Evaluation of Machine Translation with Predictive Metrics beyond BLEU/NIST: CESTA Evaluation Campaign # 1

Sylvain Surcin
HLT Evaluation Department
ELDA - 55-57 rue Brillat-Savarin
F75013 Paris - France
surcin@elda.org

Olivier Hamon
HLT Evaluation Department
ELDA - 55-57 rue Brillat-Savarin
F75013 Paris - France
hamon@elda.org

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

Martin Rajman
LIA
Ecole Polytechnique Fédérale de Lausanne Bât. INR
CH1015 Lausanne – Switzerland
martin.rajman@epfl.ch

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

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

Ismaïl Timimi
IDIST / CESARTES
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

Marianne Dabbadie
IDIST / CESARTES
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

Khalid Choukri
ELDA - 55-57 rue Brillat-Savarin
F75013 PARIS France
choukri@elda.org

Abstract

In this paper, we report on the results of a full-size evaluation campaign of various MT systems. This campaign is novel compared to the classical DARPA/NIST MT evaluation campaigns in the sense that French is the target language, and that it includes an experiment of meta-evaluation of various metrics claiming to better predict different attributes of translation quality. We first describe the campaign, its context, its protocol and the data we used. Then we summarise the results obtained by the participating systems and discuss the meta-evaluation of the metrics used.

1 Introduction

This paper aims to present the CESTA evaluation campaign, a French national project already introduced by (Mustafa el Hadi *et al.*, 2004) and focusing on the evaluation of Machine Translation systems.

The objectives of CESTA are many. First, it aims at reproducing classical evaluation campaigns as conducted by (NIST, 2003) with French, instead of English, as the target language. But more than that, one of its two main purposes is to experiment with a wider range of (semi-)automatic metrics than BLEU/NIST alone, considering types of information other than statistics over *n*-grams: in the event, syntactic and semantic information.

The other major purpose of CESTA is to conduct a meta-evaluation of the set of selected metrics, comparing their results with human judgements of fluency and adequacy, in order to detect the best correlation rates between the results of the human judgements and the metrics, taken separately or in combination.

This paper first introduces the CESTA project, its context, its objectives and its specificities compared to previous MT evaluation campaigns. Then it describes the first CESTA evaluation campaign conducted in 2004-2005. We detail the protocol of this campaign, the evaluation material

and the participating systems. Finally, we present the results obtained from the selected metrics, and the results of the meta-evaluation of the metrics in relation to human judgements. The conclusion takes the form of a critical review of the experience gained through this first campaign, in order to prepare the second evaluation campaign in the second half of 2005.

2 CESTA Evaluation Campaigns

2.1 Context

CESTA is a project approved in 2002 by the French Ministry of Research and Education within the Technolangue[1] framework. CESTA is integrated to the EVALDA evaluation platform, which aims at providing a reusable evaluation platform for eight major NLP technologies (multilingual corpora alignment, terminology extraction, machine translation, syntactic parsing, question answering, voice recognition, speech synthesis and dialogue systems).

The objectives of CESTA are twofold. It aims on the one hand to provide an evaluation of commercial and academic MT systems, and on the other to work collectively on the setting up of a new, re-usable MT evaluation protocol that is user-oriented and accommodates new metrics (relying on semantics and syntax). The motivation is to investigate if such approaches might prove more accurate and/or less costly than *n*-gram matching.

The CESTA project started in January 2003 and will last until December 2005. A panel of European experts are members of CESTA scientific committee, and have been working together to determine the CESTA protocol.

2.2 Campaign Schedule

Two evaluation campaigns are planned within CESTA. The first evaluation campaign, on which we report here, was set up to evaluate the participating systems' technological core without adapting user (terminological) dictionaries, i.e. with a default dictionary.

The second evaluation campaign – to take place after summer 2005 – will be organised after a terminological adaptation phase of the systems, since previous studies point out the gap in terms of translation quality between results obtained on target text with and without terminological enrichment (Mustafa El Hadi *et al.*, 2001, 2002; Babych *et al.*, 2004).

2.3 CESTA Protocol

The definition of the CESTA protocol took into account FEMTI, the Framework for MT Evaluation in ISLE (Hovy, King and Popescu-Belis, 2002), which offers the possibility of defining evaluation requirements, and then selecting relevant "qualities" and the metrics commonly used to score them, cf. ISO/IEC 9126 (ISO, 1999) and 14598 (ISO, 2001).

