FEMTI: creating and using a framework for MT evaluation

Margaret King

University of Geneva
ISSCO/TIM/ETI
40 Bd du Pont d'Arve
CH-1211 Geneva 4, Switzerland
margaret.king@issco.unige.ch

Andrei Popescu-Belis

University of Geneva
ISSCO/TIM/ETI
40 Bd du Pont d'Arve
CH-1211 Geneva 4, Switzerland
andrei.popescu-belis@issco.unige.ch

Eduard Hovy

University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695, USA
hovy@isi.edu

Abstract

This paper presents FEMTI, a web-based Framework for the Evaluation of Machine Translation in ISLE. FEMTI offers structured descriptions of potential user needs, linked to an overview of technical characteristics of MT systems. The description of possible systems is mainly articulated around the quality characteristics for software product set out in ISO/IEC standard 9126. Following the philosophy set out there and in the related 14598 series of standards, each quality characteristic bottoms out in metrics which may be applied to a particular instance of a system in order to judge how satisfactory the system is with respect to that characteristic. An evaluator can use the description of user needs to help identify the specific needs of his evaluation and the relations between them. He can then follow the pointers to system description to determine what metrics should be applied and how. In the current state of the framework, emphasis is on being exhaustive, including as much as possible of the information available in the literature on machine translation evaluation. Future work will aim at being more analytic, looking at characteristics and metrics to see how they relate to one another, validating metrics and investigating the correlation between particular metrics and human judgement.

1 Introduction

Yorick Wilks has often been credited with the tendentious but plausible remark that more has been written about MT evaluation than about MT itself. This paper starts by asking why that might be so, concluding that there are features of MT evaluation which make evaluators feel that each evaluation is special, tempting them to design the evaluation from scratch each time. An immediate consequence of course is that work is wasted: a deal of literature is generated, but since it starts from different presuppositions and from different pre-conditions, its utility is not immediately obvious. Furthermore, much of this literature is not easily available: for example, it was until very recently extremely difficult to get hold of a copy of the Van Slype (1979) report, a major attempt to gather together a large number of proposals for how MT evaluation should be done and subject them to systematic description[1]. The evaluator charged with designing a new evaluation thus quite naturally feels that he has neither the time nor the inclination to carry out a systematic search of the literature, decide what in the literature he can and will re-use, justify his decisions – and only then get on with designing and carrying out the evaluation for which he is responsible.

After elucidating the problem, this paper describes an attempt to alleviate it. Through collaborative work in the ISLE project, funded by the European Union, the National Science Foundation in the USA and the Federal Office for Education and Science (OFES) in Switzerland, an attempt has been made to gather into one place the accumulated experience of MT evaluation, and to describe it in such a way that future evaluators can consult and re-use this experience easily. The result is FEMTI, a framework for MT evaluation. The paper sets out the principles behind the description, illustrates it with examples and suggests how it might be used.

---

[1] This report, commissioned by the European Commission in the late 1970's, is now, by agreement with the Commission, available electronically: see bibliography.

FEMTI is meant to be a resource for the MT community as a whole: it is hoped that it will continue to grow and to be used, as members of the community find it useful and make their own contributions to its consolidation and expansion. The paper concludes with a discussion of how future work might build on what has already been done to create an-ongoing permanent source of information and of inspiration for research workers, system developers and MT system users as well as for those directly involved in MT evaluation.

2 Why is MT evaluation different?

In order to approach answering this question, we need first to say what MT evaluation is being compared with. A comparison that springs to the mind of most researchers is with the evaluation campaigns organised in several different areas of human language technology over the last two or three decades, which have proved in some cases very valuable both in advancing the core technology in question and in promoting collaborative work. The largest and best known campaigns have been organised by ARPA/DARPA in the USA, often with participation from groups from other countries. Other campaigns have been organised in France and elsewhere in Europe, occasionally by volunteer groups operating without specific funding. The campaigns most usually take the form of friendly competitions, where rival groups seek to show that their approach is superior by demonstrating that given the same input they achieve better results than do other systems. Participation in the campaigns tends to promote flexibility and open-mindedness, as success with a certain technique encourages other participants to experiment with that technique, just as getting poor results with some technique may lead to abandoning or rethinking its use.

