Assessing Human and Automated Quality Judgments in the French MT Evaluation Campaign CESTA

Olivier Hamon
ELDA
55-57, rue Brillat-Savarin
75013 Paris, France, and
LIPN, U. of Paris XIII
99, av. J.-B. Clément
93430 Villetaneuse, France
hamon@elda.org

Anthony Hartley
Centre for Translation Studies
University of Leeds
Leeds LS2 9JT
UK
a.hartley@leeds.ac.uk

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

Khalid Choukri
ELDA
55-57, rue Brillat-Savarin
75013 Paris
France
choukri@elda.org

Abstract

This paper analyzes the results of the French MT Evaluation Campaign, CESTA (2003-2006). The details of the campaign are first briefly described. The paper then focuses on the results of the two runs, which used human metrics, such as fluency or adequacy, as well as automated metrics, mainly based on n-gram comparison and word error rates. The results show that the quality of the systems can be reliably compared using these metrics, and that the adaptability of some systems to a given domain – which was the focus of CESTA's second run – is not strictly related to their intrinsic performance.

Introduction

The French MT evaluation campaign, CESTA[1], completed its two phases in 2006. The goal of the campaign was to evaluate the output quality of commercial and academic systems translating into French from English and Arabic, and to assess their adaptability to a new subject domain. CESTA also studied the reliability of automatic evaluation metrics with French as a target language and produced a number of reusable language resources for MT evaluation.

This article analyzes the scores and rankings of the systems in various conditions, from a meta-evaluation point of view. One of the main questions that are discussed is the level of agreement between human judgments of translation quality and the scores of the automated metrics. While such metrics have been developed and studied for English as a target language, this article discusses their applicability to French. Other meta-evaluation questions include the reliability of human scores, the influence of reference translations on evaluation results, and the use of automated metrics to analyze reference translations.

The article is organized as follows. First, the overall organization of the campaign is outlined, focussing on the human and automated evaluation metrics that were used. Then, the data used for the two runs is described, with automated evaluation scores being used to estimate the variability of the reference translations. The reliability of the scores is finally discussed, first intrinsically for human scores, and then in terms of correlation of automatic scores with human scores. The results show that the automatic scores are consistent with human ones, and that they are able to indicate reliably the ranking of the systems and their capacity to adapt to a new domain.

Overview of CESTA

The CESTA MT evaluation campaign (2003-2006), the first such campaign organized with French as a target language, had several objectives (Hamon et al. 2007). As stated above, the first goal was to define an evaluation protocol which includes human and automated quality metrics, and to assess its reliability on the EN/FR and AR/FR language pairs.

Two runs were organized. The first one aimed at evaluating output quality, from absolute and comparative points of view, on a general-domain reference corpus. In addition, the concrete details of the evaluation protocol were also validated and improved after the first run. The second run aimed at measuring the capacity of systems to adapt to a new domain in a very limited amount of time. The participants received "adaptation" data from a specific domain (health), and were asked to provide translation of the test data first without using the adaptation data, and then by using it to improve their systems' performances by terminological enrichment (Mustafa Hadi et al. 2002, 2004; Babych et al. 2004).

Participating Systems and Their Coding

Thirteen systems or versions of systems (Hamon et al. 2007) took part in one or both CESTA runs[2]. In order to preserve the anonymity of the results, we will adopt the CESTA notational conventions: we will refer to the systems using indices from S1 to S13, numbered in a continuous sequence throughout the two campaigns, regardless of the fact that some systems participated in both runs (so, S_x and S_y may sometimes refer to the same system). This means that it is impossible to identify the same system in each of the two campaigns; since resources and even versions changed from one campaign to the next, it would be in any case misleading to compare performances of the 'same' system in the two runs. In addition, to make notations more readable, we specify in brackets appended after the system's code, its source language ('en' or 'ar') and the number of the campaign (1, 2a or 2b) in which it participated.

[1] CESTA stands for *Campagne d'Évaluation des Systèmes de Traduction Automatique*, and was supported by the Technolangue program of the French Government.

[2] The names of the systems and/or the organizations they originate from are the following: Comprendium (Translendium SL), MLTS (CIMOS), Ramses (RALI, U. of Montreal), Reverso (Softissimo), RWTH Aachen, SDL Enterprise Translation Server, Systran, and the Polytechnic U. of Catalonia.

The first campaign thus involved systems S1(en,1) to S5(en,1) for English-to-French MT, and systems S6(ar,1) and S7(ar,1) for Arabic-to-French MT, while the second campaign involved systems S8(en,2) to S12(en,2) for English-to-French and S13(ar,2) for Arabic-to-French. However, for the second campaign, which proceeded in two phases as will be explained below, the systems are designated more specifically as '2a' (before domain adaptation) or '2b' (after adaptation), for example, S8(en,2a) or S13(ar,2b). Note that S13 was the only Arabic-to-French system in the second campaign.