2.4 Test Corpora

The test corpora for both translation directions in the first campaign contained around 20,000 source words (20,658 words for English to French, and 23,763 words for Arabic to French). The English set consisted of 15 documents selected from the JOC corpus (Journal of the European Communities), from the Questions to the European Parliament sessions of 1993. The Arabic set consisted of 16 documents selected from the UNESCO 32nd General Conference[2]. All documents were segmented at the sentence level, amounting to 790 segments in English and 298 segments in Arabic.

The documents of the test corpora were selected according to two criteria. Firstly, they should not pertain to a specific thematic area (i.e. their lexical coverage should include a minimum of technical or restricted terminology). And secondly, an authoritative French version should be available. This version constituted a first reference translation.

These documents were randomly dispersed within "masking corpora" of more than 200,000 words, both for English and for Arabic. The documents of the masking corpora were chosen to have a similar style to the test corpora to prevent the participants to isolate the test corpora and perform a manual translation or proofreading phase. The English masking corpus consisted of documents selected from the Economics and Diplomatic sections of the Financial Times

[1] http://www.technolangue.net

[2] Both corpora are available in ELDA's catalogue.

newspaper[3], and the Arabic masking corpus of documents taken from the Economics and Diplomatic sections of the Al-Hayat newspaper[4].

Thus, each evaluation corpus consisted in a test corpus, on which the scoring and judgements were performed, and a masking corpus, which was discarded. Each evaluation corpus was presented as a single file, raw text (no HTML nor rich-text tags), UTF-8 encoded, following the NIST MT format (NIST, 2003).

For each test corpus, three additional reference translations were obtained from professional translation agencies. Each reference translation was produced by a distinct translator team, without contact with the other teams. The translators were asked to stick to the source text (not to rewrite the translations). The translation guidelines were inspired by those of the last NIST MT evaluation for Arabic-to-English and Chinese-to-English. As a result, four French reference translations were available for the English test set, and another four French reference translations were available for the Arabic test set.

2.5 Selected Metrics

2.5.1 BLEU/NIST

BLEU (Bilingual Evaluation Understudy) is a semi-automated metric tuned by the US National Institute of Standards (NIST) and first developed by IBM (Papineni et al., 2001).

In simple terms, BLEU counts the number of word *n*-grams in a sentence to be evaluated which are common with one or more reference translations. A translation is considered better if it shares a greater number of *n*-grams with the reference translations. In addition, BLEU applies a penalty to those translations whose length differs significantly from that of the reference translations.

The NIST metric is an alternative to BLEU. Whereas BLEU computes the geometric mean of *n*-grams precision ($1 \leq n \leq N$) with a positive weighting, NIST computes the arithmetic mean of *n*-grams precision taking into account a comparison of the length of segments.

BLEU/NIST is in the most widespread use nowadays in the MT community. BLEU scores proved to correlate with human judgements about the fluency (Thompson and Brew, 1994) of the evaluated translation (Zhang *et al.*, 2004).

2.5.2 WNM

The Weighted N-gram Model, or WNM (Babych, 2004), is a combination of BLEU and the Legitimate Translation Variation, or LTV, metrics (Babych and Hartley, 2004a, 2004b).

For a given source text, more than one correct translation is possible. BLEU tries to cope with this by multiplying the number of reference translations to be compared to the evaluated one. But still, the fact that some *n*-gram does not occur in any reference does not mean that it is an erroneous translation, providing the meaning is the same.

Babych and Hartley's proposal is to extend BLEU and the computation of proximity scores (i.e. the distance measure between the evaluated translation and the references) by introducing weights coming from the statistical relevance of the words inside the text. Words statistically more salient will get a greater weight. The statistical salience is computed comparing the occurrence frequency of the words within segments of text and within the corpus as a whole. This computation relies on the *tf.idf* score, usually used in Information Retrieval, plus a normalisation according to the words' relative frequency (Babych *et al.*, 2003).

Typically, words such as names, events, terminological lexemes, are statistically more salient. They can be translated in a unique way only, whereas function words or expressions can have several possible correct translations.

A preliminary experiment (Babych *et al.*, 2004) proved that WNM results for recall were well correlated (even better than BLEU) to human judgements about adequacy. This was confirmed by (Babych and Hartley, 2004b).

2.5.3 X-Score

The X-Score metric (Rajman and Hartley, 2001) is based on the distribution of elementary linguistic information within a text, such as morpho-syntactic categories, or syntactic relationships. The authors' hypothesis is that this distribution of linguistic information is similar from one text to another within a given language.

