Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2016/02/jls.htm
2016 Vol. 6 (S2), pp. 670-678/Hosseini
Research Article

A SURVEY OF DIFFERENT TYPES OF EMPIRICAL MACHINE TRANSLATION AND ITS EVALUATION

***Ameneh Hosseini**
Alborz Higher Education Institute
**Author for Correspondence*

ABSTRACT
The problem of evaluating machine translation (MT) systems is more challenging than it may first appear, as diverse translations can often be considered equally correct. The task is even more difficult when practical circumstances require that evaluation be done automatically over short texts, for instance, during incremental system development and error analysis. On the one side, we have studied the problem of automatic MT evaluation. "Translation Evaluation" (TE) is a delicate process. It plays a considerable role in the process of Translation Education. TE is a tool by which translation education could get its pre determined aims. The importance of TE through the last decades has led to many studies and researches in this field of study. Various strategies using tools and models of TE, based on linguistics and interdisciplinary fields, have been presented. The stimulus of moving from one strategy to another is to objectify TE more than before so that its findings become more concrete and supportable. But such an objectivism is more challenging. This paper has paid more attention to those challenging facets and showed to what extent this objectivism has been attained. Besides, some important strategies of the past and present based on five criteria of acceptable evaluation to signalize their shortcomings in the process of TE have been analyzed. A new procedural eclectic model of TE heeding the cited criteria has been introduced at the end.

Keywords: *Translation Evaluation, Criteria of Evaluation, Eclectic Method of TE*

INTRODUCTION
Machine Translation (MT) is one of the earliest and most paradigmatic problems in Natural Language Processing (NLP) and Artificial Intelligence (AI). Although the first writings on the use of mechanical devices for translation date back from the seventeenth century, we must situate the origins of MT as a field in the late 1940's, right after World War II, with the availability of the first electronic computers in the US. In spite of their simplicity, original MT systems, based on bilingual dictionaries and manually-defined lexicalized reordering rules, obtained very promising results (Stout, 1954). Evaluating machine translation (MT) 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. Evaluation of translation aims at analyzing and marking the translation drafts of students based on a specific theory translation. What is absolute in this process is the comparison of the Target Text (TT) with Source Text (ST), while what aspects of text (Linguistic, Paralinguistic or both) to be noted, and what tools and models of evaluation to be applied is an over changing fact of TE. Nowadays, a variety of TE strategies is being used at the process of translation education, ranging from "*traditional approach*" comparing TT with its ST very subjectively and limitedly to "*modern approaches*" using new tools (e.g. parallel texts, corpora, testing frames) and models (e.g. functionalism, text typology, etc). In this article the most outstanding strategies of TE are being studied with regard to the five criteria of systematicity, comprehensiveness, validity, reliability, and objectivity. Then, a new eclectic strategy is presented to optimize these criteria in its process. McAlester (2000) introduces four criteria for TE as follows: reliability, validity, subjectivity and practicality (McAlester, 2000; Garant 2009). This study proposes that besides these four criteria, comprehensiveness and systematicity are essentially needed for TE so that it could cover all of the factors involved in translation systematically. On the other hand, practicality could be replaced with systematicity since one of the merits of systematicity is to conform the frame of TE to the situation in which the evaluation should take place in order to actualize evaluation. After all,

practicality itself is a precondition of the five other criteria. That is, we cannot talk of the existence of these criteria in a strategy, unless they have been realized in the real world.

Machine Learning

In attempting to create new automatic machine translation evaluation metrics that will address some of the recently observed outstanding difficulties, it is tempting to suggest the use of general-purpose machine-learning methods as a means of directly approximating human evaluations. However, this approach is problematic for at least two reasons, both deriving fundamentally from the resource problem. First, machine-learning methods universally adhere to a "more is better" principle with respect to the size of the training set. To successfully learn evaluation scores would require the initial development of a large set of human evaluations and, consequently, consume large amounts of time and money. If such a project were planned, it would need to involve a very carefully designed methodology in order to ensure that the evaluation data received is of the best possible quality and the greatest possible longevity. Given the extensive research history of human MT evaluation methods and practices, such a task might itself be very difficult or even prohibitive. Second, however, even if such a gigantic resource could be created, it would be necessarily static, representing a fixed distribution of MT outputs on a fixed set of language pairs. It is crucial, then, that any machine-learning approach to automatic MT evaluation sustains the ability to be retrained and itself reevaluated with respect to modern and representative translation samples at regular intervals.