Evaluation Metrics Used for CESTA

CESTA considered the human judgments of fluency and adequacy as the reference for translation quality levels. The "quality" of automated evaluation metrics was therefore assessed with respect to human scores, checking whether automatic scores reproduced the human ones or at least the rankings derived from them. CESTA has thus focused on comparative output quality that was averaged at the document level, rather than, for instance, on error analysis at the sentence level. In many respects the CESTA protocol resembled that used by the National Institute of Standards and Technology (NIST 2003).

Metrics Applied by Human Judges

The two CESTA runs included human evaluations of the quality of the output of MT systems, which represented the most costly measure of the campaign. Two well-known parameters of translation quality were assessed, namely fluency (or intelligibility) and adequacy (or fidelity), following the DARPA 1994 campaign (White et al. 1994). These parameters, together with more detailed alternatives, are given major importance in the FEMTI guidelines for MT evaluation (Hovy et al. 2002).

The human judgments were obtained using a Web-based interface (Hamon et al. 2006) which displays translated segments (generally sentences) to the users in a random order, so that quality judgments are kept as independent as possible between adjacent segments. Each segment was evaluated by two judges. Fluency and adequacy were assessed consecutively in the first run, but separately in the second run, in order to avoid biased correlations between the two aspects. In addition, in the second run, the judges also evaluated the official translation of each segment from the two parallel corpora of each direction (see the description of the data below).

For fluency, the judges were asked the following question (in French): "Is this piece of text written in good French?" Their answers were coded on a five-point scale, with nominal labels in the first run (from 'perfect French' to 'incomprehensible') but only with numerical values in the second run, to ensure uniformity of the scale.

For adequacy, the users were shown a reference translation (see below) and were asked the following question: "Is the meaning of the candidate translation the same as that of the reference translation?" As for fluency, the answers were coded on a 5-point scale.

Automated Metrics

CESTA employed five automated metrics. Three of them, referred to as BLEU, NIST and WNM, are based on a comparison between the candidate translation and one or more reference translations. The other two, the X-Score and D-score, were more experimental, and did not use reference translations, but only data from comparable corpora. However, as their results appeared to be less correlated with human reference judgments, they will not be discussed in this article; a description and commented results appear elsewhere (Hamon and Rajman 2006).

As a brief reminder, the BLEU metric (Papineni et al. 2001) and its NIST variant (Doddington 2002) make use of n-gram based comparisons ($n = 1..5$) between the candidate translation and, typically, up to four reference translations. The more n-grams the candidate segment has in common with the reference segments, the higher the score, and a penalty is introduced for much shorter candidates. The number of unigram matches has been shown to emulate fidelity scores, while higher-level n-gram matches are closer to fluency scores.

The WNM metric (Weighted N-gram Model) (Babych et al. 2004, 2005) refines the n-gram based comparison by weighting the words according to their importance, which is computed using a variant of the *tf.idf* score used in information retrieval. WNM also defines recall, precision, and f-measure; its authors show that recall emulates human judgments of adequacy, while the f-measure best corresponds to human fluency scores.

Although automated metrics have been shown to have some limits (e.g. a score increase does not always reflect an increase in translation quality), their relatively low application cost makes them widely used. The CESTA work on meta-evaluation attempts precisely to answer the question of their reliability, and to assess the limits of their usefulness.

Meta-evaluation of Scores

In order to assess the validity of the evaluation metrics, it is necessary to measure their intrinsic robustness, i.e. their capacity to generate constant scores for a constant quality of MT output, where the similarity of output quality is assessed by human judges. One of the goals of the analysis is thus to estimate the stability of scores when the test data varies while the level of MT quality remains constant. The most appropriate way to test this behaviour, on a limited amount of data, is to sample the data in order to obtain a large quantity of translation samples, and to test whether they obtain homogenous scores when evaluated with one metric.

In the experiments below, a hundred samples were created for each set of translated segments (same data and system), all with the same size as the original set. As the sets are created by random draws *with* replacement, they contain a number of duplicated segments.

The agreement between human judges was measured either as the proportion of identical scores out of the total number of segments; or as the proportion of scores that differ by at most n points ($n = 1, 2, \ldots 4$, on a scale of five). The first score is a particular case of the second one, with $n = 0$.

Finally, to compare the results of automated metrics with the human scores, the Pearson correlation of these values was computed using the scores and also using the ranks. Regarding scores, the Pearson correlation coefficients are given below, while for ranks, their proximity is computed using Hamming's distance.

Organisation of the Two CESTA Runs

The CESTA evaluation campaign was scheduled as two runs, which were conducted at an interval of some 18 months. These two test runs were preceded by a dry run designed to test the integrity of the data distribution and processing systems hosted by the CESTA organizers.

First Run

The goal of the first run was to provide an initial measure of the quality of the MT systems, using the full set of metrics selected for CESTA and texts taken from the general domain. Both translation directions mentioned above – from English and from Arabic into French – were tested (Surcin et al., 2005). Five systems translating from English into French and two systems translating from Arabic into French participated in the first campaign.

