Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations

Jesús Giménez and **Lluís Màrquez**
TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1–3, E-08034, Barcelona
{jgimenez,lluism}@lsi.upc.edu

Abstract

Combining different metrics into a single measure of quality seems the most direct and natural way to improve over the quality of individual metrics. Recently, several approaches have been suggested (Kulesza and Shieber, 2004; Liu and Gildea, 2007; Albrecht and Hwa, 2007a). Although based on different assumptions, these approaches share the common characteristic of being *parametric*. Their models involve a number of parameters whose weight must be adjusted. As an alternative, in this work, we study the behaviour of *non-parametric* schemes, in which metrics are combined without having to adjust their relative importance. Besides, rather than limiting to the lexical dimension, we work on a wide set of metrics operating at different linguistic levels (e.g., lexical, syntactic and semantic). Experimental results show that non-parametric methods are a valid means of putting different quality dimensions together, thus tracing a possible path towards *heterogeneous* automatic MT evaluation.

1 Introduction

Automatic evaluation metrics have notably accelerated the development cycle of MT systems in the last decade. There exist a large number of metrics based on different similarity criteria. By far, the most widely used metric in recent literature is BLEU (Papineni et al., 2001). Other well-known metrics are WER (Nießen et al., 2000), NIST (Doddington, 2002), GTM (Melamed et al., 2003), ROUGE (Lin and Och, 2004a), METEOR (Banerjee and Lavie, 2005), and TER (Snover et al., 2006), just to name a few. All these metrics take into account information at the lexical level[1], and, therefore, their reliability depends very strongly on the heterogeneity/representativity of the set of reference translations available (Culy and Riehemann, 2003). In order to overcome this limitation several authors have suggested taking advantage of paraphrasing support (Zhou et al., 2006; Kauchak and Barzilay, 2006; Owczarzak et al., 2006). Other authors have tried to exploit information at deeper linguistic levels. For instance, we may find metrics based on full constituent parsing (Liu and Gildea, 2005), and on dependency parsing (Liu and Gildea, 2005; Amigó et al., 2006; Mehay and Brew, 2007; Owczarzak et al., 2007). We may find also metrics at the level of shallow-semantics, e.g., over semantic roles and named entities (Giménez and Màrquez, 2007), and at the properly semantic level, e.g., over discourse representations (Giménez, 2007).

However, none of current metrics provides, in isolation, a *global* measure of quality. Indeed, all metrics focus on *partial* aspects of quality. The main problem of relying on partial metrics is that we may obtain *biased* evaluations, which may lead us to derive *inaccurate* conclusions. For instance, Callison-Burch et al. (2006) and Koehn and Monz (2006) have recently reported several problematic cases related to the automatic evaluation of systems oriented towards maximizing different quality aspects. Corroborating the findings by Culy and Riehemann (2003), they showed that BLEU overrates SMT systems with respect to other types of systems, such

[1] ROUGE and METEOR may consider morphological variations. METEOR may also look up for synonyms in WordNet.

as rule-based, or human-aided. The reason is that SMT systems are likelier to match the sublanguage (e.g., lexical choice and order) represented by the set of reference translations. We argue that, in order to perform more *robust*, i.e., less biased, automatic MT evaluations, different quality dimensions should be jointly taken into account.

A natural solution to this challenge consists in combining the scores conferred by different metrics, ideally covering a *heterogeneous* set of quality aspects. In the last few years, several approaches to metric combination have been suggested (Kulesza and Shieber, 2004; Liu and Gildea, 2007; Albrecht and Hwa, 2007a). In spite of working on a limited set of quality aspects, mostly lexical features, these approaches have provided effective means of combining different metrics into a single measure of quality. All these methods implement a *parametric* combination scheme. Their models involve a number of parameters whose weight must be adjusted (see further details in Section 2).

As an alternative path towards heterogeneous MT evaluation, in this work, we explore the possibility of relying on *non-parametric* combination schemes, in which metrics are combined without having to adjust their relative importance (see Section 3). We have studied their ability to integrate a wide set of metrics operating at different linguistic levels (e.g., lexical, syntactic and semantic) over several evaluation scenarios (see Section 4). We show that non-parametric schemes offer a valid means of putting different quality dimensions together, effectively yielding a significantly improved evaluation quality, both in terms of human likeness and human acceptability. We have also verified that these methods port well across test beds.