Empirical MT

The second part of this thesis focuses on the study of fully automatic empirical MT of written Natural Language. By fully automatic we emphasize the fact that very light human interaction is required. By written Natural Language we distinguish text translation from speech translation. Figure 1 depicts the prototypical architecture of an empirical MT system. Translation knowledge is acquired from a parallel corpus produced by human translators encoding translation examples between the languages involved. Parallel corpora are machine-readable document collections in two or more languages, such that each document is available in all languages, either as a source document or as the human translation of the associated source document. Typically, parallel corpora are automatically aligned at the paragraph or sentence level (Gale & Church, 1993). Minimal aligned units are often referred to as *segments*. Parallel corpora are also called bitexts when there are only two languages represented. Empirical systems address MT as the problem of deciding, given an input text and acquired MT knowledge models, which is the most appropriate translation according to a given optimization criterion. Pre-processing and post-processing steps (e.g., tokenization, dedicated treatment of particular expressions such as dates, etc.) are optional. Among empirical MT systems, the two well-studied paradigms are Example-based Machine Translation (EBMT) and Statistical Machine Translation (SMT). Originally, these two approaches were clearly differentiable. EBMT methods used to be linguistically guided whereas SMT methods were statistically guided.

Machine Translation Overview

MT of Natural Language (NL) is a very difficult task. It can be perceived as the simple substitution of words in one natural language for words in another. Yet it is not so simple because of the complexity of natural languages: many words have various meanings and so they can be translated in different ways. Also, the sentences might be ambiguous and have various meanings. The relationship between linguistic entities is often vague; grammatical relations can vary depending on the languages, and translating sentences from languages having different relations means reformulating the sentence. Besides, problems due to the associated world knowledge may be encountered and these are usually difficult to solve. From a linguistic point of view, we have to consider various types of dependencies:

Morphologic; Syntactic; Semantic; Pragmatic Dependencies

Machine translation of natural languages, commonly known as MT, has multiple personalities (Nirenburg and Wilks, 2000). First of all, it is a venerable scientific enterprise, a component of the larger area of studies concerned with the studies of human language understanding capacity. Indeed, computer modeling of thought processes, memory and knowledge is an important component of certain areas of

linguistics, philosophy, psychology, neuroscience, and the field of artificial intelligence (AI) within computer science. MT promises the practitioners of these sciences empirical results that could be used for corroboration or refutation of a variety of hypotheses and theories. But MT is also a technological challenge of the first order. Figure 1 (Och, 2000) gives the standard visualization of the three approaches:
1. Direct Approach; 2. Transfer Approach; 3. Interlingua Approach