Second Run

Apart from the comparative evaluation of system performance, one of the goals of the second run was to improve the robustness and reliability of the evaluation protocol (Hamon et al., 2006). However, the most innovatory objective was to attempt to assess the capacity of the systems to adapt or to be customized to the subject domains of the source texts. For this reason, the improved evaluation protocol was designed to yield two sets of scores, one using a 'generic' or 'default' version of the systems, the other using versions customized for the subject area chosen by the CESTA organizers, i.e. the health domain. Accordingly, two weeks before the evaluation proper, participants in the second run were supplied with a bilingual corpus judged representative of the health domain on which to base their customization.

Five English-to-French systems took part in the second run, of which four had already participated in the first run, and a single Arabic-to-French system. As mentioned above, codenames are different from the first run, from S8 to S13, and indicate the source language and the system version, before and after adaptation (2a or 2b). Although no inter-campaign comparisons are possible for those systems that took part in both, it is of course possible, in the second run, to compare the performance of a system before and after customization to the domain, such as the scores of S9(en,2a) with those of S9(en,2b).

Depending on the architecture of the system, domain adaptation may have entailed the use of a pre-existing dictionary in the health domain, the extraction of terms from the "adaptation" corpus provided by the CESTA organizers, or the training of the system on this corpus. Thus no constraints were imposed on the adaptation procedure, with the ensuing risk that participants might seek out data similar to the corpus provided (which was extracted from Internet sites, as we explain below) and happen to harvest the actual test data. When this actually occurred in one instance, the developers of the system agreed to revert to a version that had not been customized with the test data, and it is these results which are published below. The fact is that evaluating a system trained on the fortuitously harvested test data does not reflect the true capabilities of that system on previously unseen data, since it results in artificially high scores.

An Analysis of the CESTA Data

First Run: General Domain

The corpora used for the first run consisted of a subset of 15 documents from the *Official Journal of the European Community* (JOC) for the English-to-French direction, and a subset of 16 documents from the *Acts of the UNESCO 32nd General Conference* for the Arabic-to-French direction. The documents were segmented into sentences of variable lengths: the average was 25 words per segment for the English source and 78 words by segment for the Arabic source. Each test corpus contained around 20,000 source words and was randomly dispersed within a much larger corpus of around 200,000 words on a similar topic, in order to prevent the participants from knowing directly what were the test data, which would have enabled them to modify their system accordingly to improve performance. The whole data is UTF-8 encoded, following the tagging format defined by NIST (NIST, 2003).

For each document of the test corpora four reference translations were commissioned, to be used in the human evaluation and above all in the automated evaluation. Translations were obtained from two different origins: first, an official translation was available for each data set (as JOC and acts of UNESCO are also translated or directly written in French). In addition to the official French documents, three translations were obtained from professional translation agencies. On the whole, all four translations were used as references by the automated metrics, but only one translation was used as a reference for the human judges.

We observed that the number of words increases slightly in the reference translations with respect to the source sets, while the official translation differed significantly. Indeed, while there were 20,093 words in the English source data, the reference translations were around 22,500 words, but the official reference had some 23,581. Similarly, while there were 23,347 words in the Arabic source data, the reference translations had around 32,500 words and the official translation had 29,971 words.

The whole output of each system was evaluated by humans, for a total of 4,546 segments, i.e. five English-to-French translations of 790 segments and two Arabic-to-French translations of 298 segments. Each segment was evaluated twice by two different judges and 112 judges were recruited; thus each judge evaluated around 81 segments.

Second Run: Adaptation to Health Domain

The English-to-French test corpus for the second run belonged to the health domain and comprised sixteen documents coming from the bilingual web site *Health Canada* (http://www.hc-sc.gc.ca). The Arabic-to-French corpus, also from the health domain, comprised a set of 30 documents from the multilingual web sites of *UNICEF* (http://www.unicef.org), *World Health Organization* (http://www.who.int) and *Family Health International* (http://www.fhi.org). Each of those corpora has an official translation, as for the first run, and each source set was again composed of around 20,000 words. Documents were segmented by sentences, with a mean of 20 words per sentence for the English source and 21 words per sentence for the Arabic source. Again, the test corpora

were dispersed in a much larger masking corpus from a similar domain. In addition to the official translation, three other reference translations were provided by professional agencies. Since the quality of the official translation was found to be lower than those of the three commissioned translations, one of these was selected for the human evaluation, but the four translations were used as references for the automatic evaluation.

In addition to the test corpus, a training or "adaptation" corpus from the same specific domain was provided to the participants for each direction of translation. The goal was to allow the participants to adapt their system before the evaluation and to observe the improvement of systems after an adaptation phase. The training corpora came from the same source as the test corpora and also contained around 20,000 words.

As for the first run, the number of translated words increased in the reference translations, and the official translation differed significantly. While there were exactly 18,880 words in the English source data, the reference translation had all around 23,000 words, and the official reference had 23,411 words. Similarly, while there were 17,305 words in the Arabic source data, the reference translations had around 22,000 words and the official translation had 20,885 words.