Since comparison can only be valid if all participants are working with the same materials, organising the campaigns typically requires creating substantial resources in the form of training materials used in preparing the competing systems for the competition and testing materials, used in the competition itself. These resources have an intrinsic value, independently of the competition in which they are used: they serve subsequently as invaluable resources for developing new techniques and systems. (To put this into an MT perspective, just think of all the translation technology systems which have made use of the Canadian Hansard for development and test purposes). Evaluation campaigns have covered a wide variety of human language technology applications: speech recognition, document retrieval, information extraction, parsing, alignment algorithms, tagging and others. Why does not machine translation fit comfortably into this kind of evaluation mould?[2]

The essential clue is contained in the phrase "advancing the core technology". The purpose of all these campaigns is in a way altruistic: there is no customer waiting to buy and install the best system –the aim is to discover and encourage the most effective use of techniques still in the research phase. All that the participants may get out of the campaign is an enhanced ability to attract funding, not substantial sales of a product. Consequently, no actual practical use of the systems competing is envisaged or factored into the evaluation: the focus is entirely on whether the system can produce results that compare with the results defined to be the ideal results. To put this in ISO 9126 terms (ISO/IEC 1991), the only quality characteristic taken into account is a single sub-characteristic of functionality, *accuracy*, which is defined as the capacity to deliver the "right" or agreed upon results.

Most MT evaluations do not aim at furthering science: they are commissioned, typically, by people in the real world of translation who are faced with practical problems that they hope to be able to resolve by using an MT system. This means that other factors than accuracy enter into the picture.

We have already mentioned the ISO standard 9126-1. It concerns software in general and lays down a number of characteristics which contribute to the quality of a piece of software. We shall not repeat here the argument that ISO 9126 applies to MT software as much as to any other (Hovy, et al. in press).

---

[2] In fact, DARPA did organise a series of MT evaluation campaigns in the early 1990s. They followed the typical pattern of such campaigns by concentrating exclusively on accuracy.

However, to see how varied the factors relevant to a real world evaluation might be, let us return to the quality characteristics set out in ISO/IEC 9126-1 (ISO/IEC 2001), and briefly see how factors other than accuracy might be pertinent to some particular MT evaluation.

First, a sub-characteristic of functionality which is of particular relevance to real world evaluations is *suitability*. Earlier, we talked of *accuracy* – the capacity to produce results which conform to the specifications laid down for the system. *Suitability* has to do with whether even accurate results are suitable in the particular context in which the system is to be used. To take a caricature example, there could exist a system which produced absolutely perfect translation from Chinese into Russian: it would not be suitable for someone who needed to translate from French into English. A more likely case would be a system which produced very good results working with controlled language in a specified domain: it would nonetheless not be very suitable for free text input or for another domain.

Another sub-characteristic of functionality is *interoperability*, whether with other software or with hardware platforms. The METAL system (Slocum 1987) at a certain point in its lifetime was marketed only in a version that required a Lisp machine. Very few potential customers already owned a Lisp machine, or could see any other potential use for such a machine. Furthermore, one can easily imagine the additional complications implied at the level of organising workflow. In evaluation terms, *interoperability* here becomes at least as important as accuracy and suitability.

The other top level ISO characteristics are *reliability*, *usability*, *efficiency*, *maintainability* and *portability*. Let us imagine an MT system used to provide information to spectators in some major sporting competition, the Olympic Games for example. A system which produces impeccably accurate and suitable results will be of no use if it breaks down all the time, or takes two weeks to get running again when it does break down. *Reliability* here is extremely important. Think now of a system vendor providing machine translation free on the web: those using the system are unlikely to be prepared to fight their way through interfaces which are difficult to understand, difficult to learn, hard to operate and ugly to look at: here, *usability* is of high importance. Next imagine an MT system being used as part of the task of scanning large numbers of news reports in order to determine what items deserve thorough scrutiny: if the system cannot keep up with the flow of documents, it becomes virtually useless. Here, *efficiency* is of prime importance. Maintainability includes being able to modify the system in order to adapt it to particular user needs. For MT systems, there are many contexts in which it is important to be able to add or modify dictionary entries or, with empirically based systems, to train the system on new text. Here, *maintainability* is of the essence. *Portability* includes the ease with which one version of a system can be replaced by a new version. MT systems are rarely static: they tend to be improved over time as resources grow and bugs are fixed. It is therefore difficult to imagine a context for MT use where *portability* would not be important.

