Work-in-Progress project report : CESTA - Machine Translation Evaluation Campaign

Widad Mustafa El Hadi
IDIST / CERSATES
Université de Lille 3
Domaine universitaire
du "Pont de Bois"
rue du Barreau
BP 149
59653 Villeneuve d'Ascq
Cedex - France
mustafa@univ-lille3.fr

Marianne Dabbadie
IDIST / CERSATES
Université de Lille 3
Domaine universitaire
du "Pont de Bois"
rue du Barreau
BP 149
59653 Villeneuve d'Ascq
Cedex - France
dabbadie@univ-lille3.fr

Ismaïl Timimi
IDIST / CERSATES
Université de Lille 3
Domaine universitaire
du "Pont de Bois"
rue du Barreau
BP 149
59653 Villeneuve d'Ascq
Cedex - France
timimi@univ-lille3.fr

Martin Rajman
LIA
Ecole
Polytechnique
Fédérale de Lausanne
Bât. INR
CH-1015 Lausanne
Switzerland
martin.rajman@epfl.ch

Philippe Langlais
RALI / DIRO -
Université de Montréal
C.P. 6128,
succursale Centre-ville
Montréal (Québec) -
Canada, H3C 3J7
felipe@IRO.UMontreal.CA

Antony Hartley
University of Leeds
Centre for Translation
Studies
Woodhouse Lane
LEEDS LS2 9JT
UK
a.hartley@leeds.ac.uk

Andrei Popescu Belis
University of Geneva 40
bvd du Pont d'Arve CH-
1211 Geneva 4
Switzerland
Andrei.Popescu-Belis@issco.unige.ch

Abstract

CESTA, the first European Campaign dedicated to MT Evaluation, is a project labelled by the French Technolangue action. CESTA provides an evaluation of six commercial and academic MT systems using a protocol set by an international panel of experts. CESTA aims at producing reusable resources and information about reliability of the metrics. Two runs will be carried out: one using the system's basic dictionary, another after terminological adaptation. Evaluation task, test material, resources, evaluation measures, metrics, will be detailed in the full paper. The protocol is the combination of a contrastive reference to: IBM "BLEU" protocol (Papineni, K., S. Roukos, T. Ward and Z. Wei-Jing, 2001); "BLANC" protocol derived from (Hartley, Rajman, 2002).; "ROUGE" protocol (Babych, Hartley, Atwell, 2003). The results of the campaign will be published in a final report and be the object of two intermediary and final workshops.

1 Introduction

1.1 CESTA and the Technolangue Action in France

This article is a collective paper written by the CESTA scientific committee that aims at presenting the CESTA evaluation campaign, a project labelled in 2002 by the French Ministry of Research and Education within the framework of the Technolangue call for projects and integrated to the EVALDA evaluation platform. It reports work in progress and therefore is the description of an on-going campaign for which system results are not yet available.

In France, EVALDA is the new Evaluation platform, a joint venture between the French Ministry of Research and Technology and ELRA (European Language Resources and Evaluation Association, Paris, France). Within the framework of this initiative eight evaluation projets are being conducted: ARCADE II: campagne d'évaluation de l'alignement de corpus multilingues; CESART: campagne d'Evaluation de Systèmes d'Acquisition de Ressources Terminologiques; CESTA : campagne d'Evaluation de Systèmes de Traduction automatique; Easy: Evaluation des Analyseurs Syntaxiques du français; Campagne EQueR, Evaluation en question-réponse; Campagne ESTER, Evaluation de transcriptions d'émissions radio; Campagne EvaSY, Evaluation en synthèse vocale; and Campagne MEDIA, Evaluation du dialogue hors et en contexte.

Regarding evaluation, the objectives of the Action as Joseph Mariani pointed out in his presentation at the LREC 2002 conference are to:

- Improve the present evaluation methodologies

- Identify new (quantitative and qualitative) approaches for already evaluated technologies: socio-technical and psycho-cognitive aspects
- Identify protocols for new technologies and applications
- Identification of language resources relevant for evaluation (to promote the development of new linguistic resources for those languages and domains where they do not exist yet, or only exist in a prototype stage, or exist but cannot be made available to the interested users);

The object of the CESTA campaign is twofold. It is on the one hand to provide an evaluation of commercial Machine Translation Systems and on the other hand, to work collectively on the setting of a new reusable Machine Translation Evaluation protocol that is both user oriented and accounts for the necessity to use semantic metrics in order to make available a high quality reusable machine translation protocol to system providers.