For human evaluation, only a third of the segments translated by the systems were evaluated, mainly due to the difficulty of recruiting skilled judges for this domain. Among the 7,150 segments from the translations produced by systems plus one reference translation (i.e. six English-to-French translations of 917 segments and two Arabic-to-French translations of 824 segments), a total of 2,304 segments randomly selected were evaluated twice by 48 judges. Each segment was evaluated by two different judges, which corresponds to an evaluation of 96 segments per judge.

The protocol used for the human evaluation was similar to that of the first run, except that fluency and adequacy criteria were evaluated separately in order to avoid potential correlations of the scores due to their simultaneous evaluation by the same human judge. Furthermore, the human judges were familiar with the health domain (doctors, medical assistants, medical students, etc.), to ensure an objective judgment of the terminological quality of the translations.

In order to check whether the subset was representative of the whole set or not, we computed the BLEU scores for both sets (the entire set and the subset evaluated by human judges) and compared them using a Pearson correlation test: BLEU scores were correlated at 99.86% level and the resulting ranks were correlated at 90% level, which shows that the subset is representative of the whole corpus.

Analysis of the Reference Translations

In analysing the data, we paid special attention to the differences between the references, in order to study the intrinsic quality of the human translations. But another goal was to observe the lexical coverage of the reference translations. Indeed, the main point in producing several reference translations is to cover various ways of translating the same information, so that n-gram based comparison with the references is really informative of translation quality. To assess this desired variability, we computed the BLEU scores for each reference compared to another reference and also to the three other references.

Two main comparison criteria can be considered. First, a low BLEU score shows that the reference is translated differently from another translation, i.e. the same information is translated in two different ways. Then the scores for the overall corpus of references should be heterogeneous with sentences translated differently. In contrast, a too low score could show that the two translations differ significantly and that one of them could contain errors or wrong translations.

Second, a high BLEU score shows that the evaluated reference translation contains sentence structures and translated words that are similar to the other reference translations. This could indicate a good quality of translation, but this is also against our goal of covering different translation possibilities (estimating the source data could be translated in four different ways, which is not obvious, especially for a specific domain).

ARFR-1	**tst1**	**tst2**	**tst3**	**tst4**
ref1	-	20.04	19.08	17.96
ref2	19.98	-	23.32	26.93
ref3	19.00	23.32	-	20.20
ref4	17.91	26.93	20.20	-
3 refs	*35.37*	*43.83*	*38.29*	*41.02*
ENFR-1	**tst1**	**tst2**	**tst3**	**tst4**
ref1	-	29.84	27.45	27.84
ref2	29.81	-	41.67	64.76
ref3	27.43	41.69	-	41.33
ref4	27.82	64.76	41.30	-
3 refs	*40.56*	*75.91*	*55.62*	*74.58*
ARFR-2	**tst1**	**tst2**	**tst3**	**tst4**
ref1	-	21.18	31.27	13.97
ref2	21.21	-	31.53	27.49
ref3	31.31	31.49	-	22.66
ref4	13.95	27.52	22.68	-
3 refs	*41.61*	*49.11*	*54.34*	*38.34*
ENFR-2	**tst1**	**tst2**	**tst3**	**tst4**
ref1	-	33.18	24.01	50.70
ref2	33.17	-	29.99	33.79
ref3	23.97	29.94	-	27.85
ref4	50.69	33.79	27.90	-
3 refs	*60.56*	*50.95*	*43.96*	*64.67*

Table 1. Comparison between reference translations: (*tst1* is the evaluation of reference 1; *ref1* is reference 1 used as a reference translation; and *3 refs* means that the reference is compared with the three other references).

Table 1 shows the results obtained with the two runs and the two translation directions. Scores differ slightly for the two combinations of the same references, due to the penalty given by BLEU to the sentences with a larger number of words.

For the Arabic-to-French direction of the first run, the scores are homogeneous from one reference to another, which means the quality is similar from one translation to another. The scores are also quite low, which means the information is probably translated in four different ways.

For the English-to-French direction of the first run, two translations (the 2nd and the 4th reference) seem to be complementary since their scores are high and similar.

For the Arabic-to-French direction of the second run, scores are quite low, but two translations (3rd and 4th references) are rather different. But regarding the individual scores, it could just mean that the 3rd reference is complementary to the 1st and 2nd references, and that the 4th reference is of a lower quality.

For the English-to-French direction of the second run, scores are higher and references clearly share an identical part of the same information (especially the 1st and 4th references).

To conclude, as a general observation, cross-reference scores are higher for the English-to-French direction than for the Arabic-to-French direction. Furthermore, scores are higher for the second run, which covers a specific domain: the use of a more precise vocabulary decreases the number of different ways of translating the same items of information.

Discussion of the Results: First Run

Table 2 shows the results of the human evaluation of the translations produced by the systems. These translations were judged segment by segment (typically sentences) for fluency and adequacy (the metrics are defined above) on a scale from 1 to 5, where 5 represents an ideal translation. The table presents the scores averaged over all segments.