The important point here is not the exact definition of any one quality characteristic or of its sub-characteristics: it is rather that MT has a multitude of potential uses in a multitude of different contexts. In any specific context, some characteristics may be important, others not, to the point where a characteristic which is a sine qua non in one context may be completely irrelevant in another. And it is precisely because the relative importance of individual quality characteristics is never the same in two different work contexts that the MT evaluator is tempted to feel that he is tackling a problem which has never been tackled before, and therefore to design his evaluation from scratch.

3 Another way in which MT evaluation is different

Evaluation also involves finding ways to determine whether a given system measures up to what is required with respect to any given characteristic, in other words finding good metrics. For many of the characteristics mentioned above, it is not too difficult, with a little ingenuity, to come up with an appropriate metric. A well-known and flagrant exception in the case of MT is accuracy, the sub-characteristic of functionality beloved of evaluation campaigns. Accuracy, it will be remembered, is the capacity of the system to come up with the "right" or agreed results. This implies knowing what the agreed results should be. In the

case of many human language technology applications reaching an agreement on what the right results should be is not an insuperable problem: for translation it is simply not possible. There is no one right translation of even a banal text, and even the same translator will sometimes change his mind about what translation he prefers. There just is no gold standard. In FEMTI, for the moment at least, we have simply tried to be exhaustive and non-partisan, including all the metrics so far suggested for accuracy and leaving it to the evaluator to choose between them.

4   Alleviating the problem

In the light of all that has been said, it is easy to understand why each evaluator feels that he is faced with a new task. What we have done with FEMTI is to try to make it easier for him to profit from the work of those who have laboured in the field before him. We have tried to collect together and systematize as much as we could of all that has been written about what might be required in different work contexts, of all that has been written about different characteristics of different MT systems and of all that has been written about ways of measuring whether a given system is likely to meet the requirements of a given situation.

Stated thus, the task is enormous. We are only too aware that it is not yet finished, and perhaps never will be finished, since new work on MT evaluation appears all the time. Even so far, it has involved the work of far too many people to mention all of them individually, and we apologise for not being able properly to give credit where it is due whilst also thanking them.

The work has mainly been through a series of workshops where participants have contributed to gathering material, to investigating new metrics, to validating metrics, to investigating relationships between metrics and to systematizing descriptions. The workshop taking place during this conference is the eighth in the series. The result is a framework for machine translation evaluation, called FEMTI for short. FEMTI is, essentially, two structured descriptions: the first relates to potential user needs, the second to characteristics of systems. Pointers from user needs lead to system characteristics related to those needs, and pointers from system characteristics lead to metrics by which a system's performance with respect to

individual characteristics might be measured. The descriptions are formally taxonomies. Inspiration for their creation came from three sources. Early EAGLES work (EAGLES MT Evaluation Working Group 1996) introduced what was called the "consumer report paradigm" – the idea that it was possible to generalize the needs of individual work contexts by describing the needs of *classes* of typical users. The evaluator can then choose elements from the description which reflect his particular situation. In the JEIDA report on MT evaluation (Nomura and Isahara 1992) particular needs were picked out from a pre-defined set of possible needs and graphically represented. The representation could be compared to an independent representation of what was offered by a specific system in order to determine how closely the system matched the specific set of needs. Finally, Hovy (1999) produced a first version of a taxonomy which would represent user needs and system characteristics.

The part of the description which sets out system characteristics is articulated around the quality characteristics set out in ISO 9126-1. A first section relates to system internal characteristics, such as the model of translation on which the system is based, the linguistic resources used by the system and the way in which it is intended that the system be used, for example whether it is meant to be used interactively, whether post- or pre-editing is foreseen and so on. The second section relates to system external characteristics – the characteristics which can be observed when the system is in use – and explicitly uses the ISO quality characteristics as an organising principle. Recent ISO work as reflected in ISO 9126-1 and other documents in the 9126 and 14598 series introduces a notion of "quality in use", which is the quality of the system as perceived once it is actually being used to perform a task. For the moment at least, FEMTI does not include material on quality in use, partly because the notion has only recently been introduced, but mainly because quality in use can only really be measured once a system is actually in use in a specific situation.

It is rather difficult to give an extensive overview of FEMTI, just because it is so large. The diagram below shos the top level nodes of the taxonomy, but it should be remembered that each of the nodes which here appears to be terminal is in fact broken down into lower level nodes, in many cases

resulting in a very detailed description of the upper node.