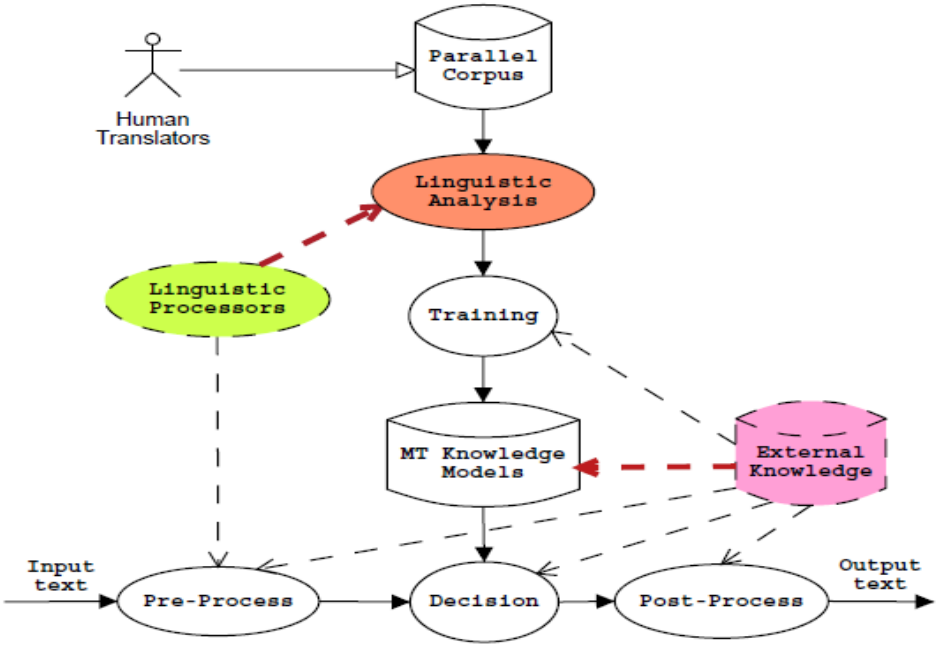

Figure 1: Architecture of an Empirical MT System

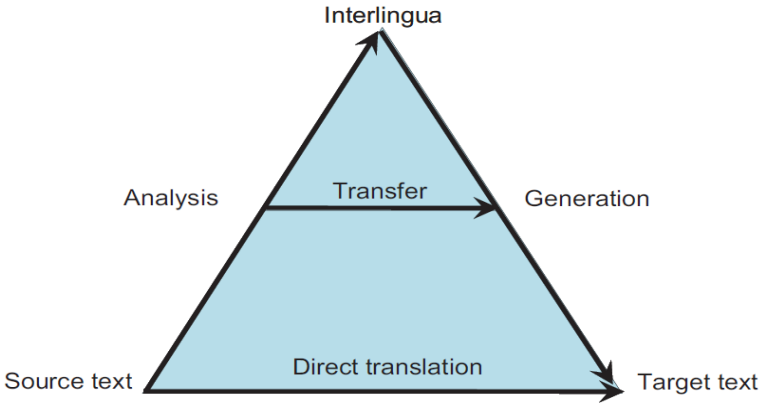

Figure 2: Different Levels of Analysis in an MT System

The Role of Evaluation Methods

The current development cycle of MT systems follows the flow chart depicted in Figure 2. In each loop of the cycle, system developers must identify and analyze possible sources of errors. Eventually, they focus on a specific sub-problem and think of possible mechanisms to address it.

Then, they implement one of these mechanisms, and test it. If the system behavior improves (i.e., the number of the selected type of errors diminishes without harming the overall system performance), the mechanism is added to the system. Otherwise, it is discarded. In the context of MT system development, evaluation methods are necessary for three main purposes:

• *Error Analysis*, i.e., to detect and analyze possible cases of error. A fine knowledge of the system capabilities is essential for improving its behavior.
• *System Comparison*, i.e., to measure the effectiveness of the suggested mechanisms. This is done by comparing different versions of the same system. It is also common to compare translations by different systems, so system developers may borrow successful mechanisms from each other. This allows the research community to advance together.
• *System Optimization*, i.e., the adjustment of internal parameters. Typically, these parameters are adjusted so as to maximize overall system quality as measured according to an evaluation method at choice.

