The Contribution of Linguistic Features to Automatic Machine Translation Evaluation

Enrique Amigó[1] **Jesús Giménez**[2] **Julio Gonzalo** [1] **Felisa Verdejo**[1]
[1]UNED, Madrid
{enrique,julio,felisa}@lsi.uned.es
[2]UPC, Barcelona
jgimenez@lsi.upc.edu

Abstract

A number of approaches to Automatic MT Evaluation based on deep linguistic knowledge have been suggested. However, n-gram based metrics are still today the dominant approach. The main reason is that the advantages of employing deeper linguistic information have not been clarified yet. In this work, we propose a novel approach for meta-evaluation of MT evaluation metrics, since correlation cofficient against human judges do not reveal details about the advantages and disadvantages of particular metrics. We then use this approach to investigate the benefits of introducing linguistic features into evaluation metrics. Overall, our experiments show that (i) both lexical and linguistic metrics present complementary advantages and (ii) combining both kinds of metrics yields the most robust meta-evaluation performance.

1 Introduction

Automatic evaluation methods based on similarity to human references have substantially accelerated the development cycle of many NLP tasks, such as Machine Translation, Automatic Summarization, Sentence Compression and Language Generation. These automatic evaluation metrics allow developers to optimize their systems without the need for expensive human assessments for each of their possible system configurations. However, estimating the system output quality according to its similarity to human references is not a trivial task. The main problem is that many NLP tasks are open/subjective; therefore, different humans may generate different outputs, all of them equally valid. Thus, language variability is an issue.

In order to tackle language variability in the context of Machine Translation, a considerable effort has also been made to include deeper linguistic information in automatic evaluation metrics, both syntactic and semantic (see Section 2 for details). However, the most commonly used metrics are still based on n-gram matching. The reason is that the advantages of employing higher linguistic processing levels have not been clarified yet.

The main goal of our work is to analyze to what extent deep linguistic features can contribute to the automatic evaluation of translation quality. For that purpose, we compare – using four different test beds – the performance of 16 n-gram based metrics, 48 linguistic metrics and one combined metric from the state of the art.

Analyzing the reliability of evaluation metrics requires meta-evaluation criteria. In this respect, we identify important drawbacks of the standard meta-evaluation methods based on correlation with human judgements. In order to overcome these drawbacks, we then introduce six novel meta-evaluation criteria which represent different metric reliability dimensions. Our analysis indicates that: (i) both lexical and linguistic metrics have complementary advantages and different drawbacks; (ii) combining both kinds of metrics is a more effective and robust evaluation method across all meta-evaluation criteria.

In addition, we also perform a qualitative analysis of one hundred sentences that were incorrectly evaluated by state-of-the-art metrics. The analysis confirms that deep linguistic techniques are necessary to avoid the most common types of error.

Section 2 examines the state of the art Section 3 describes the test beds and metrics considered in our experiments. In Section 4 the correlation between human assessors and metrics is computed, with a discussion of its drawbacks. In Section 5 different quality aspects of metrics are analysed. Conclusions are drawn in the last section.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 306–314,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

2 Previous Work on Machine Translation Meta-Evaluation

Insofar as automatic evaluation metrics for machine translation have been proposed, different meta-evaluation frameworks have been gradually introduced. For instance, Papineni et al. (2001) introduced the BLEU metric and evaluated its reliability in terms of Pearson correlation with human assessments for adequacy and fluency judgements. With the aim of overcoming some of the deficiencies of BLEU, Doddington (2002) introduced the NIST metric. Metric reliability was also estimated in terms of correlation with human assessments, but over different document sources and for a varying number of references and segment sizes. Melamed et al. (2003) argued, at the time of introducing the GTM metric, that Pearson correlation coefficients can be affected by scale properties, and suggested, in order to avoid this effect, to use the non-parametric Spearman correlation coefficients instead.

Lin and Och (2004) experimented, unlike previous works, with a wide set of metrics, including NIST, WER (Nießen et al., 2000), PER (Tillmann et al., 1997), and variants of ROUGE, BLEU and GTM. They computed both Pearson and Spearman correlation, obtaining similar results in both cases. In a different work, Banerjee and Lavie (2005) argued that the measured reliability of metrics can be due to averaging effects but might not be robust across translations. In order to address this issue, they computed the translation-by-translation correlation with human judgements (i.e., correlation at the segment level).

All that metrics were based on n-gram overlap. But there is also extensive research focused on including linguistic knowledge in metrics (Owczarzak et al., 2006; Reeder et al., 2001; Liu and Gildea, 2005; Amigó et al., 2006; Mehay and Brew, 2007; Giménez and Màrquez, 2007; Owczarzak et al., 2007; Popovic and Ney, 2007; Giménez and Màrquez, 2008b) among others. In all these cases, metrics were also evaluated by means of correlation with human judgements.