1.Evaluation requirements
  1.1 The purpose of the evaluation
  1.2 The object of the evaluation
  1.3 Characteristics of the evaluation task
        1.3.1 Assimilation
        1.3.2 Dissemination
        1.3.3 Communication
  1.4 User characteristics
        1.4.1 Machine translation user
        1.4.2 Translation consumer
        1.4.3 Organisational user
  1.5 Input characteristics (author and text)

2. System characteristics to be evaluated
  2.1 System internal characteristics
        2.1.1 MT system-specific characteristics
        2.1.2 Translation process models
        2.1.3 Linguistic resources and utilities
        2.1.4 Characteristics of process flow
  2.2 System external characteristics
        2.2.1 Functionality
                2.2.1.1 Suitability
                2.2.1.2 Accuracy
                2.2.1.3 Wellformedness
                2.2.1.4 Interoperability
                2.2.1.5 Compliance
                2.2.1.6 Security
        2.2.2 Reliability
        2.2.3 Usability
        2.2.4 Efficiency
        2.2.5 Maintainability
        2.2.6 Portability
        2.2.7 Cost

Simply reproducing the labels on the nodes of the taxonomy and showing its structure fails to give any real indication of the extent of the material accessible through the nodes. In order to give at least the flavour of what is available, in the next section we shall adopt the inverse strategy of tracking down through particular nodes and seeing where they lead. The reader can get an idea of the whole for himself by consulting the full version of the taxonomy at *http://www.issco.unige.ch/projects/isle/ taxonomy3/.*

5   FEMTI

As we have seen, the first section, reflecting user needs in the form of evaluation requirements, is organized into five main sections which treat the purpose of the evaluation, the object to be evaluated, characteristics of the translation task, user characteristics and characteristics of the input, covering both characteristics of the authors of the texts and characteristics of the text itself. Here we will choose just one of these, 1.3, characteristics of the translation task. Expanding that node, we find a definition, a list of stakeholders, a list of references and a set of notes.

Characteristics of the translation task

Definition: Characteristics of the translation task refers to the information flow intended for the output, from the point of view of the agent (human or otherwise) who receives the translation.

This part of the present taxonomy describes three principle types of use in such a way that users can identify the particular type of work they need to have done, while developers can define in strict terms what their MT system can do.

Stakeholders: Developers, research sponsors, commercial investors, buying agents, operational managers, end users.

References: (Hovy 1999, Hutchins 2001)

Notes

As was noted in (Sager 1978), for machine translation systems "two types of use [are] to be considered: (a) the non-edited output; (b) the edited output. The output may be acceptable for either use or both and the evaluation should determine this. In the case of edited output the cost of the revision, editing etc. has to be established and compared with the cost of manual translation. Since the type of use is related to the type of text, these types have to be established and taken into account."

Hovy (1999) suggests dividing all the possible translation tasks into three main groups. He noted that "in order to make the taxonomization of features useful to people who do not already know about MT and who do not wish to become experts in evaluation, it is important to articulate the layers and choices in terms that can be intuitively understood."

Before traveling further down the taxonomy from this node, let us notice that clicking on underlined elements will lead the person consulting the taxonomy to more detailed information, in the text above to the bibliographic references given.

Characteristics of the translation task is broken down into three sub-topics, assimilation, dissemination and communication. We shall choose just one of these, 1.3.1, assimilation. Expanding assimilation once again leads the user to more detailed information, this time a definition, some hints on relevant qualities and their relative importance and references, as shown below.

Assimilation

Definition: the ultimate purpose of the assimilation task (of which translation forms a part) is to monitor (relatively) large volumes of texts produced by people outside the organisation, in (usually) several languages.

Relevant qualities / how to measure: the required translation quality is not high, though translation speed and wide coverage are important.

· Production time/speed of translation – fast
· Quality of the translation
· Style – not a very important factor
· Syntax – not a very important factor
· Fidelity – important
· Field of application: one – which one; many – which are good, which are bad

References : (Hovy 1999)

Once again, clicking on underlined elements would lead to more detailed information, but we shall resist the temptation to go deeper here, and continue our exploration of the assimilation node itself, looking now at nodes which were not included in the high-level summary of the taxonomy shown earlier. Assimilation in fact breaks down into three sub-topics, document routing and sorting, information extraction and summarization, and search. Of these we shall pick on search, expanding it to find the information below.

Search

Definition: The goal of a search process is to identify a set of documents that together can satisfy an information need.