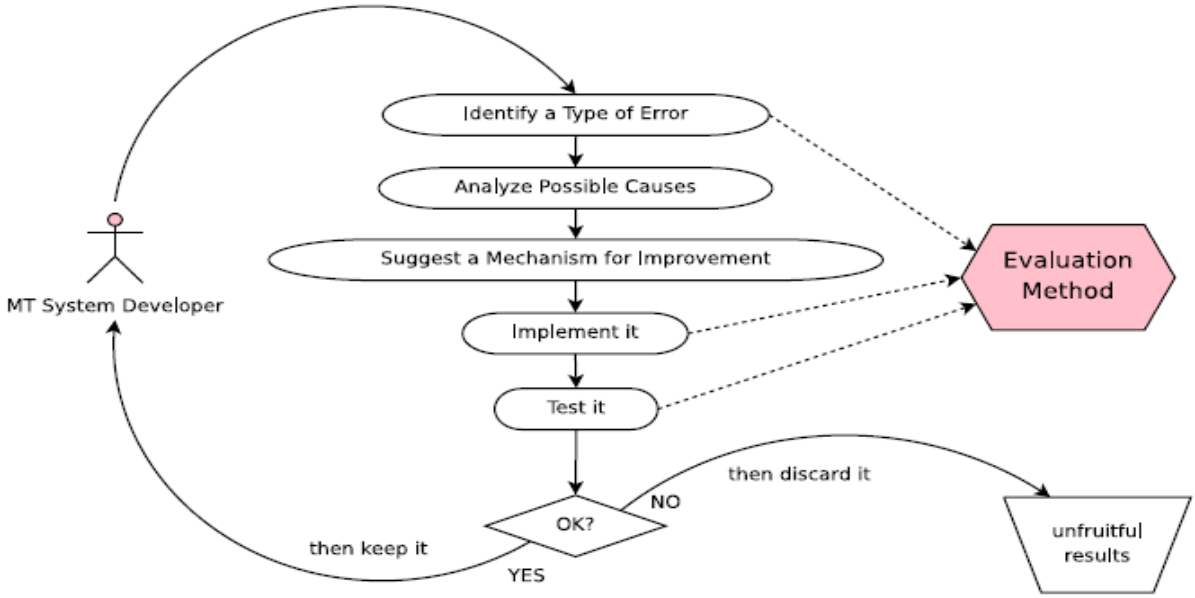

Figure 3: MT System Development Cycle

Human Likeness

A prominent alternative criterion is to evaluate metrics in terms of their ability to capture the degree of *human likeness* of automatic translations. The underlying assumption is that *good* translations should resemble human translations. When a system receives a high score according to such a metric, we can ensure that the system is able to emulate the behavior of human translators. The main advantage of human likeness is that it is a much more cost-effective alternative, since the need for human assessments disappears. Human likeness opens, thus, the path towards a new development scheme entirely based on automatic metrics. In this scheme, human subjects are only required for solving the test cases (as systems do) and, thus, to serve as models (i.e., providing human references) for the evaluation process. Avoiding human assessments eliminates also one subjective factor: the assessment evaluation guidelines. In addition, human assessments are static, while discriminative power can be updated if new human references or system outputs are incorporated to the test bed along time. However, meta-evaluation based on human likeness presents a major shortcoming; just like automatic evaluation, it depends strongly on the heterogeneity/representativeness of the test beds employed (i.e., sets of test cases, and associated automatic system outputs and human reference translations). For instance, if the set of reference translations per test case is small it may not represent well the full set of acceptable solutions, and the meta evaluation process may be biased. Therefore, the applicability of human likeness as meta-evaluation criterion must be further studied and validated. While human likeness is a sufficient condition to attain human acceptability, human acceptability does not guarantee human likeness. In other words, human judges consider acceptable translations that are human-like, but they may also consider acceptable many other automatic translations that would be rarely generated by a human translator.