Several measures have been taken to assess the reliability of these scores. The first observation is that, each segment having been judged by two evaluators, some 40% of segments (all language directions and metrics taken together) receive an identical score from both. When we take into account not only identity of scores but the distance between the scores of the two evaluators for a given segment, we see that the number of judgments which are identical or differ by no more than one point is about 84% for fluency and about 78% for adequacy, the precise figures being similar for both English and Arabic.

The results of the reliability measurements obtained by sampling of the scored segments (using the method described above) show a standard deviation ranging between 0.03 and 0.09 for fluency and adequacy. It is also possible to compute confidence intervals directly from the segment scores for each system, without recourse to sampling. These intervals range between ±0.05 and ±0.08 for all scores and all systems, with a majority of intervals in the range from ±0.06 to ±0.07. The same sampling technique allows us to calculate the probability of the observed rank of each system (also shown in Table 2) on the basis of the average scores. (The use of average ranks for each sample slightly reduces this probability.)

Human evaluation is seen to yield, with sufficient reliability, the same ranking of systems, whether by fluency or by adequacy. However, we observe that it appears impossible to distinguish between the two best systems, S2 and S3, since their confidence intervals at 95% have a non-empty intersection. Similarly, S4 and S5 do not show any significant difference in their adequacy scores, and their fluency scores are close. In the Arabic-to-French direction, S6 is seen to be clearly better than S7. Note that, although the scores for the Arabic-to-French direction are distinctly lower than those obtained for English-to-French, it is difficult to compare them as a whole since the test data and, potentially, the difficulties they present are different. If these scores had been obtained on a single, trilingual corpus, we could have concluded that the systems for Arabic-to-French are less well developed than those for English-into-French. Finally, note that for all systems the adequacy score is higher than the fluency score.

System	Fluency		Adequacy	
	Score (1-5)	Rank	Score (1-5)	Rank
S1(en,1)	2.41±.05	5 (p=1)	2.96±.06	5 (p=1)
S2(en,1)	3.04±.06	1 (p=.99)	3.54±.06	1 (p=1)
S3(en,1)	3.01±.06	2 (p=.99)	3.43±.06	2 (p=1)
S4(en,1)	2.67±.06	4 (p=1)	3.18±.07	4 (p=.89)
S5(en,1)	2.84±.07	3 (p=1)	3.24±.06	3 (p=.89)
S6(ar,1)	1.79±.08	1 (p=1)	2.24±.08	1 (p=1)
S7(ar,1)	1.33±.06	2 (p=2)	1.66±.07	2 (p=2)

Table 2. Results of human judgments for the first campaign: scores on a scale from 1 to 5 (5 is the best and 1 the worst) with confidence intervals and probability rankings.

The scores obtained by use of automated metrics are shown in Table 3. N-grams up to 4 are used by the BLEU and WNM metrics and up to 5 by the NIST metric; both also take word breaks into account. The WNM scores, for which we provide here the f-measure values (indicated as WNMf), are in fact the averages of the WNMf scores obtained by taking each human translation as the reference translation in turn. The table also shows the standard deviation for the scores, calculated by sampling segments in the manner already described.

System	BLEU		NIST		WNMf	
	%	r	value	r	%	R
S1(en,1)	37.60±2.6	5	9.03±.30	5	49.48±.33	5
S2(en,1)	44.76±1.7	3	9.78±.22	3	55.57±.39	3
S3(en,1)	57.33±2.0	1	11.03±.25	1	57.29±.39	1
S4(en,1)	46.64±1.9	2	9.98±.25	2	56.98±.35	2
S5(en,1)	43.87±2.9	4	9.65±.53	4	53.22±.42	4
S6(ar,1)	19.13±1.0	1	7.18±0.15	1	40.93±.28	1
S7(ar,1)	8.21±0.5	2	4.69±0.11	2	33.14±.28	2

Table 3. Scores for the automated metrics from the first run: figures represent percentage (%) or absolute value, standard deviation (±) and ranking (r).

The scores produced by the n-gram metrics are seen to be particularly homogeneous, less in terms of their values than in terms of the *identical* rankings that they give. For the English-to-French direction, the standard deviations, given here instead of confidence intervals, show S4, S2 and S5 to be quite close, while S3 stands apart as having the highest scores and S1 the lowest. For Arabic-to-French, BLEU, NIST and WNMf all reveal a clear difference between S6 and S7, the first being perceptibly better.

Discussion

Comparing the rankings yielded by the human evaluations with those produced by the automated metrics also enables us to gauge the confidence that we can place in the latter. Equally, it is possible to compute Pearson's correlation between the scores, which is shown in Table 4 for both fluency and adequacy. While the correlations are

acceptable, they are weaker than those reported in the literature for the same metrics but with English as the target language.

The ranking produced from the human judgments is (S2>S3)>(S5>S4)>S1, where comparable scores are bracketed together (cf. Table 2). In contrast, the ranking given by the three automated metrics is S3>S4>S2>S5>S1, which differs from the human ranking at several points. Thus, the automated metrics are not completely reliable. However, the results here support the claim of the authors of WNMf that their metric better approximates human judgments than either BLEU or NIST (Babych & Hartley, 2004).