Subtasks include refinement of the searcher's understanding of their need, refinement of the expression of that need as a query and recognition of relevant documents.

Automated components of search systems typically accomplish only portions of the required task, leaving the searcher to assess factors (e.g. veracity and completeness) that would be difficult to assess by automated means.

Searchers with limited proficiency in languages in which the documents are written will require translation support to accomplish information need refinement, query reformulation and relevant document recognition.

Relevant qualities: Functionality, Usability, Efficiency.

Stakeholders: End users with information needs, professional searchers assisting end users with their research, persons writing documents that they wish to make easily found.

References: (Oard and Gonzalez 2001)

We are nearing the bottom of the user needs taxonomy. Clicking on underlined elements will once again take us to further information. From the above, we shall choose Functionality and move into the taxonomy describing system characteristics and how they might be measured. Expanding functionality takes us to the following information, taken mainly from the ISO 9126 standard.

Functionality

Definition: the capability of the software product to provide functions which meet stated and implied needs when the software is used under specified conditions.

References: (ISO/IEC 2001:6.1)

Note: This characteristic is concerned with what the software does to fulfill needs whereas the other characteristics are mainly concerned with where and how it fulfils needs.

Functionality, as we have already seen, has a number of sub-nodes. Since we spent some time in the earlier part of this paper talking about accuracy, let us choose to expand that node. We find first that it breaks down into three sub-topics, fidelity, consistency and terminology. Of these we shall choose to expand fidelity. Fidelity is a terminal node in the whole FEMTI framework, and, as such, is required to carry information about how the system characteristic in question can be measured. Thus for the first time we find quite extensive information on pertinent metrics. We also find a definition, as always, and a set of notes.

Fidelity

Definition: Subjective evaluation of the degree to which the information contained in the original text has been reproduced without distortion in the translation (Van Slype 1979). Measurement of the correctness of the information transferred from the source language to the target language (Halliday in Van Slype's critical report).

Metrics:
· Carroll (in Van Slype's critical report): Rating of sentences read out of context on a 9-point scale
· Crook and Bishop (in Van Slype's critical report): Rating on a 25-point scale
· Halliday (in Van Slype's critical report): Assessment of the correctness of the information transferred
· Leavitt (in Van Slype's critical report): Rating of text units on a 9-point scale
· Miller and Beebe-Center (in Van Slype's critical report): Rating of a text on a 100-point scale
· Miller and Beebe-Center (in Van Slype's critical report): Shannon measurement of the quality of information transferred
· Sinaiko (in Van Slype's critical report): Re-translation
Van Slype (in Van Slype's critical report): Rating of sentences read on a 4-point scale
· Rating of 'adequacy' on a 5-point scale (White and O'Connell, 1994)
· BLEU evaluation toolkit (Papineni, et al. 2001): Automatic n-gram comparison of translated sentences with one or more human reference translations
· Rank-order evaluation of MT systems (Rajman and Hartley 2002): Correlation of automatically computed semantic and syntactic attributes of the MT output with human scores for fidelity (and adequacy and informativeness)
· Automated word-error-rate evaluation (Och, et al. 1999)
· Automated metric using head transducers (Alshawi, et al. 2000)

Notes: The fidelity rating has been found to be equal to or lower than the comprehensibility rating, since the unintelligible part of the message is not found in the translation. Any variation between the comprehensibility rating and the fidelity rating is due to additional distortion of the information which can arise from:
· Loss of information (silence) – example: word not translated
· Interference (noise) – example: word added by the system
· Distortion from a combination of loss and interference – example: word badly translated
Detailed analysis of the fidelity of a translation is very difficult to carry out, since each sentence conveys not a single item of information or a series of elementary items of information, but rather a portion of message or a series of complex messages whose relative importance in the sentence is not easy to appreciate.
Some automated metrics assume a fidelity evaluation as a human ground truth, or are relevant to fidelity evaluation.

Working through even a partial example of the contents of FEMTI has taken us into rather a lot of detail, although not, we hope, so much that the reader has been swamped. The essential point to hold on to is that each node in the framework can be expanded in two ways: one expansion leads to information pertinent to the node itself, typically including a definition, a list of stakeholders whose interests are connected to the node, literature references and a set of notes. The second expansion leads further into the taxonomy. Any node can indicate a passage from user needs to system characteristics, and the definition of quality characteristics bottoms out in (a choice of) metrics pertinent to the characteristic being described.