A Prelude to Eclectic Method of TE

We claim that TE is a long-term process consisting of some essential stages. Each of these stages includes different combination of tools, models and aims of evaluation in a meaningful and continuous mood. These stages are designed according to syllabus and predetermined purpose of each stage of translation education process. The dominant factor in the process of translation education is "*Translation competence"*. Toury (1984) divides translation competence into three parts: (1) Bilingual competence which refers to mastering to both SL and TL in all linguistics levels, (2) Interlingual proficiency means to know how the two languages are similar and different from one another, (3) Intercultural transfer competence that is to be able to transfer the ST to TT in a sociocultural context (Toury 1984; Malmakjar, 2008). Translation competence evolves via the process of *theorizing* (borrowed from Robinson 1997) from one level to another. According to Robinson (1997), the central idea of theorizing phenomena is the fact that the students being exposed to different situations of translation activate and reconstruct their own knowledge and understanding of translation phenomena. Thus, the higher the level of translation competence of the students, the more linguistic and non linguistic aspects of translation will be exploited by them. We borrowed Toury's three part division of translation competence as the base of three stages of translation education process. We believe that as the overall process of translation education at last leads to the final goal (that is competent translator), each individual stage follows its certain aim (mastering bilingual competence, interlingual proficiency, and intercultural transfer) parallel to that final one. In each stage different tools and models may be applied. Through adopting the same policy to TE, we design a stage by stage evaluation process that accurately fits the frame and strategies taken for translation education since translation competence is the most related factor to TE as well. That is, we indeed evaluate the translation competence of the students. So, in order to evaluate the ever-changing aspects of competence we should address them one by one, and use the appropriate tool and model of evaluation at any stage.

The Eagles Guidelines for NLP Software Evaluation

The European EAGLES initiative (Expert Advisory Group on Language Engineering Standards) came into being as an attempt to create standards for language engineering. The initiative was born out of a perception that linguistic resources were essential to progress in the area, but were expensive and time-consuming to create. Agreed standards for the form and content of resources would facilitate resource transfer across projects, product development, and different applications. The first areas to be attacked in the first phase of the initiative (1993–1995) were corpora, lexicons, grammar formalisms, and evaluation methodologies. 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 to which the system was to be put and on its potential users. However, it did seem possible to create what was called 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 a report by Sparck Jones and Galliers (1993), later reworked and published in book form (Sparck Jones and Galliers, 1996). Influenced by earlier work in evaluation, including the ISO/IEC 9126 standard published in 1991 (see next section), the EAGLES Evaluation Work Group (EWG) proposed the creation of a quality model for NLP systems in general, in terms of a hierarchically structured classification of features and attributes, where the leaves of the hierarchy were measurable attributes, to which specific metrics were associated. The quality model in itself was intended to be very general, covering any feature which might potentially be of interest to any user. 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 measurements to be combined in different ways (EAGLES, 1996). These first attempts at providing a theoretical framework were validated by application to quite simple examples of language technology: spelling checkers were examined fairly thoroughly, and preliminary work was done on drawing up quality models for grammar checkers and translation memory systems (TEMAA, 1996). Whilst the case studies tended to confirm the utility of the theoretical framework, they also stressed the attention that had to be paid to sheer meticulous detail in designing a valid evaluation. In the second phase

of the EAGLES initiative (1995–1996), work on evaluation was essentially limited to consolidation and dissemination of the guidelines (EAGLES, 1999). During this time, the EAGLES methodology was used outside the project to design evaluations of a dialog system (Blasband, 1999) and of a speech recognition system (in a private company), as well as a comparative evaluation of a number of dictation systems (Canelli *et al.*, 2000). The designers of these evaluations provided useful feedback and encouragement. Also during the second phase, the EWG came into closer contact with the ISO/IEC work on the evaluation of software in general. When the ISLE project was proposed in 1999, it transpired that the American partners had also been working along the lines of taxonomies of features (Hovy, 1999), focusing explicitly on MT and developing with the same formalism a taxonomisation of user needs, along the lines suggested by the JEIDA study (Nomura, 1992). The EWG of the ISLE project therefore decided to concentrate on MT systems, refining and extending the taxonomies that had been proposed. It is essentially this work which is described here.