2 Related Work

Approaches to metric combination require two important ingredients:

Combination Scheme, i.e., how to combine several metric scores into a single score. As pointed out in Section 1, we distinguish between parametric and non-parametric schemes.

Meta-Evaluation Criterion, i.e., how to evaluate the quality of a metric combination. The two most prominent meta-evaluation criteria are:

- *Human Acceptability:* Metrics are evaluated in terms of their ability to capture the degree of acceptability to humans of automatic translations, i.e., their ability to emulate human assessors. The underlying assumption is that *'good'* translations should be acceptable to human evaluators. Human acceptability is usually measured on the basis of *correlation* between automatic metric scores and human assessments of translation quality[2].
- *Human Likeness:* Metrics are evaluated in terms of their ability to capture the features which distinguish human from automatic translations. The underlying assumption is that *'good'* translations should resemble human translations. Human likeness is usually measured on the basis of *discriminative power* (Lin and Och, 2004b; Amigó et al., 2005).

In the following, we describe the most relevant approaches to metric combination suggested in recent literature. All are parametric, and most of them are based on machine learning techniques. We distinguish between approaches relying on human likeness and approaches relying on human acceptability.

2.1 Approaches based on Human Likeness

The first approach to metric combination based on human likeness was that by Corston-Oliver et al. (2001) who used decision trees to distinguish between human-generated ('good') and machine-generated ('bad') translations. They focused on evaluating only the well-formedness of automatic translations (i.e., subaspects of fluency), obtaining high levels of classification accuracy.

Kulesza and Shieber (2004) extended the approach by Corston-Oliver et al. (2001) to take into account other aspects of quality further than fluency alone. Instead of decision trees, they trained Support Vector Machine (SVM) classifiers. They used features inspired by well-known metrics such as BLEU, NIST, WER, and PER. Metric quality was evaluated both in terms of classification accuracy and correlation with human assessments at the sentence level.

[2] Usually adequacy, fluency, or a combination of the two.

A significant improvement with respect to standard individual metrics was reported.

Gamon et al. (2005) presented a similar approach which, in addition, had the interesting property that the set of human and automatic translations could be independent, i.e., human translations were not required to correspond, as references, to the set of automatic translations.

2.2 Approaches based on Human Acceptability

Quirk (2004) applied supervised machine learning algorithms (e.g., perceptrons, SVMs, decision trees, and linear regression) to approximate human quality judgements instead of distinguishing between human and automatic translations. Similarly to the work by Gamon et al. (2005) their approach does not require human references.

More recently, Albrecht and Hwa (2007a; 2007b) re-examined the SVM classification approach by Kulesza and Shieber (2004) and, inspired by the work of Quirk (2004), suggested a regression-based learning approach to metric combination, with and without human references. The regression model learns a continuous function that approximates human assessments in training examples.

As an alternative to methods based on machine learning techniques, Liu and Gildea (2007) suggested a simpler approach based on linear combinations of metrics. They followed a *Maximum Correlation Training*, i.e., the weight for the contribution of each metric to the overall score was adjusted so as to maximize the level of correlation with human assessments at the sentence level.

As expected, all approaches based on human acceptability have been shown to outperform that of Kulesza and Shieber (2004) in terms of human acceptability. However, no results in terms of human likeness have been provided, thus leaving these comparative studies incomplete.

3 Non-Parametric Combination Schemes

In this section, we provide a brief description of the QARLA framework (Amigó et al., 2005), which is, to our knowledge, the only existing non-parametric approach to metric combination. QARLA is non-parametric because, rather than assigning a weight to the contribution of each metric, the evaluation of a given automatic output a is addressed through a set of independent probabilistic tests (one per metric) in which the goal is to falsify the hypothesis that a is a human reference. The input for QARLA is a set of test cases A (i.e., automatic translations), a set of similarity metrics X, and a set of models R (i.e., human references) for each test case. With such a testbed, QARLA provides the two essential ingredients required for metric combination:

Combination Scheme Metrics are combined inside the QUEEN measure. QUEEN operates under the *unanimity* principle, i.e., the assumption that a 'good' translation must be similar to all human references according to all metrics. $\text{QUEEN}_X(a)$ is defined as the probability, over $R \times R \times R$, that, for every metric in X, the automatic translation a is more similar to a human reference r than two other references, r' and r'', to each other. Formally:

$$\text{QUEEN}_{X,R}(a) = Prob(\forall x \in X : x(a,r) \geq x(r',r''))$$

where $x(a,r)$ stands for the similarity between a and r according to the metric x. Thus, QUEEN allows us to combine different similarity metrics into a single measure, without having to adjust their relative importance. Besides, QUEEN offers two other important advantages which make it really suitable for metric combination: (i) it is *robust* against metric redundancy, i.e., metrics covering similar aspects of quality, and (ii) it is not affected by the scale properties of metrics. The main drawback of the QUEEN measure is that it requires at least three human references, when in most cases only a single reference translation is available.

Meta-evaluation Criterion Metric quality is evaluated using the **KING** measure of human likeness. All human references are assumed to be equally optimal and, while they are likely to be different, the best similarity metric is the one that identifies and uses the features that are common to all human references, grouping them and separating them from automatic translations. Based on QUEEN, KING represents the probability that a human reference

does not receive a lower score than the score attained by *any* automatic translation. Formally:

$$\text{KING}_{A,R}(X) = Prob(\forall a \in A : \text{QUEEN}_{X,R-\{r\}}(r) \geq \text{QUEEN}_{X,R-\{r\}}(a))$$

KING operates, therefore, on the basis of discriminative power. The closest measure to KING is ORANGE (Lin and Och, 2004b), which is, however, not intended for the purpose of metric combination.

Apart from being non-parametric, QARLA exhibits another important feature which differentiates it form other approaches; besides considering the similarity between automatic translations and human references, QARLA also takes into account the distribution of similarities among human references.

However, QARLA is not well suited to port from human likeness to human acceptability. The reason is that QUEEN is, by definition, a very restrictive measure —a 'good' translation must be similar to *all* human references according to *all* metrics. Thus, as the number of metrics increases, it becomes easier to find a metric which does not satisfy the QUEEN assumption. This causes QUEEN values to get close to zero, which turns correlation with human assessments into an impractical meta-evaluation measure.

We have *simulated* a non-parametric scheme based on human acceptability by working on uniformly averaged linear combinations (ULC) of metrics. Our approach is similar to that of Liu and Gildea (2007) except that in our case all the metrics in the combination are equally important[3]. In other words, ULC is indeed a particular case of a parametric scheme, in which the contribution of each metric is not adjusted. Formally:

$$\text{ULC}_X(a, R) = \frac{1}{|X|} \sum_{x \in X} x(a, R)$$

where X is the metric set, and $x(a, R)$ is the similarity between the automatic translation a and the set of references R, for the given test case, according to the metric x. Since correlation with human assessments at the system level is vaguely informative (it is often estimated on very few system samples), we evaluate metric quality in terms of correlation with human assessments at the sentence level (R_{snt}). We use the sum of adequacy and fluency to simulate a global assessment of quality.

[3]That would be assuming that all metrics operate in the same range of values, which is not always the case.

	AE$_{04}$	**CE$_{04}$**	**AE$_{05}$**	**CE$_{05}$**
#human references	5	5	5	4
#system outputs	5	10	7	10
#outputs$_{\text{assessed}}$	5	10	6	5
#sentences	1,353	1,788	1,056	1,082
#sentences$_{\text{assessed}}$	347	447	266	272

Table 1: Description of the test beds

4 Experimental Work

In this section, we study the behavior of the two combination schemes presented in Section 3 in the context of four different evaluation scenarios.

4.1 Experimental Settings

We use the test beds from the 2004 and 2005 NIST MT Evaluation Campaigns (Le and Przybocki, 2005)[4]. Both campaigns include two different translations exercises: Arabic-to-English ('AE') and Chinese-to-English ('CE'). Human assessments of adequacy and fluency are available for a subset of sentences, each evaluated by two different human judges. See, in Table 1, a brief numerical description including the number of human references and system outputs available, as well as the number of sentences per output, and the number of system outputs and sentences per system assessed.