[3] Part of the MLCC corpus, available at ELDA.

[4] "Arabic Data Set" corpus, available at ELDA.

Depending on the nature of the linguistic information selected to work with, the metric's precision will vary. For instance, working with syntactic dependencies will be much more precise that working with morphosyntactic categories only. Whichever type of information is selected, the X-Score measures the grammaticality of a text, comparing the distribution of the selected linguistic information within this text to a representative measure of the same information distribution within the whole language.

At an initial learning stage, a typical representation of the linguistic information within a very large corpus (representing the language) is computed. The corpus is composed of documents for which a fluency score is available. Fluency is used because it is held to be very similar to grammaticality. Then the frequencies of the different categories of the selected linguistic information are computed and used to train a linear predictor able to compute a predicted fluency score for any new input frequency list.

This linear predictor will then be used for evaluation. It is noteworthy that only the target language and the target translated documents have to be modelled.

This metric remains experimental and it can be expected to be highly dependent on many parameters. In particular, it depends on the nature of the selected linguistic information, on the tool used to extract this information, and on the training corpus, to cite only the main factors.

CESTA will investigate this metric for different types of linguistic information, and with different tools.

2.5.4 D-Score

The D-Score (Rajman and Hartley, 2001) measures the preservation of a text's semantic content throughout the translation process.

First the authors create semantic vector space models, of both the source language and of the target language. Then the "position" of any given source document is computed within the source language vector space, and the "position" of its translation is also computed within the target language vector space. Finally, the distance between these two positions is used to compute the D-Score measure.

The authors' hypothesis is that the translation process is semantically conservative. This means that the structure of a source text's semantic vector representation (i.e. its position within the source language model) shall be preserved and be almost identical to the structure of its translation's semantic vector representation (its position within the target language model).

It is still a highly experimental metric, subject to high variations due to a number of parameters. In particular, it is highly dependent on the method used to reduce the representation space of the terms, the usage of tools such as stemmers or lemmatisers to normalise the terms, and the training corpus used to build the source and target language models.

2.6 Participants

The participants in the CESTA Evaluation Campaign # 1 were a mix of commercial and academic organisations. Participation was already opened to organisations outside of the project consortium, and one external organisation participated in the English-to-French task.

There were five participants to the English-to-French task: Systran, Softissimo (with the Reverso system), SDL International, the RALI Laboratory of the University of Montréal, and Comprendium S.L.

There were two participants to the Arabic-to-French task: Systran and CIMOS.

The systems are rule-based, except RALI's system, which is statistical.

The results presented hereafter are anonymised.

2.7 Human Judgements

The CESTA protocol comprises a human judgement phase. All submitted translations were evaluated by two human judges, for both adequacy and fluency.

A total of 112 human judges participated in the human judgement phase. They were paid a modest sum for their participation, and were recruited among students of French universities, and through advertising on specialised e-mail distribution lists. We did not take into account their age nor gender, but rather focused on the fact they were all native French speakers and not native speakers of English or of Arabic (in order not to introduce favourable biases).

The inter-judge agreement was checked on the initial downsized evaluation run by ELDA, using a Pearson's correlation test. For adequacy, correlation was $R = 0.9793$ with confidence interval at 95% of $0.978 < R < 0.981$. For fluency the correlation was $R = 0.9574$ with a confidence interval at 95% of $0.953 < R < 0.961$. The inter-judge agreement will be recomputed at the end of the Evaluation Run # 1 human judgement phase.

For fluency, the judges are asked to answer the question "is this text written in good French?" by giving a score on a 5-grades scale from "native French" to "non understandable".

For adequacy, they are asked to compare the meaning of the evaluated segment to that of a reference translation and score adequacy on a 5-grade scale from "whole meaning is present" to "nothing in common".

The judges are provided with guidelines before they start. These guidelines stipulate that they must react as instinctively as possible and not spend more than 30 seconds in total on any segment.

To distribute the segments among the judges, all the submitted translations of the two tasks are merged (as they all are in French). Then the following procedure is applied:

1. Create 2 sets of tokens
   a. Set 1 contains one token per judge.
   b. Set 2 contains one token per translated document (i.e. the translation of a given document by a given system).
2. As long as there remains a token in set 2
   a. Select a translation in set 2.
   b. Select two judges in set 1. If it is empty, fill it anew.
   c. Assign the evaluation of this translation to those judges.