	BLEU	**NIST**	**WNMf**
Fluency	0.69	0.63	0.72
Adequacy	0.69	0.64	0.72

Table 4. Pearson's correlation (scale -1 to 1) between the automated metrics and human judgments, first run, English-to-French direction.

Lastly, to explain the observed difference in ranking between human judges and automated metrics – (S2 > S3) > (S5 > S4) > S1 compared with S3 > S4 > S2 > S5 > S1 – we hypothesize that the n-gram-based automated metrics favour system S4, which in fact uses a language model to select translated segments, and that the performances of S2 and S3 are objectively too close to be distinguished. In every case, the major differences between systems seem to be well captured by the automated metrics, even if their reliability appears to decrease when the target language is French rather than English, in comparison with results published in the literature.

Discussion of the Results: Second Run

We first present the performances recorded by the automated metrics for the 'generic' systems (run 2a). However, only the output of the customized systems was evaluated by human judges also in order to feed into a detailed analysis of the metrics' reliability. Any comparison with the customized systems must, therefore, be made through the scores from the automated metrics, which the figures for run 2b showed to be acceptably robust.

System	**BLEU**		**NIST**		**WNMf**	
	%	**rank**	**value**	**rank**	**%**	**rank**
S8(en,2a)	32.83	4	7.76	4	48.09	4
S9(en,2a)	37.96	1	9.14	1	51.37	1
S10(en,2a)	33.80	3	8.58	3	50.02	2
S11(en,2a)	35.19	2	8.71	2	49.79	3
S12(en,2a)	25.61	5	7.38	5	48.06	5
S13(ar,2a)	36.71	1	8.72	1	54.29	1

Table 5. Scores for the automated metrics for the results of the second run, before adaptation to the domain (2a).

The scores for run 2a before domain adaptation (see Table 5) given by the n-gram-based automated metrics agree once again on a rank ordering, which entitles us to have reasonable confidence in the results. One system, S9(en,2a) stands out as the best, followed by the group S10-S11, then the group S8-S12. Within these two pairs it appears hard to distinguish between systems reliably. For the pair S8-S12, a marked difference is nonetheless signalled by BLEU, but this is not confirmed by WNMf.

The translations generated by the customised systems (run 2b) were subjected to much more detailed evaluation, starting with human judges using an interface which had been improved in the light of the first run. The averages of the scores obtained, with their confidence intervals computed directly on the full set of segments translated by each system, are given in Table 6, together with their probabilities (or frequencies) calculate by sampling. Inter-evaluator agreement increased overall with respect to the first run, insofar as the identical scores assigned for fluency and adequacy stand at, respectively, 43% and 46% of segments, compared with 42% and 38% for the first run. When a difference of one point between judgments is allowed, the similarity reaches some 80% for both attributes, which represents an improved agreement on adequacy relative to the first run, but not on fluency.

System	**Fluency**		**Adequacy**	
	Score (1-5)	**Rank**	**Score (1-5)**	**Rank**
S8(en,2b)	2.28±.10	5 (*p*=1)	2.84±.11	5 (p=1)
S9(en,2b)	3.19±.11	3* (*p*=.51)	3.15±.10	4 (p=1)
S10(en,2b)	3.30±.10	2 (p=.95)	3.44±.11	2 (p=.88)
S11(en,2b)	3.19±.10	3* (p=.51)	3.38±.11	3 (p=.88)
S12(en,2b)	3.57±.09	1 (p=1)	3.78±.09	1 (p=1)
S13(ar,2b)	3.08±.11	1 (p=1)	2.70±.12	1 (p=1)

Table 6. Results of the human judgments for the second run, after adaptation to the domain: scores on a scale from 1 to 5 with their confidence intervals, and ranks with their probabilities (* marks a tied score).

Table 6 shows that the highest scores are achieved by system S12, while the performances of systems S9, S10 and S11 are very close, even indistinguishable, and that system S8 trails far behind. In most cases, the adequacy of the translations is greater than their fluency.

In addition, the human judges evaluated (without being explicitly told) the official translation of each segment (i.e. the translation published on the source website), in order to enable comparison with the system outputs. For the English-to-French direction, the official translation achieved a fluency score of 4.55 and an adequacy score of 4.20, while for the Arabic-to-French direction the values were, respectively, 4.70 and 3.51. Given that the maximum score is 5, we see that fluency scores approach this maximum, but that adequacy is judged much lower, with values that are surprisingly low for a human translation. Several explanations can be advanced, with little to choose between them: harshness of the judges, a too constraining understanding by the judges of the notion of adequacy, low inherent quality of the official translations, or linguistic differences between the French judges and the translations from Quebec (for the English-to-French direction).

The fluency and adequacy scores for the official translations always surpass those of the best systems in Table 6, even if the distance between them is not always great. Unlike the MT system output, the human translations always receive much better scores for

adequacy, which may lend support to the second hypothesis above, namely a too constraining understanding by the judges of the notion of adequacy. This difference may also be due to the higher frequency of reformulations by human translators, whereas MT systems tend to preserve the structure of the source text.