Definition of a Quality Model

According to ISO/IEC 14598-1 (1999a), the software life-cycle starts with the analysis of the user needs that will be answered by the software, which determine a set of software specifications. From the point of view of quality, these are the *external quality requirements*. During the design and development phase software quality becomes an *internal* matter related to the characteristics of the software itself. Once a product is obtained, it becomes 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 its origins, *quality in use* is the extent to which the software really helps users fulfill their tasks (ISO/IEC, 2001b). According to ISO/IEC, 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 the case of MT software, an important element of the life-cycle must be taken into account: there is no straight forward link, in the conception phase, from the external quality requirements to the internal structure of a system. To use an example, it may be possible to design a valid system that sells books online, based on the requirements that the system manage a database of at least one billion titles and 100,000 customers. However, it is at present impossible to infer the design of a system that translates the books, from requirements that the target must be 100 languages. Research in MT (and in other branches of NLP and Artificial Intelligence) must deal with the specification-to-design phase empirically, with the result that evaluators have then to define their own contexts of use and select external quality requirements. Therefore, the relation between external and internal qualities is quite loose in the case of MT. According to ISO/IEC (2001b), software quality results in general from six *quality characteristics*:

Functionality, *Reliability*, *Usability*, *Efficiency*, *Maintainability*, *Portability*

Already present in ISO/IEC (1991), these characteristics have been refined in the more recent version of the standard, thanks to a loose hierarchy of sub characteristics (still domain-independent) that may contain some overlapping (ISO/IEC, 2001b). The terminal entries are always measurable features of the software, that is, *attributes*. Conversely, 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 (ISO/IEC, 1999a, emphases original).

Stages in the Evaluation Process

The ISO/IEC standards also outline the evaluation process, generalizing a proposal already present in the first ISO/IEC 9126. The five consecutive phases are emphasized differently according to whom the initiators of the evaluation are, developers, acquirers, or evaluators.

1. Establish the quality requirements, i.e. the list of required quality characteristics.
2. Specify the evaluation, i.e. the measurements and their mapping to the requirements.
3. Design the evaluation, producing the evaluation plan, i.e. the documentation of the procedures used to perform measurements.
4. Execute the evaluation, producing a draft evaluation report.
5. Conclude the evaluation.

According to ISO/IEC 14598-5, the specification of measurements starts with a distribution of the evaluation requirements over the components of the evaluated system. Then, each required quality characteristic must be decomposed into the relevant sub-characteristics, and so on. Metrics must be specified for each of the attributes arrived at in this decomposition process. More precisely, three stages must be distinguished in the specification, design, and execution of an evaluation. The following order applies to execution:

a. application of a metric b. rating of the measured value c. integration or assessment of the various ratings. 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: The integration of measurements towards a *final score* is not an ISO/IEC priority, unlike the case of many NLP and MT evaluation campaigns (one has to turn back to the first ISO/IEC 9126 standard to find a mention of the integration stage). The three-stage distinction above, advocated also by EAGLES (1996, 1999), seems to us particularly useful regarding the concrete evaluations of systems.

Influence of the Evaluation Levels

The guidelines for evaluators (ISO/IEC, 1998) provide a similar, though less developed example of the influence of the intended context of use on evaluation choices. Here, a parameter somewhat parallel to integrity is defined, namely *evaluation levels*. These levels range from A (most critical) to D (least critical) and concern four classes of risks: environment, safety (people), economy (companies) and security (data). The standard provides a ranking of "evaluation techniques" (for each of the six quality characteristics) based on the required evaluation level. More demanding techniques should be used for higher levels. For instance, for efficiency, from less to more demanding levels, one should carry out: execution time measurements; benchmark testing; an analysis of the design to determine the algorithmic complexity. It is obvious that only very specific factors are taken into account here (those associated with risks). But usability studies point out the strong influence of the context of use on the quality model.

Process of Eclectic Method of TE

This study divides translation education course into two major parts: (1) Microlinguistic and (2) Macrolinguistic. At the first part the language systems of SL and TL and their differences and similarities at (semantics, syntax, lexis, etc.) levels would be taught to students so that the bilingual and interlingual competences of them improve. Because of emphasizing textual features by students through the process of translation, theories to be applied for this part are generally structural ones like that of Jacobson`s (1959); Catford's (1965), etc.

