support answering said questions. For instance, one might want to know the people, places and organizations mentioned in an article. This is covered by the named entity metric. Yet, it is really only the first stage of measurement. The secondary measure would be to look to determine if the entity relationships are also preserved by the translation - that is, who belongs to what organization or who did what to whom. This is the question we began to study at MT Evaluation workshop organized at NAACL 2001, when we asked participants to fill in templates based on specific kinds of questions. The better systems would enable the successful template filling and scoring would follow Message Understanding (MUC) guidelines. It is this type of exercise you will be asked to do at this time. The previously identified named entities will be used here. You will fill out templates to answer specific details of events or relationships between parties.

3. References

D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, and M. Vilain. 1997. Mixed-Initiative Development of Language Processing Systems. In *Fifth Conference on Applied Natural Language Processing*, Washington, D.C.

G. Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *HLT 2002, Human Language Technology Conference*, San Diego, CA.

J. Doyon, K. Taylor, and J.S. White. 1998. The DARPA MT Evaluation Methodology: Past and Present. In *Proceedings of the AMTA Conference*, Philadelphia, PA.

A. Hartley and M. Rajman. 2001. Automatically Predicting MT Systems Rankings Compatible with Fluency, Adequacy or Informativeness Acores. In *Workshop on MT Evaluation "Who did what to whom?" at MT Summit VIII*, Santiago de Compostela, Spain, p.29-34. See `http://www.issco.unige.ch/projects/isle/MT-Summit-wsp.html`.

K. J. Miller. 2000. *The Machine Translation of Prepositional Phrases*. Unpublished PhD Dissertation. Georgetown University. Washington, DC.

W. Mustafa El Hadi, I. Timimi and M. Dabbadie. 2001. Setting a Methodology for Machine Translation Evaluation. In *Workshop on MT Evaluation "Who did what to whom?" at MT Summit VIII*, Santiago de Compostela, Spain, p.49-54. See `http://www.issco.unige.ch/projects/isle/MT-Summit-wsp.html`.

W. Mustafa El Hadi, I. Timimi, and M. Dabbadie. 2002. Terminological Enrichment for non-Interactive MT Evaluation. In *LREC 2002, Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

S. Niessen, F.J. Och, G. Leusch, H. Ney. 2000 An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *LREC 2000, 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, pp. 39-45.

K. Papineni. 2002. Machine Translation Evaluation: N-grams to the Rescue. In *LREC 2002, Third*

International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain.

K. Papineni, S. Roukos, T. Ward, J. Henderson, and F. Reeder. 2002. Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results. In *HLT 2002, Human Language Technology Conference*, San Diego, CA.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. BLEU: a method for automatic evaluation of MT. Research Report, Computer Science RC22176 (W0109- 022), IBM Research Division, T.J.Watson Research Center, 17 September 2001. See `http://domino.watson.ibm.com/library/CyberDig.nsf/home`, and search for 'RC22176'.

M. Rajman and A. Hartley. 2002. Automatic Ranking of MT Systems In *LREC 2002, Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

F. Reeder, K.J. Miller, J. Doyon, and J.S. White. 2001. The Naming of Things and the Confusion of Tongues: an MT Metric. In *Workshop on MT Evaluation "Who did what to whom?" at MT Summit VIII*, Santiago de Compostela, Spain, p.55-59. See `http://www.issco.unige.ch/projects/isle/MT-Summit-wsp.html`.

H. Somers and N. Prieto-Alvarez. 2000. Multiple Choice Reading Comprehension Tests for Comparative Evaluation of MT Systems. In *Workshop on MT Evaluation at AMTA-2000.*

M. Tomita. 1992. Application of the TOEFL Test to the Evaluation of Japanese-English MT. In *MT Evaluation Workshop at AAMT.*

G. Van Slype. 1979. Critical Study of Methods for Evaluating the Quality of Machine Translation. Technical Report BR19142, Bureau Marcel van Dijk / European Commission (DG XIII), Brussels. See `http://issco-www.unige.ch/projects/isle/van-slype.pdf`.

M. Vanni and K. J. Miller. 2002. Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Evaluation Metrics across Languages. In *LREC 2002, Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain..