1.2 Object of the campaign

The object of the CESTA campaign is to evaluate technologies together with metrics, i.e. to contribute to the setting of a state of the art within the field of Machine Translation systems evaluation.

1.3 CESTA user oriented protocol

The campaign will last three years, starting from January 2003. A board of European experts are members of CESTA Scientific committee and have been working together in order to determine the protocol to use for the campaign. Six systems are being evaluated. Five of these systems are commercial MT systems and one is a prototype developed at the university of Montreal by the RALI research centre. Evaluation is carried out on text rather than sentences. Text approximate width will be 400 words. Two runs will be carried out. For industrial reasons, systems will be made anonymous.

2 State-of-the-art in the field of Machine Translation evaluation

In 1966, the ALPAC report draws light on the limits of Machine Translation systems. In 1979, the Van Slype report presented a study dedicated to Machine Translation metrics.

In 1992, the JEIDA campaign puts the user at the center of evaluator's preoccupation. JEIDA proposed to draw human measures on the basis of three questionnaires:

- One destined to users (containing a hundred questions)
- Other questionnaires are destined to system Machine translation systems editors (three different questionnaires),
- And a set of other questionnaires reserved to Machine Translation systems developers.

Scores are worked out on the background of fourteen categories of questions. From these scores, graphs are produced according to the answers obtained. A comparison of different graphs for each systems is used as a basis for systems classification.

The first DARPA Machine Translation evaluation campaign (1992-1994) makes use of human judgments. It is a very expensive method but interesting however, as regards the reliability of the evaluation thus produced. This campaign is based on tests carried out from French, Spanish and Japanese as source languages and English as a target language. The measures used for each of the following criteria are:

- Fidelity – a proximity distance is worked out between a source sentence and a target sentence on a 1 to 5 scale.
- Intelligibility, that corresponds to linguistic acceptability of a translation is measured on a 1 to 5 evaluation scale.
- Informativeness: the test is carried out on reading of the target text alone. A questionnaire on text informative content is displayed allowing to work out a measure calculated on the basis of the percentage of good answers provided in system translation.

In 1995, the OVUM report proposes to compare commercial Machine Translation systems on the basis of ten criteria.

In 1996, the EAGLES report (EAGLES, 1999) sets new standards for Natural Language

Processing software evaluation on the background of ISO 9126.

Initiated in 1999, and coordinated by Pr Antonio Zampolli, the ISLE project is divided into three working groups, one being a Machine Translation group.

Starting from ISO 9126 standard (King, 1999b), the aim of the project is to produce two taxonomies (c.f. section 3 of this article) and :

- One defining quality subcriteria with the aim of refining the six criteria defined by ISO 9126 (i.e. functionality, reliability, user-friendliness, efficiency, maintenance portability)
- The second one specifying use contexts that define the type of task induced the use of a by Machine Translation system, the types of users and input data. This taxonomy uses contextual parameters to select and order the quality criteria subject to evaluation. This taxonomy can be viewed and downloaded on the ISSCO website at the following address : http://www.issco.unige.ch/projects/isle/femti/

The second DARPA campaign (Papineni, K., S. Roukos, T. Ward and Z. Wei-Jing, 2001), making use of the IBM BLEU metric is mentioned in the CESTA protocol (c.f. section 8.1 of this article).

3 User-oriented evaluations

An emerging evaluation methodology in NLP technology focuses on quality requirements analysis. The needs and consequently the satisfaction of end-users, and this will depend on the tasks and expected results requirement domains, which we have identified as diagnostic quality dimensions. One of the most suitable methods in this type of evaluation is the adequacy evaluation that aims at finding out whether a system or product is adequate to someone's needs (see Sparck-Jones & Gallier, 1996 and King, 1996 among many others for a more detailed discussion of these issues). This approach encourages communication between users and developers.