For metric computation, we have used the IQ$_{\text{MT}}$ v2.1, which includes metrics at different linguistic levels (lexical, shallow-syntactic, syntactic, shallow-semantic, and semantic). A detailed description may be found in (Giménez, 2007)[5].

4.2 Evaluating Individual Metrics

Prior to studying the effects of metric combination, we study the isolated behaviour of individual metrics. We have selected a set of metric representatives from each linguistic level. Table 2 shows meta-evaluation results for the test beds described in Section 4.1, according both to human likeness (KING)

[4]`http://www.nist.gov/speech/tests/summaries/2005/mt05.htm`

[5]The IQ$_{\text{MT}}$ Framework may be freely downloaded from `http://www.lsi.upc.edu/~nlp/IQMT`.

Level	Metric	KING				R_{snt}			
		AE_{04}	**CE_{04}**	**AE_{05}**	**CE_{05}**	**AE_{04}**	**CE_{04}**	**AE_{05}**	**CE_{05}**
Lexical	1-WER	0.70	0.51	0.48	0.61	0.53	0.47	0.38	0.47
	1-PER	0.64	0.43	0.45	0.58	0.50	0.51	0.29	0.40
	1-TER	0.73	0.54	0.53	0.66	0.54	0.50	0.38	0.49
	BLEU	0.70	0.49	0.52	0.59	0.50	0.46	0.36	0.39
	NIST	**0.74**	0.53	**0.55**	**0.68**	0.53	**0.55**	0.37	0.46
	GTM.e1	0.67	0.49	0.48	0.61	0.41	0.50	0.26	0.29
	GTM.e2	0.69	0.52	0.51	0.64	0.49	0.54	0.43	0.48
	$ROUGE_L$	0.73	**0.59**	0.49	0.65	**0.58**	**0.60**	0.41	**0.52**
	$ROUGE_W$	**0.75**	**0.62**	**0.54**	**0.68**	**0.59**	**0.57**	**0.48**	**0.54**
	$METEOR_{wnsyn}$	**0.75**	0.56	**0.57**	**0.69**	**0.56**	**0.56**	0.35	0.41
Shallow Syntactic	SP-O_p-*	0.66	0.48	0.49	0.59	0.51	**0.57**	0.38	0.41
	SP-O_c-*	0.65	0.44	0.46	0.59	**0.55**	**0.58**	0.42	0.41
	SP-$NIST_l$	0.73	0.51	**0.55**	**0.66**	0.53	0.54	0.38	0.44
	SP-$NIST_p$	**0.79**	**0.60**	**0.56**	**0.70**	0.46	0.49	0.37	0.39
	SP-$NIST_{iob}$	0.69	0.48	0.49	0.59	0.32	0.36	0.27	0.26
	SP-$NIST_c$	0.60	0.42	0.39	0.52	0.26	0.27	0.16	0.16
Syntactic	DP-HWC_w	0.58	0.40	0.42	0.53	0.41	0.08	0.35	0.40
	DP-HWC_c	0.50	0.32	0.33	0.41	0.41	0.17	0.38	0.32
	DP-HWC_r	0.56	0.40	0.37	0.46	0.42	0.16	0.39	0.43
	DP-O_l-*	0.58	0.48	0.41	0.52	0.52	0.48	0.36	0.37
	DP-O_c-*	0.65	0.45	0.44	0.55	0.49	0.51	0.43	0.41
	DP-O_r-*	0.71	**0.57**	**0.54**	0.64	**0.55**	**0.55**	**0.50**	**0.50**
	CP-O_p-*	0.67	0.47	0.47	0.60	0.53	**0.57**	0.38	0.46
	CP-O_c-*	0.66	0.51	0.49	0.62	**0.57**	**0.59**	**0.45**	0.50
	CP-STM	0.64	0.42	0.43	0.58	0.39	0.13	0.34	0.30
Shallow Semantic	NE-O_e-**	0.65	0.45	0.46	0.57	0.47	0.56	0.32	0.39
	SR-O_r-*	0.48	0.22	0.34	0.41	0.28	0.10	0.32	0.21
	SR-O_{rv}	0.36	0.13	0.24	0.27	0.27	0.12	0.25	0.24
Semantic	DR-O_r-*	0.62	0.47	0.50	0.55	0.47	0.46	0.43	0.37
	DR-O_{rp}-*	0.58	0.42	0.43	0.50	0.37	0.35	0.36	0.26
Optimal Combination		**0.79**	**0.64**	**0.61**	**0.70**	**0.64**	**0.63**	**0.54**	**0.61**