Observing the good control of students on the microlinguistic aspects of SL and TL, teacher embarks on the second part. Here a translation phenomenon is provided with a situational and functional approach and the focus shifts from the translation text to the process of translating. In this process the students become aware of the fact that translating is a socio cultural activity. Now they practice to enhance their intercultural transfer competence which aims at contextualization of translation into the receptor culture and, as a result, the application of functional theories (See House 1977/1997, Reiss 1977/1989, Reiss & Vermeer 1989) becomes essentially necessary to this stage. In this stage they learn how to produce functional equivalence.

Subsequently this would be explained more. Teachers usually use both traditional and modern strategies to evaluate the students` translation competence. But none of the strategies could satisfy the teacher or the students. What is essentially ignored while being vital is the *conformity* of evaluation strategy to the principles of each education stage.

Evaluation of Bilingual Competence

The first step of evaluation in translation classroom is done like that of language education. Testing is a good option for assessing the students' abilities on SL and TL separately. As mentioned before, testing is linguistically objective, valid, reliable, and a systematic strategy. Thus, it is suitable for assessing the novices translation. What is considered here is such topics as structural, lexical, and orthographic correctness. Theses will be evaluated well by different types of language tests.

Evaluation of Interlingual Competence

After relative mastering of structural, lexical, semantic and syntactic levels of both languages, students should go ahead and perceive the differences and similarities between the source and target languages. The more different languages, the more challenging this stage would be. This stage, welcoming to compare and contrastive analysis, is the starting point at which the two languages involved in translation education encounter each other. At the first phase of this stage integration of traditional approach into testing format is recommended. Since the selected texts to be translated by the students are structurally and lexically the simplest ones, by relying on his knowledge the teacher could identify mistakes committed by the students. Integration of the traditional approach is vital to provide the TE with systematic and concrete marking of student`s drafts. It should be noted that the purpose of this stage is nothing more than the sheer assessing of students' realization of simple differences of the two language structures and lexis. At the more developed phase the integration of bilingual corpus-based approach with testing is suitable. Bilingual corpus is electronic collections of STs and TTs. The purpose of evaluation is to see how much the students are aware of textual equivalence at micro linguist level between ST and TT. Because the selected texts to be translated are more complicated and realistic than those used in previous stages, the teacher should be supported by resources like corpus texts emerged from the real world. Like previous steps, these texts should be designed in testing formats.

Evaluation of Intercultural Transfer Competence

Competences of the students at this stage guide them to contextualize the translation process. They see that translation follows a special skopos and function, has its particular readership, is affected by the power relations in the society, and is an ideological affair supported (or rejected) by social institutions. Thus, assessing the students' awareness and application of "contextualization" in translation should not be missed out. This would be done through placing the students in challenging contexts of translation, where they should identify the source text function and reproduce it in the target text. Considering factors like pragmatics, subject field, text function, text type, and readership is necessary in the macro linguistic level of translation education. Since these factors are "invisible" for the evaluator, and not accessible to him. Therefore, they are more challenging than the concrete textual factors. To evaluate these factors there is no way except assessing their "representatives" in the translation draft. These representatives may be special terminologies, linguistic, esthetic, or stylistic patterns, etc. So, the teacher should recognize them and consider them into his evaluation, that is, to see to what extent the students are aware of these contextual factors, and how this knowledge is represented in their translations. But as mentioned before, the teacher could not do it by himself. So, he should resort to some sources. The best source is parallel texts in the form of corpora. Due to its lack of systematicity, these corpora should be integrated with the testing format. This means that the corpora samples should be presented to students in the form of tests. Up to now testing has been used just for assessing the linguistic elements, but now its application will be extended to nonlinguistic ones as well. For example, if the purpose of evaluation is to assess the functional dimension of the translation, it is appropriate to use the text typology model and, therefore, extract some texts from the parallel corpus and tailor them in a testing mood. As said before, both students and teachers benefit from the corpora as a whole. They become familiar with the special textual patterns, terminologies and concepts that are used (differently in two languages) for a special function. As a last word, the two approaches of testing and corpora could work together and form an eclectic method of objective, comprehensive, systematic, valid, and reliable.