The definition of the CESTA evaluation protocol took into account the Framework for MT Evaluation in ISLE (FEMTI), available online. FEMTI offers the possibility to define evaluation requirements, then to select relevant 'qualities', and the metrics commonly used to score them (cf. ISO/IEC 9126, 14598). The CESTA evaluation methodology is founded on a black box approach.

CESTA evaluators considered a generic user, which is interested in general-purpose, ready-to-use translations, preferably using an off-the-shelf system. In addition, CESTA aims at producing reusable resources, and providing information about the reliability of the metrics (validation), while being cost-effective and fast.

With these evaluation requirements in mind (FEMTI-1), it appears that the relevant qualities (FEMTI-2) are 'suitability', 'accuracy' and 'well-formedness'. Automated metrics best meet the CESTA needs for reusability, among which BLEU, X-score and D-score (chosen for internal reasons). Their validation requires the comparison of their scores with recognised human scores for the same qualities (e.g., human assessment of fidelity or fluency). 'Efficiency', measured through post-editing time, was also discussed. For the evaluation, first a general-purpose dictionary could be used, then a domain-specific one.

3.1 An approach based on use cases

ISO 14598 directives for evaluators put forth as a prequisite for systems development the detailed identification of user needs that ought to be specified through the use case document. Moreover, conducting a full evaluation process involves going through the establishment of an evaluation requirements document. ISO 14598 document specifies that quality requirements should be identified "according to user needs, application area and experience, software integrity and experience, regulations, law, required standards, etc.".

The evaluation specification document is created using the Software Requirement Specifications (SRS) and the Use-Case document. The CESTA protocol relies on a use case that refers to a translation need grounded on basic syntactic correctness and simple understanding of a text, as required by information watch tasks for example, and excludes making a direct use of the text for post editing purposes.

4 Two campaigns

4.1 Specificities of the CESTA campaign

Two campaigns are being organised :

The first campaign is organised using a system's default dictionary. After systems terminological adaptation a second campaign will be organised. Two studies previously carried out and presented respectively at the 2001 MT Summit (Mustafa El Hadi, Dabbadie, Timimi, 2001) and at the 2002 LREC conference (Mustafa Mustafa El Hadi, Dabbadie, Timimi, 2002) allowed us to realise the gap in terms in terms of quality between results obtained on target text after terminological enrichment.

4.2 First campaign

The organisation of the campaign implies going through several steps :

- Identification of potential participants
- Original protocol readjustement,
- The setting of a specific test tool that is currently being be implemented in conformity with protocol specifications validated by CESTA scientific committee. CESTA protocol specifications have been communicated to participants in particular as regards data formatting, test schedule, metrics and adaptation phase. For cost requirements, CESTA will not include a training phase. The first run will start during autumn 2004

4.3 Second campaign

The systems having already been tuned, an adaptation phase will not be carried out for the second campaign. However terminological adaptation will be necessary at this stage. The second series of tests being carried out on a thematically homogeneous corpus, the thematic domain only will be communicated to participants for terminological adaptation. For thematic adaptation, and in order to avoid system optimisation after the first series of tests, a new domain specific 200.000 word hiding corpus will be used.

The terminological domain on which evaluation will be carried out will then have to be defined. This terminological domain will be communicated to participants but not the corpus used itself. On the other hand, participants will be asked to send organisers a written agreement by which they will commit themselves to provide organisers with any relevant information regarding system tuning and specific adaptations that have made on each of the participating MT systems, in order to allow the scientific committee to understand and analyse the origin of the potential system ranking changes. The second run will start during year 2005.

Organisers have committed themselves not to publish the results between the two campaigns.

After the training phase, the second campaign will take place. Participants will be given a fifteen days delay to send the results. An additional three months period will be necessary to carry out result analysis and prepare data publication and workshop organisation.

CESTA scientific committee also decided in parallel with the two campaigns, to evaluate systems capacity to process formatted texts including images and HTML tags. Participants who do not wish to participate to this additional test have informed the scientific committee. Most of the time the reason is that their system is only capable of processing raw text. This is the case mainly for academic systems involved in the campaign, most of the commercial systems being nowadays able to process formatted text.

5 Contrastive evaluation

One of the particularities of the CESTA protocol is to provide a Meta evaluation of the automated metrics used for the campaign – a kind of state of the art of evaluation metrics. The robustness of the metrics will be tested on minor language pairs through a contrastive evaluation against human judgement.