System	BLEU		NIST		WNMf	
	%	r	value	r	%	r
S8(en,2b)	33.04±3.00	2	8.35±0.40	5	50.05±0.66	4
S9(en,2b)	38.07±2.70	4	9.13±0.34	2	51.50±0.71	3
S10(en,2b)	36.60±2.40	5	8.97±0.31	3	52.47±0.68	2
S11(en,2b)	35.74±4.60	3	8.77±0.49	4	50.59±0.66	5
S12(en,2b)	40.43±1.00	1	9.27±0.17	1	56.25±0.77	1
S13(ar,2b)	40.82	1	8.95	1	54.15	1

Table 7. Scores for the automated metrics for the second run (2b), after domain adaptation, in percentage or absolute value, with standard deviation (±) and rank (r).

The scores computed by the automated metrics (see Table 7) reproduce, with some minor exceptions, the ranking given (at much higher cost) by the human judgments for the English-to-French direction. The lead of the first-ranked system S12 is even more commanding, while the scores of the group S9-S10-S11 are more heterogeneous and closer to those of S8. Thus S8 appears to be favoured by the automated n-gram metrics compared with the human judgments. The possibility that this system was optimized for BLEU cannot be ruled out, but this would need to be substantiated by a comprehensive human analysis of the system's output.

The confidence intervals computed by sampling do not generally allow us to distinguish between the four systems S8 to S11 using n-gram-based automated metrics.

For the Arabic-to-English direction, the scores of the sole participating system appear to be in the same range as those of the English-to-French systems, but this comparison cannot be very precise, since the reference data are different.

When the scores given by the human judges for system S13 are compared with the scores for the human reference translation – fluency 3.08 *vs* 4.70 and adequacy 2.70 *vs* 3.51 – we can see that the system better approximates human performance for adequacy, i.e. its ability to reproduce the content of the source sentences, than it does for fluency. This may also reflect a bias of human judges, who are less tolerant of errors of French.

Meta-evaluation

The correlations between the automated metrics and the human judges improve noticeably in the second run (2b) for English-to-French, as a comparison of Table 4 with Table 8 shows. NIST and WNMf tie in achieving the best correlation – 0.87/0.86 for fluency and 0.95 for adequacy – thus confirming the experience of the authors of the latter metric (Babych & Hartley, 2004).

Although these figures are slightly lower than those achieved with English as the target language, they nonetheless confirm that such automated metrics can be substituted when necessary for much more costly human metrics. The increased correlation is probably explained by an improved protocol for collecting human judgments, as well as by the use of better quality – that is, more homogeneous – reference translations. The nature of the source texts used, in particular their less wide-ranging domain coverage, may also explain the increased stability of the human quality judgments.

	BLEU	NIST	WNMf
Fluency	0.85	0.87	0.86
Adequacy	0.94	0.95	0.95

Table 8. Pearson's correlation (scale -1 to 1) between the automated metrics and the human judges, second run (2b), English-to-French direction.

The second campaign also aimed to assess the capacity of the systems for rapid adaptation to a specific domain, in this case health. Although investigated by (Babych et al., 2004), this remains an often neglected aspect of MT evaluation. Table 9 reproduces for ease of comparison the automated scores generated before (2a) and after (2b) domain adaptation. This comparison illustrates the difficulties of appropriate customization to a specialized domain. While most scores improve considerably after customization, for systems S9, S10 and S11 the improvement remains slight. This could be due to an already high initial quality level, or to difficulties in adapting the system stemming from the small size of the customization data set and the relatively short period of time allowed for customization.

On the other hand, systems S11 and, to a lesser extent, S8 adapt very effectively to the domain. The case of S11, which rises from last to first position, offers the particularly edifying prospect of this system achieving an excellent performance provided the application domain is properly characterized at the outset.

Sys.	BLEU (%)		NIST		WNMf (%)	
	before	after	before	after	before	after
S8	32.83	33.04	7.76	8.35	48.09	50.05
S9	37.96	38.07	9.14	9.13	51.37	51.50
S10	33.80	36.60	8.58	8.97	50.02	52.47
S11	35.19	35.74	8.71	8.77	49.79	50.59
S12	25.61	40.43	7.38	9.27	48.06	56.25

Table 9. Comparison of scores for automated metrics in the second run, before an after domain adaptation, English/French direction.

The differences between the reference data for the first and second runs means that the results are not directly comparable, all the more so because the participating systems were not exactly the same. Consequently, general lessons can be drawn only at the level of a meta-evaluation of the metrics, that is, via reliability measures which enable us to better understand the systems' qualities and limitations for future use. Although widely used in the community, the BLEU and NIST metrics have not entirely lived up to expectations. Correlations and distance measures relative to human judgments appear broadly acceptable, yet less good than anticipated.