Towards a corpus of corrected human translations

Andrei Popescu-Belis, Margaret King, Houcine Benantar

ISSCO/TIM/ETI, University of Geneva,
École de Traduction et d'Interprétation
40 Bvd. du Pont d'Arve
CH-1211 Geneva 4 – Switzerland
andrei.popescu-belis@issco.unige.ch
margaret.king@issco.unige.ch
benanta4@etu.unige.ch

Abstract

This section of the workbook describes the test data that is proposed to the participants. The data is part of a broader-scope corpus containing translations produced by students and corrected by their professors. Such a corpus will be used in automatic evaluation of MT systems. This section describes the structure of the corpus and provides some sample data. The full workshop data can be downloaded from: `http://www.issco.unige.ch/projects/isle/mteval-may02/`.

1. Introduction

Several automatic measures for MT evaluation have been proposed, and computational tools to carry them on effectively are now available. From Henry Thompson's (1992) proposal to IBM's BLEU, through Niessen et al.'s (2000) proposal and NIST's 2001 MT Evaluation, all of these measures make heavy use of large sets of reference data (or golden standard).

It is indeed acknowledged that, while a unique "correct translation" of a source is insufficient for evaluation (since another perfectly acceptable translation can differ substantially from the first one), the solution may reside in the use of a set of reference translations, which will hopefully encompass the range of possible variations among acceptable translations. Once such a set available, the quality of candidate translations can be judged with respect to it, by automatically computing a similarity distance between the candidate and the set. Evaluation is thus greatly accelerated.

However, producing such resources is quite expensive. A team of professional translators must be hired and asked to translate a number of reference texts. The quality of the reference translations thus produced would be high, but maybe some more simplistic formulations, acceptable from an MT system, would not be present in the corpus, thus biasing the results.

We propose here to build a corpus of translations using translations exams from the Ecole de Traduction et d'Interprétation (University of Geneva). These translations are encoded using markup, together with the corrections made by professors, and most important, with the *grade* that has been decided. We describe below this construction effort, than describe the data that will be used in the LREC 2002 MT Evaluation Workshop.

2. Description of the corpus

2.1. Structuring the data

One of the principles underlying the encoding of the data is to encode the most part of the information present on the paper version of the exam. This includes mainly the text produced by each student, the corrections added by the professors grading the exam, and the final grade.

We chose an XML-based annotation format, with one file per translation. Each file has a header containing useful data (except the name of the student, who is never typed in), and a <content> element with the translation. Instead of giving the DTD that was written, here is an example of exam file.

```
<?xml version="1.0"
      encoding="iso-8859-1"
      standalone="no" ?>
<!DOCTYPE exam SYSTEM "exam.dtd">
<exam>
  <header>
    <index>101</index>
    <author>101</author>
    <date>11/02/2002</date>
    <source-language>en</source-language>
    <target-language>fr</target-language>
    <level>2e cycle (years 3-4)</level>
    <exam-title>Traduct. FR/EN</exam-title>
    <comments>Exam graded by two
independent reviewers. This is a non-native
English speaker. Teacher's comments: "Your
style was confident, your English
idiomatic. Only minor mistakes appear in
the flow of your translation. Good work."
    </comments>
    <grade max="6.0" pass="4.0">5.0</grade>
  </header>
  <contents>
    <title-zone>
      <s>...</s>
    </title-zone>
    <p>
    <s>...</s>
    ...
    </p>
  </contents>
</exam>
```

Figure 1. Example of translation header.

Together with the DTD, we also use tools to validate each XML file, as well as a simple XSL file (stylesheet) that extracts the original text and discards the markup (this stylesheet is used to produce the workshop data described in the next section).

The innovative part of this corpus of “imperfect” translations is the encoding of the mistakes, together with their corrections. This requirement renders the typing of the data a bit more tedious, but increases the value of the resource, since the erroneous fragments of the texts can be discarded (or given a lower weight) when computing the distance between a candidate translation and the corpus.