Table 2: Metric Meta-evaluation

and human acceptability (R_{snt}), computed over the subsets of sentences for which human assessments are available.

The first observation is that the two meta-evaluation criteria provide very similar metric quality rankings for a same test bed. This seems to indicate that there is a relationship between the two meta-evaluation criteria employed. We have confirmed this intuition by computing the Pearson correlation coefficient between values in columns 1 to 4 and their counterparts in columns 5 to 8. There exists a high correlation (R = 0.79).

A second observation is that metric quality varies significantly from task to task. This is due to the significant differences among the test beds employed. These are related to three main aspects: language pair, translation domain, and system typology. For instance, notice that most metrics exhibit a lower quality in the case of the 'AE_{05}' test bed. The reason is that, while in the rest of test beds all systems are statistical, the 'AE_{05}' test bed presents the particularity of providing automatic translations produced by *heterogeneous* MT systems (i.e., systems belonging to different paradigms)[6]. The fact that most systems are statistical also explains why, in general, lexical metrics exhibit a higher quality. However, highest levels of quality are not in all cases attained by metrics at the lexical level (see highlighted values). In fact, there is only one metric, '$ROUGE_W$' (based on lexical matching), which is consistently among the top-scoring in all test beds according to both meta-evaluation criteria. The underlying cause is simple: current metrics do not provide a global measure of quality, but account only for partial aspects of it. Apart from evincing the importance of the meta-evaluation process, these results strongly suggest the need for conducting heterogeneous MT evaluations.

[6] Specifically, all systems are statistical except one which is human-aided.

$Opt.K(AE.04)$	$=\{$*SP-NIST$_p$*$\}$
$Opt.K(CE.04)$	$=\{$*ROUGE$_W$*, *SP-NIST$_p$*, *ROUGE$_L$*$\}$
$Opt.K(AE.05)$	$=\{$*METEOR$_{wnsyn}$*, *SP-NIST$_p$*, *DP-O$_r$*-*$\}$
$Opt.K(CE.05)$	$=\{$*SP-NIST$_p$*$\}$
$Opt.R(AE.04)$	$=\{$*ROUGE$_W$*, *ROUGE$_L$*, *CP-O$_c$*-*, *METEOR$_{wnsyn}$*, *DP-O$_r$*-*, *DP-O$_l$*-*, *GTM.e2*, *DR-O$_r$*-*, *CP-STM*$\}$
$Opt.R(CE.04)$	$=\{$*ROUGE$_L$*, *CP-O$_c$*-*, *ROUGE$_W$*, *SP-O$_p$*-*, *METEOR$_{wnsyn}$*, *DP-O$_r$*-*, *GTM.e2*, *1-WER*, *DR-O$_r$*-*$\}$
$Opt.R(AE.05)$	$=\{$*DP-O$_r$*-*, *ROUGE$_W$*$\}$
$Opt.R(CE.05)$	$=\{$*ROUGE$_W$*, *ROUGE$_L$*, *DP-O$_r$*-*, *CP-O$_c$*-*, *1-TER*, *GTM.e2*, *DP-HWC$_r$*, *CP-STM*$\}$

Table 3: Optimal metric sets

4.3 Finding Optimal Metric Combinations

In that respect, we study the applicability of the two combination strategies presented. Optimal metric sets are determined by maximizing over the corresponding meta-evaluation measure (KING or R_{snt}). However, because exploring all possible combinations was not viable, we have used a simple algorithm which performs an *approximate* search. First, individual metrics are ranked according to their quality. Then, following that order, metrics are added to the optimal set only if in doing so the global quality increases. Since no training is required it has not been necessary to keep a held-out portion of the data for test (see Section 4.4 for further discussion).