Results vary somewhat depending on the size of the n-grams considered, as is documented in the full CESTA report[3]. Thus, NIST achieves better correlations than

[3] www.technolangue.net/IMG/pdf/Rapport_final_CESTA_v1.04.pdf

BLEU with unigrams and bigrams, while the position is reversed for trigrams and quadrigrams.
Finally, WNMf achieves better correlations than either BLEU or NIST, justifying the search for new metrics that improve on BLEU. However, WNMf behaves differently depending on which reference translation is used; for example, its correlation with human judgments goes from around 40% to around 80% depending whether the reference translation is, respectively, the official translation or one of the translations produced by the agencies (observation derived from the first run). Using several reference translations, if available, and averaging the scores seems a viable compromise, and was the approach taken within CESTA.
In addition, the mWER and mPER metrics were tested, yielding correlations with human judgments comparable to those of BLEU and NIST.

Evaluation Package

The evaluation protocols created by CESTA – notably, data and metrics – are available to both MT developers and users. These resources are, to our knowledge, a first for the English-to-French and English-to-Arabic directions. The package is distributed on CD-ROM by ELDA and includes all documentation and reports.

Conclusion

The two runs of the CESTA campaign have been hugely informative, as the volume of numerical data in the final report demonstrates. An evaluation protocol has been put in place, both to conduct automated evaluations and to attempt to confirm their validity by human evaluations, which remain the reference for assessing the ability to translate. The human evaluations proved costly, in time and in money, so confirming the usefulness of automated evaluation, which is less expensive in both respects.
It is nevertheless important to set the automated results in context and to acknowledge their limitations before aspiring to replace human evaluation altogether. CESTA has shown that the relative subjectivity of human evaluation can be mitigated by increasing the number of judgments for a given segment. Across the two runs, the human evaluators were seen to produce consistent judgments, strict agreement on scores (on a five-point scale) being just under 50% and the similarity between judgments being far higher.
As for the MT systems, two distinguish themselves in the first run, while one lags well behind. In the second run, a hierarchy again emerges, with one system clearly ahead of the field and another system trailing. Three systems derive benefit from the domain data, recording significant progress after customization.
An evaluation package including the data produced and used for the CESTA evaluation campaign is available from ELDA[4].

Acknowledgments

We express our warm thanks to the developers who adapted the participating systems to the CESTA specifications, and to the representatives who attended CESTA working meetings.

[4] http://catalog.elra.info

References

Babych B., Hartley A. and Elliott D. (2005). Estimating the predictive power of n-gram evaluation metrics across languages and text types. *Proc. of MT Summit X*, Phuket, Thailand, pp. 412-418.

Babych B., Elliott D., Hartley A. (2004). Extending MT evaluation tools with translation complexity metrics. In *Proc. of the Coling 2004 (20th International Conference on Computational Linguistics)*, Geneva, Switzerland, August 2004. pp. 106-112.

Babych B. and Hartley T. (2004). Extending the BLEU MT Evaluation Method with Frequency Weightings. *Proc. of ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics: ACL 2004)*, Barcelona, Spain, pp. 621-628.

Dabbadie M., Mustafa El Hadi W. and Timimi I. (2004). CESTA: The European MT Evaluation Campaign. *Multilingual Computing and Technology*, vol. 15, n° 5, pp. 10-12.

Doddington G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *Proc. of HLT 2002 (2nd Conference on Human Language Technology)*, San Diego, CA, pp. 128-132.

Hamon O. et al. (2006). CESTA: First Conclusions of the Technolangue MT Evaluation Campaign. *Proc. of LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Genova, Italy, pp. 179-184.

Hamon O. and Rajman M. (2006). X-Score: Automatic Evaluation of Machine Translation Grammaticality. *Proc. of LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Genova, Italy.

Hamon O., Popescu-Belis A., Hartley A., Mustafa El Hadi W. and Rajman M. (2007) - CESTA: Campagne d'Evaluation des Systèmes de Traduction Automatique. In Chaudiron S. et al., eds., *Bilan de l'action Technolangue (2003-2006)*, Hermès, Paris, 24 p.

Hovy E. H., King M. and Popescu-Belis A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, vol. 17 (1), pp. 1-33.

Mustafa El Hadi W. et al. (2004). CESTA: Machine Translation Evaluation Campaign. *Proc. of Coling 2004 Workshop on Language Resources for Translation Work, Research and Training*, Geneva, pp. 8-17.

Papineni K. et al. (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).

Surcin S. et al. (2005). Evaluation of Machine Translation with Predictive Metrics beyond BLEU/NIST: CESTA Evaluation Campaign #1. *Proceedings of Machine Translation Summit X*, Phuket, Thailand, pp. 117-124.

White J. S. (2001). Predicting Intelligibility from Fidelity in MT Evaluation. *Proceedings of Workshop on MT Evaluation "Who did what to whom?" at Mt Summit VIII*, Santiago de Compostela, Spain.

White J. S. and O'Connell T. A. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of AMTA Conference, 5-8 October 1994*, Columbia, MD, USA.

Zhang Y., Vogel S. and Waibel A. (2004). Interpreting BLEU/NIST Scores: How Much Improvement do We Need to Have a Better System? *Proceedings of LREC 2004* Lisbon, Portugal, vol. VI/VI, pp. 2051-2054.