In a different research line, several authors have suggested approaching automatic evaluation through the combination of individual metric scores. Among the most relevant let us cite research by Kulesza and Shieber (2004), Albrecht and Hwa (2007). But finding optimal metric combinations requires a meta-evaluation criterion. Most approaches again rely on correlation with human judgements. However, some of them measured the reliability of metric combinations in terms of their ability to discriminate between human translations and automatic ones (*human likeness*) (Amigó et al., 2005). .

In this work, we present a novel approach to meta-evaluation which is distinguished by the use of additional easily interpretable meta-evaluation criteria oriented to measure different aspects of metric reliability. We then apply this approach to find out about the advantages and challenges of including linguistic features in meta-evaluation criteria.

3 Metrics and Test Beds

3.1 Metric Set

For our study, we have compiled a rich set of metric variants at three linguistic levels: lexical, syntactic, and semantic. In all cases, translation quality is measured by comparing automatic translations against a set of human references.

At the lexical level, we have included several standard metrics, based on different similarity assumptions: edit distance (WER, PER and TER), lexical precision (BLEU and NIST), lexical recall (ROUGE), and F-measure (GTM and METEOR). At the syntactic level, we have used several families of metrics based on dependency parsing (DP) and constituency trees (CP). At the semantic level, we have included three different families which operate using named entities (NE), semantic roles (SR), and discourse representations (DR). A detailed description of these metrics can be found in (Giménez and Màrquez, 2007).

Finally, we have also considered ULC, which is a very simple approach to metric combination based on the unnormalized arithmetic mean of metric scores, as described by Giménez and Màrquez (2008a). ULC considers a subset of metrics which operate at several linguistic levels. This approach has proven very effective in recent evaluation campaigns. Metric computation has been carried out using the IQ*MT* Framework for Automatic MT Evaluation (Giménez, 2007)[1]. The simplicity of this approach (with no training of the metric weighting scheme) ensures that the potential advantages detected in our experiments are not due to overfitting effects.

[1] http://www.lsi.upc.edu/~nlp/IQMT

	2004		2005	
	AE	CE	AE	CE
#references	5	5	5	4
#systems$_{\text{assessed}}$	5	10	5+1	5
#cases$_{\text{assessed}}$	347	447	266	272

Table 1: NIST 2004/2005 MT Evaluation Campaigns. Test bed description

3.2 Test Beds

We use the test beds from the 2004 and 2005 NIST MT Evaluation Campaigns (Le and Przybocki, 2005)[2]. Both campaigns include two different translations exercises: Arabic-to-English ('AE') and Chinese-to-English ('CE'). Human assessments of adequacy and fluency, on a 1-5 scale, are available for a subset of sentences, each evaluated by two different human judges. A brief numerical description of these test beds is available in Table 1. The corpus AE05 includes, apart from five automatic systems, one human-aided system that is only used in our last experiment.

4 Correlation with Human Judgements

4.1 Correlation at the Segment vs. System Levels

Let us first analyze the correlation with human judgements for linguistic vs. n-gram based metrics. Figure 1 shows the correlation obtained by each automatic evaluation metric at system level (horizontal axis) versus segment level (vertical axis) in our test beds. Linguistic metrics are represented by grey plots, and black plots represent metrics based on n-gram overlap.

The most remarkable aspect is that there exists a certain trade-off between correlation at segment versus system level. In fact, this graph produces a negative Pearson correlation coefficient between system and segment levels of 0.44. In other words, depending on how the correlation is computed, the relative predictive power of metrics can swap. Therefore, we need additional meta-evaluation criteria in order to clarify the behavior of linguistic metrics as compared to n-gram based metrics.

However, there are some exceptions. Some metrics achieve high correlation at both levels. The first one is ULC (the circle in the plot), which combines both kind of metrics in a heuristic way (see Section 3.1). The metric nearest to ULC is DP-O_r-$\star$, which computes lexical overlapping but on dependency relationships. These results are a first evidence of the advantages of combining metrics at several linguistic processing levels.

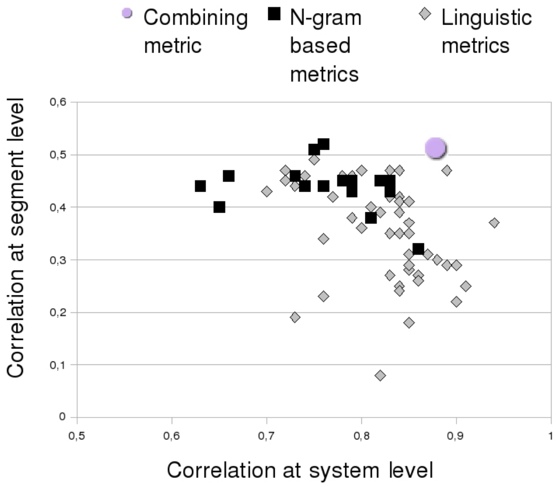

Figure 1: Averaged Pearson correlation at system vs. segment level over all test beds.