A reasonable constraint is that a maximum of around 80 segments in total are assigned to a judge, in order not to need more than 1 hour to judge all segments. In our campaign, the total number of translated segments was 4,546 (both translation directions). Given our self-imposed time constraint, we needed 112 judges, each one being assigned between 81 and 82 segments.

A special web application was developed for human judgements. They could be done remotely, from any computer with a web browser connected to the Internet.

2.8 Schedule

A first dry-run took place in August 2004 with real size corpora. This dry-run aimed at checking the protocol and data flows and formats between the Evaluation agency (ELDA) and the participants. For this reason, no reference translations of the dry-run corpora were produced on this occasion.

About the same time, a downsized campaign was organised within ELDA in order to check the whole feasibility of the evaluation and scoring process.

Evaluation Campaign # 1 took place in February 2005. Participants had one week to return to ELDA their translations, which were automatically scored straight away. The human judgement phase (for meta-evaluation) required much more time and took place between March and June 2005.

3 Evaluation Results

The results for English-to-French and Arabic-to-French (respectively labelled 'EN' and 'AR') are presented together.

First, we present BLEU/NIST using 4-gram precision. This precision proved to best correlate with human judgements of fluency (Zhang *et al.* 2004) for English texts, and it will be one of the meta-evaluation goals to find the value of *n* for optimal n-gram precision for French. The results are presented with and without case sensitivity (Tables 1 and 2). More generally, all results are presented with a confidence interval of 0.7%.

Table 1: BLEU Results using 4-grams precision

System	4-gram with case sensitive	4-gram Without case sensitive
System 1 - EN	0.26	0.27
System 2 - EN	0.28	0.30
System 3 - EN	0.20	0.21
System 4 - EN	0.27	0.28
System 5 - EN	0.41	0.43
System 6 - AR	0.06	0.07
System 7 - AR	0.01	0.02

Table 2: NIST Results using 4-grams precision

System	4-gram With case sensitive	4-gram Without case sensitive
System 1 - EN	0.17	0.17
System 2 - EN	0.18	0.20
System 3 - EN	0.13	0.14

System 4 - EN	0.17	0.18
System 5 - EN	0.24	0.26
System 6 - AR	0.04	0.05
System 7 - AR	0.01	0.01

Table 3 presents WNM results, with precision, recall, and fluency scores. WNM evaluation was performed with only one reference translation, for each translation direction.

Table 3: WNM Results

System	WNM Precision	WNM Recall	WNM Fluency
System 1 - EN	2.26	0.51	0.83
System 2 - EN	2.27	0.56	0.90
System 3 - EN	2.06	0.47	0.77
System 4 - EN	2.29	0.54	0.88
System 5 - EN	2.35	0.59	0.94
System 6 - AR	0.44	0.61	0.51
System 7 - AR	0.41	0.79	0.54

Table 4 shows results of the two experimental metrics: D-Score and X-Score.

Table 4: D-Score & X-Score Result

System	D-Score	X-Score
System 1 - EN	0.016	0.407
System 2 - EN	0.019	0.394
System 3 - EN	0.022	0.391
System 4 - EN	0.014	0.418
System 5 - EN	0.019	0.420
System 6 - AR	0.016	0.383
System 7 - AR	0.015	0.391

4 Meta-evaluation of Metrics

4.1 Automatic Meta-evaluation

In order to compare metrics, Table 5 presents the results we have obtained with WER and PER metrics, and Table 6 proposes a system ranking in order to see differences between systems.

Table 5 WER & PER Results

System	WER	PER
System 1 - EN	67.74	48.13
System 2 - EN	65.23	41.87
System 3 - EN	70.92	50.01
System 4 - EN	71.70	44.19
System 5 - EN	64.46	44.40
System 6 - AR	74.63	41.35
System 7 - AR	98.53	49.13

Table 6: System Ranking

System	En→Fr					Ar→Fr	
	1	2	3	4	5	6	7
BLEU	4	2	5	3	1	1	2
NIST	4	2	5	3	1	1	2
WNM fluency	4	2	5	3	1	2	1
WNM recall	4	2	5	3	1	2	1
D-Score	4	2	1	5	2	1	2
X-Score	3	4	5	2	1	2	1
WER	3	2	4	5	1	1	2
PER	4	1	5	2	3	1	2

As we can see, the results of BLEU, NIST and the two WNM scores are correlated, but only for English-to-French. Hence we have two possibilities: either Arabic-to-French produces fewer salient words, which invalidates the WNM metric, or the WNM statistical salience of words produces more pertinent scores.