The scientific committee has decided to use Arabic→French as a minor language pair. Evaluation on the minor language pair will be carried directly on two of the participating systems and using English as a pivotal language on the other systems. Translation through a pivotal language will then be the following : Arabic→English→French.

Organiser are, of course, perfectly aware of the potential loss of quality provoked by the use of a pivotal language but recall however that, contrarily to the major language pair, evaluation carried out on the minor language pair through a pivotal system will not be used to evaluate these systems

themselves, but metric robustness. Results of metric evaluation and systems evaluation will, of course, be obtained and disseminated separately.

During the tests of the first campaign, the French→English system obtaining the best ranking will be selected to be used as a pivotal system for metrics robustness Meta evaluation.

6 Test material

The required material is a set of corpora as detailed in the following section and a test tool that will be implemented according to metrics requirements and under the responsibility of CESTA organisers.

6.1 Corpus

The evaluation corpus is composed of 50 texts, each text length is 400 words to be translated twice, considering that a translation already exists in the original corpus. The different corpora are provided by ELRA. The masking corpus has 250.000 words and must be thematically homogeneous.

For each language pair the following corpora will be used:

Adaptation

- This 200.000 à 250.000 word corpus is a bilingual corpus. It is used to validate exchanges between organisers and participants and for system tuning.

First Campaign

- One 20.000 word evaluation corpus will be used (50 texts of 400 words each)
- One 200.000 to 250.000 word masking corpus that hides the evaluation corpus.

Second campaign

- One new 20.000 word corpus will be used but it will have to be thematically homogeneous (on a specific domain that will be communicated to participants a few months before the run takes place)
- One masking corpus similar to the previous one.

Additional requirement

The BLANC metric requires the use of a bilingual aligned corpus at document scale.

Three human translations will be used for each of the evaluation source texts. Considering that the corpora used, already provide one official translations, only two additional human translations will be necessary. These translations will be carried out under the organisers responsibility. Within the framework of CESTA use cases, evaluation is not made in order to obtain a ready to publish target language translation, but rather to provide a foreign user a simple access to information within the limits of basic grammatical correctness, as already mentioned in this article.

7 The BLEU, BLANC and ROUGE metrics

Three types of metrics will be tested on the corpus, the CESTA protocol being the combination of a contrastive reference to three different protocols:

7.1 The IBM "BLEU" protocol (Papineni, K., S. Roukos, T. Ward and Z. Wei-Jing, 2001).

The IBM BLEU metric used by the DARPA for its 2001 evaluation campaign, uses co-occurrence measures based on N-Grams. The translation in English of 80 Chinese source documents by six different commercial Machine Translation systems, was submitted to evaluation. From a reference corpus of translations made by experts, this metric works out quality measures according to a distance calculated between an automatically produced translation and the reference translation corpus based on shared N-grams (n=1,2,3…). The results of this evaluation are then compared to human judgments.

- NIST now offers an online evaluation of MT systems performance, i.e.:
    - A program that can be downloaded for research aims. The user then provides source texts and reference translations for a determined pair of languages.
    - An e-mail evaluation service, for more formal evaluations. Results can be obtained in a few minutes.

7.2 The "BLANC" protocol

It is a metric derived from a study presented at the LREC 2002 conference (Hartley A., Rajman M., 2002). We only take into account a part of the

protocol described in the referred paper, i.e. the X score, that corresponds to grammatical correctness.

We will not give an exhaustive description of this experience and shall only detail the elements that are relevant to the CESTA evaluation protocol.

The protocol has been tested on the following languages.

- Source language: French
- Target language: English
- Source corpus : 100 texts – domain : newspaper articles

Human judgements for comparison referential:

- 12 English monolingual students.
- No human translation reference corpus.
- Three criteria were tested: Fluency, Adequacy, Informativeness

Six systems were submitted to evaluation :
Candide (CD), Globalink (GL), MetalSystem (MS), Reverso (RV), Systran (SY), XS (XS)

- Each of the systems is due to translate a hundred source texts ranging from 250 to 300 words each. A corpus of 600 translations is thus produced.
- For each of the source texts, a corpus of 6 translations is produced automatically. These translations are then regrouped by series of six texts.
- According to the protocol initiated by (White & Forner, 2001) these series are then ranked by medium adequacy score.
- Every 5 series, a series is extracted from the whole. Packs of twenty series of target translations are thus obtained and submitted to human evaluators.