Several conventions have been used to encode the mistakes and their correction: the *<m>* tag denotes a mistake, and the attributes encode its correction. The ‘t’ attribute encodes the type, as noted by the professor (‘–‘ means a fragment to be deleted), while the ‘w’ attribute encodes the replacement string. Missing parts are encoded as an empty *<m/>* element, with t=”miss” and w=”the missing string”. A sample corrected paragraph is shown below.

```
<p>
   <s>Just like you, we feel convinced
   that the prevention of drug addiction
   <m t="-" w="none">s</m> starts at
   home, through <m t="-">the</m> <m
   t="miss" w="a good"/> <m t="w"
   w="relationship">relation</m> between
   adults and children, by strengthening
   self-esteem.</s>
   <s>The findings of recent studies
   clearly show that the earlier the
   prevention, the <m t="gr" w="more">
   most</m> efficient it is.</s>
</p>
<p>
   <s>You do not necessarily need to be a
   specialist in drug addiction <m t="-">
   s</m> to talk over this issue with
   your children.</s>
   <s> The most important thing <m t="-"
   w="is">lies in</m> dialog, <m t="-">
   in</m> attentive listening, <m t="-">
   in</m> reciprocal confidence.</s>
</p>
```

Figure 2. Translated paragraph and annotated mistakes.

2.2. Present state of the corpus

The corpus presented above is still under construction. As members of the Translation Faculty at the University of Geneva, we have been granted access to the written examinations of translations students (anonymized). We are focusing, for this corpus, on pure translations: the students are required to produce, in a limited amount of time and without dictionary, a translation of a piece of text – in general an excerpt from an article or essay, broadly speaking with a “general” vocabulary (through more specific exams, such as law translation, do exist).

Several language pairs are tested for at our faculty. The best represented ones, in terms of number of translations, are translations from English into French. However, given that a majority of researchers focuses on translation *into* English, we collect also French-to-English translations (less numerous).

The quality level of these translations is quite variable, as well as the difficulty of the source text. A considerable part of the corpus comes from entry-level examinations, but there are also translations from students that are close to graduation; in this case, the source texts are more “difficult” (a notion that must still be quantified).

The corrections are done on the paper version by two graders, teachers of the faculty. Their annotations are by no means standardized, but we attempt to grasp them in the most precise manner using the annotation format described above. The encoding principle is that *stripping a text from its XML annotation must yield exactly the text produced by the candidate.* The consistency and correction of the typed texts are checked by a second annotator, and the validity of the XML mark-up is checked against the DTD using a parser (Xalan-Java).

For the time being, a total of about 50 translations of two texts have been encoded. The public distribution of this data is still under consideration.

2.3. Possible uses of the corpus

The construction of this corpus is part of a long-term effort in MT evaluation at ISSCO/TIM/ETI, University of Geneva. The main use of the corpus is as a resource for automatic evaluation, where the cost of the resource lies in typing and encoding the data, rather than asking professional translators to translate a given source text. Given that this is a corpus of “imperfect” translations, we must encode also the corrections that were made by the graders (teachers). This increases the reliability of the corpus when used for automatic evaluation, since the erroneous fragments of the student translations can be discarded or given less confidence. The grades obtained by each translation can also be used to modulate the confidence attributed to each translation.

The corpus can also be used, of course, to extract statistics about the types of translations mistakes, and the correlation between the distribution of mistakes in a translation and the grade scored by that translation. Of course, the corpus could serve also to explore automatic techniques to grade human translations, which differ quite strongly from machine translations (translation quality, proximity to source structures, etc.).

3. Description of test data for the workshop

For the present workshop, the organizers provide test data consisting in two sets of translations extracted from the corpus, enriched with machine translations of the same text. The test data is available at the workshop’s site: `http://www.issco.unige.ch/projects/isle/mteval-may02/`.

- The source texts (*10S.txt* and *20S.txt*) are excerpts from two longer essays, originally in French – the source is of course provided, as well as a reference translation for each text (*10A.txt* and *20A.txt*) constructed from the best student translations, using also the teacher’s corrections. Of course, these aren’t meant to be “the perfect translation”, but only correct translations that are close enough to the source text to help evaluators that do not understand French

For each of the two source texts, we provide about a dozen translations in English, some of them by translation students and some by commercial systems available over the Internet. Translations are numbered *101.txt* through *113.txt* and *201.txt* through *213.txt* (three numbers are missing from the second list, for technical reasons). There is no particular order, and in particular *1XY.txt* vs. *2XY.txt* are not necessarily translated by the same translator (human or system).