Optimal metric sets are displayed in Table 3. Inside each set, metrics are sorted in decreasing quality order. The *'Optimal Combination'* line in Table 2 shows the quality attained by these sets, combined under QUEEN in the case of KING optimization, and under ULC in the case of optimizing over R_{snt}. In most cases optimal sets consist of metrics operating at different linguistic levels, mostly at the lexical and syntactic levels. This is coherent with the findings in Section 4.2. Metrics at the semantic level are selected only in two cases, corresponding to the R_{snt} optimization in 'AE$_{04}$' and 'CE$_{04}$' test beds. Also in two cases, corresponding to the KING optimization in 'AE$_{04}$' and 'CE$_{05}$' test beds, it has not been possible to find any metric combination which outperforms the best individual metric. This is not a discouraging result. After all, in these cases, the best metric alone achieves already a very high quality (0.79 and 0.70, respectively). The fact that a single feature suffices to discern between manual and automatic translations indicates that MT systems are easily distinguishable, possibly because of their low quality and/or because they are all based on the same translation paradigm.

4.4 Portability

It can be argued that metric set optimization is itself a training process; each metric would have an associated binary parameter controlling whether it is selected or not. For that reason, in Table 4, we have analyzed the portability of optimal metric sets (i) across test beds and (ii) across combination strategies. As to portability across test beds (i.e., across language pairs and years), the reader must focus on the cells for which the meta-evaluation criterion guiding the metric set optimization matches the criterion used in the evaluation, i.e., the top-left and bottom-right 16-cell quadrangles. The fact that the 4 values in each subcolumn are in a very similar range confirms that optimal metric sets port well across test beds. We have also studied the portability of optimal metric sets across combination strategies. In other words, although QUEEN and ULC are thought to operate on metric combinations respectively optimized on the basis of human likeness and human acceptability, we have studied the effects of applying either measure over metric combinations optimized on the basis of the alternative meta-evaluation criterion. In this case, the reader must compare top-left vs. bottom-left (KING) and top-right vs. bottom-right (R_{snt}) 16-cell quadrangles. It can be clearly seen that optimal metric sets, in general, do not port well across meta-evaluation criteria, particularly from human likeness to human acceptability. However, interestingly, in the case of 'AE$_{05}$' (i.e., heterogeneous systems), the optimal metric set ports well from human acceptability to human likeness. We speculate that system heterogeneity has contributed positively for the sake of robustness.

5 Conclusions

As an alternative to current parametric combination techniques, we have presented two different meth-

Metric Set	KING				R_{snt}			
	$\mathbf{AE_{04}}$	$\mathbf{CE_{04}}$	$\mathbf{AE_{05}}$	$\mathbf{CE_{05}}$	$\mathbf{AE_{04}}$	$\mathbf{CE_{04}}$	$\mathbf{AE_{05}}$	$\mathbf{CE_{05}}$
$\boldsymbol{Opt.K(AE.04)}$	0.79	0.60	0.56	0.70	0.46	0.49	0.37	0.39
$\boldsymbol{Opt.K(CE.04)}$	0.78	0.64	0.57	0.67	0.49	0.51	0.39	0.43
$\boldsymbol{Opt.K(AE.05)}$	0.74	0.63	0.61	0.66	0.48	0.51	0.39	0.42
$\boldsymbol{Opt.K(CE.05)}$	0.79	0.60	0.56	0.70	0.46	0.49	0.37	0.39
$\boldsymbol{Opt.R(AE.04)}$	0.62	0.56	0.52	0.49	0.64	0.61	0.53	0.58
$\boldsymbol{Opt.R(CE.04)}$	0.68	0.59	0.55	0.56	0.63	0.63	0.51	0.57
$\boldsymbol{Opt.R(AE.05)}$	*0.75*	*0.64*	*0.59*	*0.69*	0.62	0.60	0.54	0.57
$\boldsymbol{Opt.R(CE.05)}$	0.64	0.56	0.51	0.52	0.63	0.57	0.53	0.61

Table 4: Portability of combination strategies

ods: a genuine non-parametric method based on human likeness, and a parametric method based human acceptability in which the parameter weights are set equiprobable. We have shown that both strategies may yield a significantly improved quality by combining metrics at different linguistic levels. Besides, we have shown that these methods generalize well across test beds. Thus, a valid path towards heterogeneous automatic MT evaluation has been traced. We strongly believe that future MT evaluation campaigns should benefit from these results specially for the purpose of comparing systems based on different paradigms. These techniques could also be used to build better MT systems by allowing system developers to perform more accurate error analyses and less biased adjustments of system parameters.