7.2.1 Evaluators' tasks

- Each evaluator reads 10 series of 6 translations i.e. 60 texts.
- Each of these series is then read by six different evaluators
- The evaluators must observe a ten minute compulsory break every two series.
- The evaluators do not know that the texts have been translated automatically.

The directive given to them is the following:
« rank these six texts from best to worst. If you cannot manage to give a different ranking to two texts, regroup them under the same parenthesis and give them the same score, as in the following example : 4 [1 2] 6 [3 5]."

The aim of this instruction is to produce rankings that are similar to the rankings attributed automatically.
Human judgement that ranks from best to worse corresponds in reality to a set of the fluency, adequacy and Informativeness criteria that can be attributed to the texts translated automatically.

7.2.2 Automatically generated scores

- X-score : syntactic score
- D-score : semantic score

Within the framework of the CESTA evaluation campaign the scientific committee decided to make use of the X-score only, the semantic D-score having proved to be unstable and that it could be advantageously replaced by the a metric based on (Bogdan, B.; Hartley, A.; Atwell, 2003), a reformulation of the D-score developed by (Rajman, M. and T. Hartley, 2001), and which we refer to as the ROUGE metric in this article.

7.2.3 X-score: definition

- This score corresponds to a grammaticality metric
- Each of the texts is previously parsed with XELDA Xerox parser.
- 22 types of syntactic dependencies identified through the corpus of automatic translations.
- The syntactic profile of each source document is computed. This profile is then used to derive the X-score for each document, making use of the following formula:
- X-score = (#RELSUBJ+#RELSUBJPASS-#PADJ-#ADVADJ)

7.3 The "ROUGE" protocol

This protocol, developed by Anthony Hartley in (Bogdan, B.; Hartley, A.; Atwell, 2003), is a semantic score. It is the result of a reformulation of the D-Score, the semantic score initiated through previous collaboration with Martin Rajman (Rajman, M. and T. Hartley, 2001), as explained in the previous section.

The original idea on which this protocol is based relies on the fact that MT evaluation metrics that "are based on comparing the distribution of statistically significant words in corpora of MT output and in human reference translation corpora".

The method used to measure MT quality is the following: a statistical model for MT output corpora and for a parallel corpus of human translations, each statistically significant word being highlighted in the corpus. On the other hand, a statistical significance score is given for each highlighted word. Then statistical models for MT target texts and human translations are compared, special attention being paid to words that are automatically marked as significant in MT outputs, whereas they do not appear to be marked as significant in human translations. These words are considered to be "over generated". The same operation is then carried out on "under generated words". At this stage, a third operation consists in the marking of the words equally marked as significant by the MT systems and the human translations. The overall difference is then calculated for each pair of texts in the corpora. Three measures specifying differences in statistical models for MT and human translations are then implemented : the first one aiming at avoiding "over generation", the second one aiming at avoiding "under generation" and the last one being a combination of these two measures. The average scores for each of the MT systems are then computed.

As detailed in (Bogdan, B.; Hartley, A.; Atwell, 2003):

"1. The score of statistical significance is computed for each word (with absolute frequency ≥ 2 in the particular text) for each text in the corpus, as follows:

$$S_{word[text]} = \ln \frac{\left(P_{word[text]} - P_{word[rest-corp]}\right) \times N_{word[txts-not-found]}}{P_{word[all-corp]}}$$

where:

$S_{word[text]}$ is the score of statistical significance for a particular word in a particular text□

$P_{word[text]}$ is the relative frequency of the word in the text;

$P_{word[rest-corp]}$ is the relative frequency of the same word in the rest of the corpus, without this text;

$N_{word[txt-not-found]}$ is the proportion of texts in the corpus, where this word is not found (number of texts, where it is not found divided by number of texts in the corpus)□

$P_{word[all-corp]}$ is the relative frequency of the word in the whole corpus, including this particular text

2. In the second stage, the lists of statistically significant words for corresponding texts together with their $S_{word[text]}$ scores are compared across different MT systems. Comparison is done in the following way:

For all words which are present in lists of statistically significant words both in the human reference translation and in the MT output, we compute the sum of changes of their $S_{word[text]}$ scores:

$$S_{text.diff} = \sum \left| \left(S_{word[text.reference]} - S_{word[text.MT]} \right) \right|$$

The score $S_{text.diff}$ is added to the scores of all "over-generated" words (words that do not appear in the list of statistically significant words for human reference translation, but are present in such list for MT output). The resulting score becomes the general "over-generation" score for this particular text:

$$S_{over-generation.text} = S_{text.diff} + \sum_{words.text} S_{word.over-generated[text]}$$

The opposite "under-generation" score for each text in the corpus is computed by adding $S_{text.dif}$ and all $S_{word[text]}$ scores of "under-generated" words – words present in the human reference translation, but absent from the MT output.

$$S_{under-generation.text} = S_{text.diff} + \sum_{words.text} S_{word.undergenerated[text]}$$

It is more convenient to use inverted scores, which increases as the MT system improves. These scores, $S_{o.text}$ and $S_{u.text}$, could be interpreted as scores for ability to avoid "over-generation" and "under-generation" of statistically significant words. The combined (o&u) score is computed similarly to the F-measure, where Precision and Recall are equally important:

$$S_{o.text} = \frac{1}{S_{over-generation.text}};$$

$$S_{u.text} = \frac{1}{S_{under-generation.text}};$$

$$S_{o\&u.text} = \frac{2 S_{o.text} S_{u.text}}{S_{o.text} + S_{u.text}}$$

The number of statistically significant words could be different in each text, so in order to make the scores compatible across texts we compute the average over-generation and under-generation scores per each statistically significant word in a given text. For the o_{text} score we divide $S_{o.text}$ by the number of statistically significant words in the MT text, for the u_{text} score we divide $S_{u.text}$ by the number of statistically significant words in the human (reference) translation:

$$o_{text} = \frac{S_{o.text}}{n_{statSignWordsInMT}};$$

$$u_{text} = \frac{S_{u.text}}{n_{statSignWordsInHT}};$$

$$u \,\&\, o_{text} = \frac{2 o_{text} u_{text}}{o_{text} + u_{text}}$$

The general performance of an MT system for IE tasks could be characterised by the average o-score, u-score and u&o-score for all texts in the corpus".

8 Time Schedule and result dissemination

The CESTA evaluation campaign started in January 2003 after having been labeled by the French Ministry of Research. During year 2003 CESTA scientific committee went through protocol detailed redefinition and specification and a time schedule was agreed upon.

2004 first semester is being dedicated to corpus untagging and the programming of CESTA evaluation tool. Reference human translations will also have to be produced and the implemented evaluation tool submitted to trial and validation.

After this preliminary work, the first run will start during autumn 2004. At the end of the first campaign, result analysis will be carried out. A workshop will then be organized for CESTA participants. Then the second campaign will take place at the end of Spring 2005, the terminological adaptation phase being scheduled on a five month scale.

After carrying out result analysis and final report redaction, a public workshop will be organized and the results disseminated and subject to publication at the end of 2005.

9 Conclusion

CESTA is the first European Campaign dedicated to MT Evaluation. The results of the campaign will be published in a final report and be the object of an intermediary workshop between the two campaigns and a final workshop at the end of the campaign.

It is a noticeable point that the CESTA campaign aims at providing a state of the art of automated metrics in order to ensure protocol reusability. The originality of the CESTA protocol lies in the combination and contrastive use of three different types of measures carried out in parallel with a Meta evaluation of the metrics.

It is also important to note that CESTA aims at providing a black box evaluation of available Machine Translation technologies, rather than a comparison of systems and interfaces, that can be tuned to match a particular need. If systems had to be compared, the fact that these applications should be compared including all software lawyers and ergonomic properties, ought to be taken into consideration.