4.2 Drawbacks of Correlation-based Meta-evaluation

Although correlation with human judgements is considered the standard meta-evaluation criterion, it presents serious drawbacks. With respect to correlation at system level, the main problem is that the relative performance of different metrics changes almost randomly between testbeds. One of the reasons is that the number of assessed systems per testbed is usually low, and then correlation has a small number of samples to be estimated with. Usually, the correlation at system level is computed over no more than a few systems.

For instance, Table 2 shows the best 10 metrics in CE05 according to their correlation with human judges at the system level, and then the ranking they obtain in the AE05 testbed. There are substantial swaps between both rankings. Indeed, the Pearson correlation of both ranks is only 0.26. This result supports the intuition in (Banerjee and Lavie, 2005) that correlation at segment level is necessary to ensure the reliability of metrics in different situations.

However, the correlation values of metrics at segment level have also drawbacks related to their interpretability. Most metrics achieve a Pearson coefficient lower than 0.5. Figure 2 shows two possible relationships between human and metric

[2] http://www.nist.gov/speech/tests/mt

CE 2005			AE 2005	
SP-pNIST-5	0,88		SR-Mrv-_	0,95
SP-iobNIST-5	0,87		SP-pNIST-5	0,92
DR-STM-4	0,76		DR-Orp-_-b	0,91
DR-Orp-_	0,75		DP-HWC-r-4	0,89
DR-Orp-_-b	0,74		SP-cNIST-5	0,89
DP-HWC-r-4	0,73		SP-iobNIST-5	0,86
SP-cNIST-5	0,72		SR-Orv	0,82
SR-Mrv-_	0,71		SR-Orv-b	0,81
DR-STM-5	0,69		DR-STM-5	0,81
SR-Orv	0,67		DR-Orp-_	0,8
SR-Or-b	0,67		DR-STM-4	0,77
SR-Orv-b	0,67		SR-Or-b	0,73

Table 2: Metrics rankings according to correlation with human judgements using CE05 vs. AE05

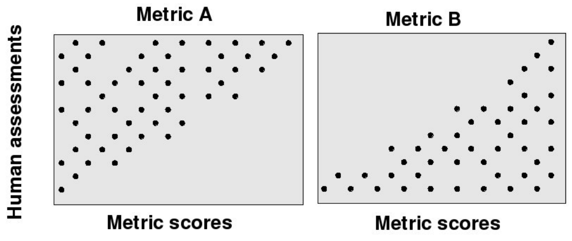

Figure 2: Human judgements and scores of two hypothetical metrics with Pearson correlation 0.5

produced scores. Both hypothetical metrics A and B would achieve a 0.5 correlation. In the case of Metric A, a high score implies a high human assessed quality, but not the reverse. This is the tendency hypothesized by Culy and Riehemann (2003). In the case of Metric B, the high scored translations can achieve both low or high quality according to human judges but low scores ensure low quality. Therefore, the same Pearson coefficient may hide very different behaviours. In this work, we tackle these drawbacks by defining more specific meta-evaluation criteria.

5 Alternatives to Correlation-based Meta-evaluation

We have seen that correlation with human judgements has serious limitations for metric evaluation. Therefore, we have focused on other aspects of metric reliability that have revealed differences between n-gram and linguistic based metrics:

1. Is the metric able to accurately reveal improvements between two systems?

2. Can we trust the metric when it says that a translation is very good or very bad?

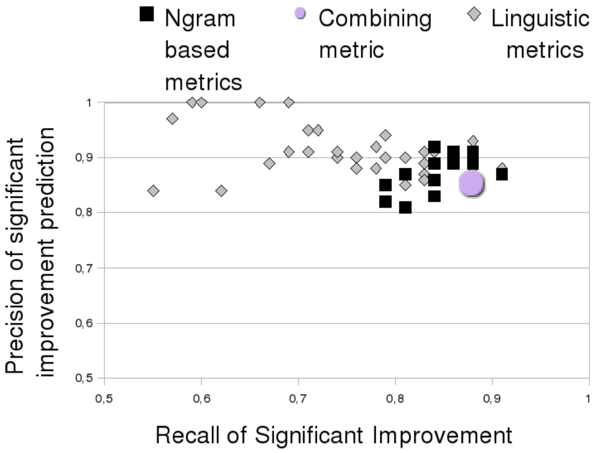

Figure 3: SIP versus SIR

3. Are metrics able to identify good translations which are dissimilar from the models?

We now discuss each of these aspects separately.