As an additional result, we have found that there is a tight relationship between human acceptability and human likeness. This result, coherent with the findings by Amigó et al. (2006), suggests that the two criteria are interchangeable. This would be a point in favour of combination schemes based on human likeness, since human assessments —which are expensive to acquire, subjective and not reusable— are not required. We also interpret this result as an indication that human assessors probably behave in many cases in a discriminative manner. For each test case, assessors would inspect the source sentence and the set of human references trying to identify the features which 'good' translations should comply with, for instance regarding adequacy and fluency. Then, they would evaluate automatic translations roughly according to the number and relevance of the features they share and the ones they do not.

For future work, we plan to study the integration of finer features as well as to conduct a rigorous comparison between parametric and non-parametric combination schemes. This may involve reproducing the works by Kulesza and Shieber (2004) and Albrecht and Hwa (2007a). This would also allow us to evaluate their approaches in terms of both human likeness and human acceptability, and not only on the latter criterion as they have been evaluated so far.

Acknowledgements

This research has been funded by the Spanish Ministry of Education and Science, project OpenMT (TIN2006-15307-C03-02). Our NLP group has been recognized as a Quality Research Group (2005 SGR-00130) by DURSI, the Research Department of the Catalan Government. We are thankful to Enrique Amigó, for his generous help and valuable comments. We are also grateful to the NIST MT Evaluation Campaign organizers, and participants who agreed to share their system outputs and human assessments for the purpose of this research.

References

Joshua Albrecht and Rebecca Hwa. 2007a. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. In *Proceedings of ACL*, pages 880–887.

Joshua Albrecht and Rebecca Hwa. 2007b. Regression for Sentence-Level MT Evaluation with Pseudo References. In *Proceedings of ACL*, pages 296–303.

Enrique Amigó, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo. 2005. QARLA: a Framework for the Evaluation of Automatic Sumarization. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics*.

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. 2006. MT Evaluation: Human-Like vs. Human Acceptable. In *Proceedings of COLING-ACL06*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of EACL*.

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, pages 140–147.

Christopher Culy and Susanne Z. Riehemann. 2003. The Limits of N-gram Translation Evaluation Metrics. In *Proceedings of MT-SUMMIT IX*, pages 1–8.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd IHLT*.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-Level MT evaluation without reference translations: beyond language modeling. In *Proceedings of EAMT*.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proceedings of the ACL Workshop on Statistical Machine Translation*.

Jesús Giménez. 2007. IQMT v 2.1. Technical Manual. Technical report, TALP Research Center. LSI Department. http://www.lsi.upc.edu/˜nlp/IQMT/-IQMT.v2.1.pdf.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of NLH-NAACL*.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*.

Audrey Le and Mark Przybocki. 2005. NIST 2005 machine translation evaluation official results. Technical report, NIST, August.

Chin-Yew Lin and Franz Josef Och. 2004a. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of ACL*.

Chin-Yew Lin and Franz Josef Och. 2004b. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of COLING*.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Ding Liu and Daniel Gildea. 2007. Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation. In *Proceedings of the 2007 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-07)*.

Dennis Mehay and Chris Brew. 2007. BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of HLT/NAACL*.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd LREC*.

Karolina Owczarzak, Declan Groves, Josef Van Genabith, and Andy Way. 2006. Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 148–155.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-Based Automatic Evaluation for Machine Translation. In *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, RC22176, IBM. Technical report, IBM T.J. Watson Research Center.

Chris Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Metric. In *Proceedings of LREC*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, , and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, pages 223–231.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In *Proceedings of EMNLP*.