6 Future work

The example above gives a fairly clear idea of the current state of FEMTI. Much remains to be done. First, it is certain that there are gaps and that there are incoherencies. It is hoped that the MT community as a whole will help us to remedy these defects. Access to FEMTI via the web is free, and we hope easy. Each node is provided with a comment button: we hope that those who consult the framework, whether out of sheer curiosity or out of a desire to make use of the information it contains, will use the comment facility to signal weaknesses and to make suggestions.

In its current state, FEMTI is both catholic and agnostic. Emphasis has mainly gone into trying to be exhaustive, and comparatively little effort has been devoted to comparing or validating metrics. However, some of the ISLE workshops organised around the construction of FEMTI have stimulated the invention of new metrics and have encouraged comparison and cross validation of individual metrics. Much more could be done in this area, and again it is hoped that the community at large will contribute.

In an ideal world, the framework would be totally automated. That is, a would-be evaluator would be presented with a set of needs where he could check those relevant to him and perhaps indicate their relative importance, press a button and be presented with a recipe for carrying out his evaluation. This is a long way off and is perhaps not even totally feasible, even though the passage from evaluation requirements to system characteristics and from these to metrics shows how automation might be done. We hope to be able to investigate full automation further: achieving it serves as the guiding light towards which we strive.

7 Bibliography

Alshawi Hiyan, Srinivas Bangalore and Shona Douglas 2000, Head Transducer Models for Speech Translation and their Automatic Acquisition from Bilingual Data, *Machine Translation, 15*, 1, p. 105-124.

EAGLES MT Evaluation Working Group 1996, *EAGLES Evaluation of Natural Language Processing Systems*, Final Report Center for Sprogteknologi, EAG-EWG-PR.2 (ISBN 87-90708-00-8).

Hovy Eduard H. 1999, Toward Finely Differentiated Evaluation Metrics for Machine Translation, *Proceedings EAGLES Workshop on Standards and Evaluation*, Pisa, Italy.

Hovy Eduard H., Margaret King and Andrei Popescu-Belis in press, Principles of Context-Based Machine Translation Evaluation, *Machine Translation*, p. 38.

Hutchins John 2001, Machine translation and human translation: in competition or in complementation?, *International Journal of Translation, 13*, 1-2, p. 5-20.

ISO/IEC 1991, *ISO/IEC 9126: Information Technology -- Software Product Evaluation / Quality Characteristics and Guidelines for Their Use*, International Organization for Standardization / International Electrotechnical Commission, Geneva.

ISO/IEC 2001, *ISO/IEC 9126-1: Software Engineering -- Product Quality -- Part 1:Quality Model*, International Organization for Standardization / International Electrotechnical Commission, Geneva.

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

Oard D. and J. Gonzalez 2001, The Clef-2001 Interactive Track, Proceedings *Proceedings of the 2001 Cross-Language Evaluation Forum Workshop*, Darmstadt, Germany.

Och Franz Josef, Christoph Tillmann and Hermann Ney 1999, Improved Alignment Models for Statistical Machine Translation, *Proceedings Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, University of Maryland, College Park, MD, USA, p. 20-28.

Papineni Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu 2001, *BLEU: a Method for Automatic Evaluation of Machine Translation*, Research Report, Computer Science IBM Research Division, T.J.Watson Research Center, RC22176 (W0109-022).

Rajman Martin and Tony Hartley 2002, Automatic Ranking of MT *Systems, Proceedings Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas de Gran Canaria, Spain, volume 4, p. 1247-1253.

Sager J.C. 1978, Criteria for Machine Translation Evaluation, *Proceedings Workshop on Evaluation Problems in Machine Translation*, Luxemburg.

Slocum Jonathan 1987, METAL: the LRC machine translation system, *Machine Translation Today: The State of the Art, Proceedings of the Third Lugano Tutorial, 1984*, Edinburgh University Press / University of Texas, Edinburgh, UK, p. 319-350.

Van Slype Georges 1979, *Critical Study of Methods for Evaluating the Quality of Machine Translation*, European Commission / Directorate for General Scientific and Technical Information Management (DG XIII), BR 19142. *http://www.issco.unige.ch/projects/ isle/van-slype.pdf.*

White John S. and Theresa A. O'Connell 1994, The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches, *Proceedings AMTA Conference, 5-8 October 1994*, Association for Machine Translation in the Americas (AMTA), Columbia, MD, USA.