The human translators were not instructed to use either of the particular varieties of English (British vs. American), hence some slight spelling variations. The systems were simply those made available over the Internet by various providers, as listed for instance on the following page, compiled by Laurie Gerber: `http://www.lim.nl/eamt/resources/`. We do not wish to disclose the names of the systems that produced the various translations, since the evaluations produced in this workshop do not claim commercial-level reliability.

A sample of the translations produced for the first text (including source and reference) is provided for visual comparison in the table below.

Subject to availability, and depending on decisions that will be made after the time of writing, extra data will be made available at the workshop’s website (`http://www.issco.unige.ch/projects/isle/mteval-may02/`), and the participants will be informed as soon as possible about updates.

<table>
<tr><td colspan="2"><i>Source text</i>
Comme vous, nous sommes convaincus que la prévention des toxicomanies commence dans la famille, dans la relation entre adultes et enfants, à travers le renforcement de l'estime de soi.
Les résultats d'études récentes le démontrent clairement : plus la prévention commence tôt, plus elle est efficace.
Il n'est pas forcément nécessaire d'être un spécialiste des toxicomanies pour aborder ce sujet avec vos enfants.
L'essentiel est ailleurs, dans le dialogue, dans l'écoute, la confiance réciproque.</td></tr>
<tr><td colspan="2"><i>Reference translation</i>
Like you, we are convinced that the prevention of dependence begins at home, through the relationship between adults and children. This is done through reinforcing the child's self-esteem.
The findings of recent studies clearly show that the earlier prevention starts, the more efficient it will be.
You do not necessarily need to be an expert in drug dependence to talk about this issue with your children.
What really matters is talking together, listening to each other, and having mutual confidence in one another.</td></tr>
<tr><td><i>Translation 101</i>
Just like you, we feel convinced that the prevention of drug addictions starts at home, through the relation between adults and children, by strengthening self-esteem.
The findings of recent studies clearly show that "the earlier the prevention, the most efficient it is."
You do not necessarily need to be a specialist in drug addictions to talk over this issue with your children.
The most important thing lies in dialog, in attentive listening, in reciprocal confidence.</td><td><i>Translation 108</i>
As you, we are convinced that the prevention of the drug addiction begins in the family, in the relation among adults and children, through the intensification of the respect of one.
The results of recent studies demonstrate him(it) clearly: the more the prevention begins early, the more it is effective.
It is not necessarily necessary to be a specialist of the drug addiction to approach this subject with your children.
The main part is somewhere else, in the dialogue, in the listening, the mutual confidence.</td></tr>
<tr><td><i>Translation 102</i>
One thing is sure, we both agree: prevention of drug addiction starts at home, through the relationships between adults and children where the self-esteem has to be strengthened.
Outcomes of recent studies carried out recently, clearly demonstrate that the sooner the prevention begins, the better and the more successful it will be.
You needn't be a specialist in drugs to talk about it with your children.
It is necessary to listen to them, you must establish a real dialogue based on reciprocal confidence.</td><td><i>Translation 109</i>
As you, we are convinced that the prevention of the drug addiction begins in the family, in the relation between adults and children, through the intensification of the self-respect.
The results of recent studies demonstrate him(it) clearly: the more the prevention begins early, the more it is effective.
It is not necessarily necessary to be a specialist of the drug addiction to approach this subject with your children.
The main part is somewhere else, in the dialogue, in the listening, the mutual confidence.</td></tr>
<tr><td><i>Translation 103</i>
Like you, we are convinced that drug prevention begins within the family, in the relationship between grown-ups and children, through the encouragement of</td><td><i>Translation 110</i>
Like you, we are convinced that the prevention of drug-addiction starts in the family, in the relation between adults and children, through the reinforcement</td></tr>
</table>