The BLEU/NIST metrics are more closely correlated with the WNM precision score than with the two other WNM scores.

In the same way, the WNM recall score is close to the X-Score, which is another recall evaluation metric.

4.2 Correlation with Human Judgements

Table 7 shows the decreasing ranking established by the judges for fluency and adequacy criteria. It is noticeable that both criteria are correlated: the correlation between human judgements of fluency and adequacy is 0.538 ± 0.133 (95% confidence).

This leads us to wonder if the method we used to prompt the judges for their judgment was really sound. The fact that the interface first asked for the fluency judgment, then for the adequacy (still displaying the segment) could have introduced a precedence bias.

Table 7: System Human Ranking

Rank	Fluency	Adeq.
1st – EN	System 4—EN	System 4—EN
2nd - EN	System 5 – EN	System 5 – EN
3rd - EN	System 1 – EN	System 1 – EN
4th - EN	System 2 – EN	System 2 – EN
5th - EN	System 3 – EN	System 3 – EN
1st - AR	System 6 – AR	System 6 – AR
2nd - AR	System 7 - AR	System 7 - AR

Tables 8 and 9 show the decreasing mean values for fluency and adequacy together with the results of the automatic metrics. The ranks given by the metrics are shown between parentheses.

Table 8: Human Ranking vs Statistical Metrics

S.	Fl.	Ad.	BLEU	NIST	WNM
S1	0.46	0.56	0.44 (4)	9.74 (4)	2.26 (4)
S2	0.42	0.54	0.49 (2)	10.22 (2)	2.27 (3)
S3	0.35	0.49	0.39 (5)	9.19 (5)	2.06 (5)
S4	0.51	0.64	0.46 (3)	9.97 (3)	2.29 (2)
S5	0.50	0.61	0.59 (1)	11.28 (1)	2.35 (1)
S6	0.20	0.31	0.21 (1)	7.42 (1)	0.44(1)
S7	0.08	0.17	0.09 (2)	4.79 (2)	0.35(2)

Only the values for WNM precision are given here for brevity's sake, as it best correlates with human judgements.

Table 9: Human Ranking vs Knowledge Metrics

Sys.	Fl.	Ad.	X-Score	D-Score
S1	0.46	0.56	0.407 (3)	0.0159 (4)
S2	0.42	0.54	0.394 (4)	0.0186 (2)
S3	0.35	0.49	0.391 (5)	0.0222 (1)
S4	0.51	0.64	0.418 (2)	0.0139 (5)
S5	0.50	0.61	0.420 (1)	0.0186 (2)
S6	0.20	0.31	0.383 (2)	0.0156 (1)
S7	0.08	0.17	0.391 (1)	0.0154 (2)

In these results, the X-Score was implemented using Treetagger (Schmid, 1994) for POS tagging, and the D-Score was implemented with lemmatisation only to reduce the representation space.

A number of analyses and experiments are still being conducted on the collected data, but we can already present the 1-to-1 correlations between the human judgements and the scores from the automatic metrics.

Table 10: Correlations

Metrics	Fluency	Adequacy	Fluency / Adequacy Ranking
BLEU (1-gram)	0.57	0.50	0.3
BLEU (2-gram)	**0.72**	**0.66**	0.5
BLEU (3-gram)	0.71	0.64	0.5
BLEU (4-gram)	0.70	0.64	0.5
BLEU (sum)	0.68	0.62	0.5
NIST (1-gram)	0.69	0.64	0.5
NIST (2-gram)	**0.75**	**0.69**	**0.8**
NIST (3-gram)	0.66	0.59	0.5
NIST (4-gram)	0.70	0.64	0.5
NIST (sum)	0.71	0.65	0.5
WNM precision	**0.89**	**0.84**	**0.8**
WNM recall	0.69	0.69	0.5
WNM fluency	0.72	0.71	0.5
X-Score	0.95	0.93	0.9
D-Score	-0.81	-0.82	-0.9

First, we observe that in our experiments, BLEU and NIST proved better correlated for 2-grams with human judgements than for longer N-grams. But the major observation is that, among the 3 statistical metrics we used, the better correlated (and hence the most predictive) with the human judgments is the precision score for WNM.

Second, we observe that the X-Score correlated closely with human judgements, while the D-Score is inversely correlated.