Conclusion

Applying machine learning to the problem of machine translation evaluation is primarily difficult because the desired human judgment targets are too expensive to make available in large quantities, particularly as the population of MT outputs may be changing constantly. The solution proposed here is to approximate human judgments with a binary decision variable that reports whether a translation was produced by a human or by a machine, thereby eliminating the need for user data collection. By utilizing measures of concordance, continuous judgments can be learned from this simple binary classification. Results indicate that this approach is effective, improving sentence level correlation between metric scores and human judgments from 0.29 to 0.38, and that for a fixed set of input features, the classification criterion is

strongly linked with such correlation. The method thus provides the opportunity to obtain improved correlation even without access to a human evaluated test set.

REFERENCES

Arnold D, Lee Humphreys R and Sadler L (edition) (1993). Special Issue on Evaluation of MT Systems. *Machine Translation* **8** 1–2.

Blasband M (1999). Practice of Validation: The ARISE Application of the EAGLES Framework. In *Proceedings of the European Evaluation of Language Systems Conference (EELS)*, Hoevelaken, The Netherlands. Available at: http://www.computeer.nl/eels.htm.

Canelli M, Grasso D and King M (2000). Methods and Metrics for the Evaluation of Dictation Systems: A Case Study'. In *LREC 2000: Second International Conference on Language Resources and Evaluation*, Athens 1325–1331.

Church KW and Hovy EH (1993). Good Applications for Crummy Machine Translation. *Machine Translation* **8** 239–258.

Crook M and Bishop H (1965). Evaluation of Machine Translation, Final report, Institute for Psychological Research, Tufts University, Medford, MA.

Daly-Jones O, Bevan N and Thomas C (1999). *Handbook of User- Centred Design*, Deliverable 6.2.1, INUSE European Project IE-2016. Available: http://www.ejeisa.com/nectar/inuse/

Eagles E (1999). EAGLES Evaluation of Natural Language Processing Systems, Final Report EAG-II-EWG-PR.2, Project LRE-61-100, Center for Sprogteknologi, Copenhagen, Denmark. Available: http://www.issco.unige.ch/projects/eagles/.

Flanagan MA (1994). Error Classification for MT Evaluation. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland 65–72.

Halliday T and Briss E (1977). The Evaluation and Systems Analysis of the Systran Machine Translation System, Report RADC-TR-76-399, Rome Air Development Center, Griffiss Air Force Base, NY.

Hovy EH (1999). Toward Finely Differentiated Evaluation Metrics for Machine Translation. In *Proceedings of EAGLES Workshop on Standards and Evaluation*, Pisa, Italy.

Pfafflin S (1965). Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments. *Mechanical Translation* **8** 2–8.

Popescu-Belis A (1999b). 'L'évaluation en Génie Linguistique: Un Modèle pour Vérifier la Cohérence des Mesures' [Evaluation in Language Engineering: A Model for Coherence Verification of Measures]. *Langues* **2** 151–162.

Popescu-Belis A, Manzi S and King M (2001). Towards a Two-stage Taxonomy for Machine Translation Evaluation. In *MT Summit VIII Workshop on MT Evaluation "Who did what to whom?"* Santiago de Compostela, Spain 1–8.

Sinaiko HW (1979). Measurement of usefulness by performance test. In Van Slype, G, In: *Critical Methods for Evaluating the Quality of Machine Translation*. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management, Report BR 19142, Bureau Marcel van Dijk 91.

Taylor KB and White JS (1998). Predicting What MT is Good for: User Judgements and Task Performance', in David Farwell, Laurie Gerber and Eduard H. Hovy (edition), *Machine Translation and the Information Soup*, (Germany, Berlin: Springer-Verlag) 364–373.

TEMAA (1996). TEMAA Final Report, LRE-62-070, Center for Sprogteknologi, Copenhagen, Denmark. Available: http://cst.dk/temaa/D16/d16exp.html.

Vanni M and Miller KJ (2002). Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Metrics across Languages, In *LREC 2002: Third International* Conference on Language Resources and Evaluation, *Las Palmas de Gran Canaria, Spain, pp. 1254–1262.*