self-esteem. Recent studies have clearly shown that the earlier the prevention begins, the more efficient it is. It is not unavoidably necessary to be a specialist in drug addictions to talk about this subject with your children. What matters more is discussion, attentive listening and mutual trust.	of the regard of oneself. The results of the recent studies show it clearly: the more the prevention starts early, the more it is effective. It is not inevitably necessary to be a specialist in drug-addiction to tackle this subject with your children. Essence is elsewhere, in the dialogue, in listening, reciprocal confidence.
Translation 104 Like you, we are convinced that the prevention of dependences begins at home, through the relationship of parents with their children. This is done through the reinforcement of the child's self-esteem. As recent studies have clearly shown, the earlier prevention starts, the more efficient it will be. You do not necessarily need to be an expert in dependences to talk about this issue with your children. What really matters is talking together, listening to each other, and having confidence in one another.	*Translation 111* Like you, we are convinced that the prevention of drug-addiction starts in the family, in the relation between adults and children, through the reinforcement of the regard of oneself. The results of the recent studies show it clearly: the more the prevention starts early, the more it is effective. It is not inevitably necessary to be a specialist in drug-addiction to tackle this subject with your children. Essence is elsewhere, in the dialogue, in listening, reciprocal confidence.
Translation 105 Like you, we are convinced that prevention starts at home: the relationship between parents and children as well as the child's self-esteem are of great importance. Recent studies have shown very clearly that the earlier prevention starts, the more effective it will prove. You do not necessarily need to be an expert in addictions to talk about that issue with your children. Exchanging thoughts, listening to each other as well as mutual trust is much more important.	*Translation 112* As you, we are convinced of the prevention of the drug addictions beginning in the family, in the relationship between adults and children, through the reinforcement of the esteem of themselves. The results of recent studies demonstrate it clearly : the earlier the prevention begins, the more efficient it is. Him n ' is not inevitably necessary of to be a specialist of the drug addictions to approach this subject with your children. The essential is elsewhere, in the dialogue, in the listening, the reciprocal trust.
Translation 106 Like you, we are convinced that the prevention of drug addiction begins within the family, in the relationship between adults and children, through the reinforcement of self-confidence. Recent study results show this clearly: the earlier the prevention starts, the more efficient it is. It is not completely necessary to be a specialist on drug addiction to discuss this subject with your children. The importance is elsewhere: it is in the discussion, in the listening, in the mutual confidence.	*Translation 113* Like you, we are convinced that the prevention of drug-addiction starts in the family, in the relation between adults and children, through the reinforcement of the regard of oneself. The results of the recent studies show it clearly: the more the prevention starts early, the more it is effective. It is not inevitably necessary to be a specialist in drug-addiction to tackle this subject with your children. Essence is elsewhere, in the dialogue, in listening, reciprocal confidence.
Translation 107 As you, we are convinced that the prévention of the toxicomanies begin in the family, in the relation between adults and children, through the reinforcement of the esteem of oneself. The results of recent studies show it clearly: more the prévention begin early, more she is effective. It is not necessarily necessary be a specialist of the toxicomanies to approach this subject with your children. The essential is elsewhere, in the dialog, in the listen, reciprocal confidence.	

Figure 3. Excerpt from the test data: source text (French), reference translation, candidate translations from humans and from commercial systems available over the Internet.

The references of the two source texts are the following:

- Excerpts from the brochure "Prévenir ses enfants des problèmes de drogue", Institut Suisse de Prévention de l'Alcoolisme et Autres Toxicomanies (ISPA), 24 p., 1999. (Free, order at *http://www.sfa-ispa.ch*
- Micheline Centlivres-Demont, "Hommes combattants, femmes discrètes : aspects des résistances subalternes dans le conflit et l'exil afghan" (p.169-182, excerpt at p. 178). In "Hommes armés, femmes aguerries : rapports de genre en situations de conflit armé", Fenneke Reysoo, editor, DDC/Unesco/IUED, Geneva, 2001, 250 p.
  Proceedings of a colloquium held at the Institut Universitaire des Études du Développement, Geneva, 23-24 January 2001.
  Available freely at the IUED's press service or at: http://www.unige.ch/iued/new/information/publications/yp_tm_hommes_armes_femmes.html).

4. References

G. Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *HLT 2002, Human Language Technology Conference*, San Diego, CA.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. BLEU: a method for automatic evaluation of MT. Research Report, Computer Science RC22176 (W0109-022), IBM Research Division, T.J.Watson Research Center, 17 September 2001. See `http://domino.watson.ibm.com/library/CyberDig.nsf/home`, and search for 'RC22176'.

H. S. Thompson, ed., 1992. The Strategic Role of Evaluation in Natural Language Processing and Speech Technology. Record of a workshop sponsored by DANDI, ELSNET and HCRC, University of Edinburgh, Technical Report, May 1992.