Moreover apart from providing a state of the art through a Meta evaluation of the metrics used in its protocol, thanks to the setting of this original protocol that relies on the contrastive use of complementary metrics, CESTA aims at protocol reusability. One of the outputs of the campaign will be the creation of a Machine Translation evaluation toolkit that will be put at users and system developers' disposal.Acknowledgements

References

Besançon, R. and Rajman, M., (2002). Evaluation of aVector Space similarity measure in a multilingual framework. Procs. 3rd International Conference on Language Resources and Evaluation, Las Palmas,Spain,.1252

Bogdan, B.; Hartley, A.; Atwell E.; Statistical modelling of MT output corpora for Information Extraction Proceedings Corpus Linguistics 2003, Lancaster, UK, 28-31 March 2003, pp. 62-70

Chaudiron, S. Technolangue. In: http://www.apil.asso.fr/metil.htm, mars 2001

Chaudiron, S. L'évaluation des systèmes de traitement de l'information textuelle : vers un changement de paradigmes, Mémoire pour l'habilitation à diriger des recherches en sciences de l'information, présenté devant l'Université de Paris 10, Paris, novembre 2001

Dabbadie, M, Mustafa El Hadi, W., Timimi, I. (2001). Setting a Methodology for Machine Translation Evaluation. In: Machine Translation Summit VIII, ISLE/EMTA, Santiago de Compestela, Spain, 18-23 October 2001, pp. 49-54.

Dabbadie, M., Mustafa El Hadi, W., Timimi, I., (2002). Terminological Enrichment for non-Interactive MT Evaluation. In: LREC 2002 Proceedings – Las Palmas de Gran Canaria, Spain – 29th – 31st May 2002 – vol 6 – 1878-1884

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

EAGLES (1999). EAGLES Reports (Expert Advisory Group on Language Engineering Standards)http://www.issco.unige.ch/projects/eagles/ewg99.

ISLE (2001). MT Evaluation Classification, Expanded Classification. http://www.isi.edu/natural-language/mteval/2b-MT-classification.htm.

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

ISO (1999). Standard ISO/IEC 9126-1 Information Technology – Software Engineering – Quality characteristics and sub-characteristics. Software Quality Characteristics and Metrics - Part 1

ISO (1999). Standard ISO/IEC 9126-2 Information Technology – Software Engineering – Software products Quality : External Metrics - Part 2

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.

ISSCO (2001) Machine Translation Evaluation : An Invitation to Get Your Hands Dirty!, ISSCO, University of Geneva, Workshop organised by M. King (ISSCO) & F. Reed, (Mitre Corporation), April 19-24 2001.

King (1999a) EAGLES Evaluation Working Group, report,http://www.issco.unige.ch/projects/eagles.

King, M. (1999b). "ISO Standards as a Point of Departure for EAGLES Work in EELS Conference (European Evaluation of Language Systems), 12-13 April 1999.

Mariani, Joseph. «Language Technologies : Technolangue Action ». Presentation. In: LREC'2002 International Strategy Panel17, Las Palmas, May 2002.

Nomura, H. 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).

Popescu-Belis, A. 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.

Rajman, M. and T. Hartley, (2001). Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores. Procs. 4th ISLE Workshop on MT Evaluation, MT Summit VIII, 29-34.

Rajman, M. and T. Hartley, (2002). Automatic ranking of MT systems. In: LREC 2002 Proceedings – Las Palmas de Gran Canaria, Spain – 29th – 31st May 2002 – vol 4 – 1247-1253

Reeder, F., K. Miller, J. Doyon, and J. White, J. (2001). The naming of things and the confusion of tongues: an MT metric. Procs. 4th ISLE Workshop on MT Evaluation, MT Summit VIII, 55-59.

Sparck-Jones K., Gallier, J.R. (1996). Evaluating Natural Language Processing Systems: An Analysis and Review, Springer, Berlin.

TREC, NIST Website, last updated, August 1st, 2000, visited by the authors, 23-03-2003

Vanni, M. and K. Miller (2001). Scaling the ISLE framework: validating tests of machine translation quality for multi-dimensional measurement. Procs. 4th ISLE Workshop on MT Evaluation, MT Summit VIII, 21-27.

VanSlype., G. (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).

Véronis, J., Langlais, Ph. (2000). ARCADE: évaluation de systèmes d'alignement de textes multilingues. In Chibout, K., Mariani, J., Masson, N., Neel, F. éds., (2000). Ressources et évaluation en ingénierie de la langue, Duculot, Coll. Champs linguistiques, et Collection Universités Francophones (AUF).