5 Prospects

CESTA is the first European evaluation campaign dedicated to MT. The complete results of the two campaigns will be published in a final report and will be the object of a public workshop at the end of the campaign. It is noteworthy that CESTA aims at providing state-of-the-art 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 MT technologies, rather than a comparison of systems and interfaces that can be tuned to match a particular need. If technologies rather than outputs had to be compared, all their software layers and ergonomic properties should be taken into consideration.

The second CESTA evaluation campaign will be conducted in September-October 2005, and will be open to external participants. The evaluation protocol will be revised by the project's scientific committee in order to take into account the experience learnt from the first campaign. For example, we are already reconsidering the wording of the instructions given to the judges for the fluency evaluation task.

In addition, the evaluation material will consist of texts with a strong lexical specialisation (texts

describing a technical specialism, with its own terminology). The terminological domain on which the evaluation will be carried out will be communicated to the participants. If required by participants using a statistical MT system, a training corpus of the domain will be provided.

The participants will be asked to commit to provide the organisers with any relevant information regarding system tuning and specific adaptations they have made.

Organisations interested in participating in this next campaign are invited to contact ELDA.

After each campaign, ELDA will issue *evaluation packages*. These packages will consist of DVDs containing all the materials (corpus, tools, documentation and results of the campaign) necessary for reproducing the campaign independently.

6 Bibliographical References

Babych B., 2004. Weighted N-gram model for evaluating Machine Translation output. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*. Birmingham, January 2004. pp. 15-22.

Babych B., Elliott D. Hartley A., 2004. Calibrating resource-light automatic MT evaluation: a cheap approach to ranking MT systems by the usability of their output. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, May 2004. pp. 2031-2034.

Babych B., Hartley A. 2004a. Modelling legitimate translation variation for automatic evaluation of MT quality. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, May 2004. pp. 833-836.

Babych B., Hartley A. 2004b. Extending the BLEU MT Evaluation Method with Frequency Weightings. In *ACL 2004 Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, July 2004. pp. 622-629.

Babych B., Hartley A., Atwell E., 2003. Statistical modelling of MT output corpora for Information Extraction. In *Proceedings of the Corpus Linguistics 2003 Conference (COLING)*, ed. Archer D., Rayson P., Wilson A., McEnery T. Lancaster, March 2003. pp. 62-70.

I Hovy E., King M., Popescu-Belis A., 2002. Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, vol. 17, n. 1, p.43-75.

ISO 1999. Standard ISO/IEC 9126. Part 1: Information Technology – Software Engineering – Quality characteristics and sub-characteristics. Software Quality Characteristics and Metrics. Part 2: Information Technology – Software Engineering – Software Products Quality: External Metrics.

Mustafa El Hadi W., Timimi I., Dabbadie M., 2001. Setting a Methodology for Machine Translation Evaluation. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, MT Summit VIII*, Santiago de Compostela, September 2001. pp. 49-54.

Mustafa El Hadi W., Timimi I., Dabbadie M., 2002. Terminological Enrichment for non-Interactive MT Evaluation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas de Gran Canarias, May 2002. pp. 1878-1884.

Mustafa El Hadi W., Dabbadie M., Timimi I., Rajman M., Langlais P., Hartley A., Popescu-Belis A., 2004. CESTA – Machine Translation Evaluation Campaign. In *Proceedings of the LR4Trans Workshop of the 20th International Conference on Computational Linguistics, COLING'2004,* Geneva, August 2004. pp. 8-17.

NIST, 2003. The 2004 NIST Machine Translation Evaluation Plan (MT-04), v2.1. http://www.nist.gov/speech/tests/mt.

Papineni K., Roukos S., Ward T. and Zhu W.-J. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation, *IBM Research Report RC22176 (W0109-022)*. In *Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL)*, Philadelphia, July 2002, pp. 311-318.

Rajman M., Hartley A., 2001. Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, MT Summit VIII*, Santiago de Compostela, September 2001. pp. 29-34.

Schmid H., 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, 1994.

Thompson H., Brew C. 1994. Automatic Evaluation of Computer Generated Text. In *Proceedings of the ARPA/ISTO Workshop on Human Language Technology*, 1994. pp. 104-109.

Zhang Y., Vogel S., Waibel A. 2004. Interpreting BLEU/NIST Scores: How Much Improvement? Do We Need to Have a Better System? In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, May 2004. pp. 2051-2054.