5.1 Ability of metrics to Reveal System Improvements

We now investigate to what extent a significant system improvement according to the metric implies a significant improvement according to human assessors, and viceversa. In other words: are the metrics able to detect any quality improvement? Is a metric score improvement a strong evidence of quality increase? Knowing that a metric has a 0.8 Pearson correlation at the system level or 0.5 at the segment level does not provide a direct answer to this question.

In order to tackle this issue, we compare metrics versus human assessments in terms of precision and recall over statistically significant improvements within all system pairs in the test beds. First, Table 3 shows the amount of significant improvements over human judgements according to the Wilcoxon statistical significant test ($\alpha \leq 0.025$). For instance, the testbed CE2004 consists of 10 systems, i.e. 45 system pairs; from these, in 40 cases (rightmost column) one of the systems significantly improves the other.

Now we would like to know, for every metric, if the pairs which are significantly different according to human judges are also the pairs which are significantly different according to the metric.

Based on these data, we define two meta-metrics: *Significant Improvement Precision* (SIP) and *Significant Improvement Recall* (SIR). SIP

	Systems	System pairs	Sig. imp.
CE$_{2004}$	10	45	40
AE$_{2004}$	5	10	8
CE$_{2005}$	5	10	4
AE$_{2005}$	5	10	6
Total	25	75	58

Table 3: System pairs with a significant difference according to human judgements (Wilcoxon test)

(precision) represents the reliability of improvements detected by metrics. SIR (recall) represents to what extent the metric is able to cover the significant improvements detected by humans. Let I_h be the set of significant improvements detected by human assessors and I_m the set detected by the metric m. Then:

$$\text{SIP} = \frac{|I_h \cap I_m|}{|I_m|} \qquad \text{SIR} = \frac{|I_h \cap I_m|}{|I_h|}$$

Figure 3 shows the SIR and SIP values obtained for each metric. Linguistic metrics achieve higher precision values but at the cost of an important recall decrease. Given that linguistic metrics require matching translation with references at additional linguistic levels, the significant improvements detected are more reliable (higher precision or SIP), but at the cost of recall over real significant improvements (lower SIR).

This result supports the behaviour predicted in (Giménez and Màrquez, 2009). Although linguistic metrics were motivated by the idea of modeling linguistic variability, the practical effect is that current linguistic metrics introduce additional restrictions (such as dependency tree overlap, for instance) for accepting automatic translations. Then they reward precision at the cost of recall in the evaluation process, and this explains the high correlation with human judgements at system level with respect to segment level.

All n-gram based metrics achieve SIP and SIR values between 0.8 and 0.9. This result suggests that n-gram based metrics are reasonably reliable for this purpose. Note that the combined metric, ULC (the circle in the figure), achieves results comparable to n-gram based metrics with this test[3]. That is, combining linguistic and n-gram based metrics preserves the good behavior of n-gram based metrics in this test.

5.2 Reliability of High and Low Metric Scores

The issue tackled in this section is to what extent a very low or high score according to the metric is reliable for detecting extreme cases (very good or very bad translations). In particular, note that detecting wrong translations is crucial in order to analyze the system drawbacks.

In order to define an accuracy measure for the reliability of very low/high metric scores, it is necessary to define quality thresholds for both the human assessments and metric scales. Defining thresholds for manual scores is immediate (e.g., lower than 4/10). However, each automatic evaluation metric has its own scale properties. In order to solve scaling problems we will focus on equivalent rank positions: we associate the ith translation according to the metric ranking with the quality value manually assigned to the ith translation in the manual ranking.

Being $Q_h(t)$ and $Q_m(t)$ the human and metric assessed quality for the translation t, and being rank$_h(t)$ and rank$_m(t)$ the rank of the translation t according to humans and the metric, the normalized metric assessed quality is:

$$Q_{N_m}(t) = Q_h(t')|\ (\text{rank}_h(t') = \text{rank}_m(t))$$

In order to analyze the reliability of metrics when identifying wrong or high quality translations, we look for contradictory results between the metric and the assessments. In other words, we look for metric errors in which the quality estimated by the metric is low ($Q_{N_m}(t) \leq 3$) but the quality assigned by assessors is high ($Q_h(t) \geq 5$) or viceversa ($Q_{N_m}(t) \geq 7$ and $Q_h(t) \leq 4$).

The vertical axis in Figure 4 represents the ratio of errors in the set of low scored translations according to a given metric. The horizontal axis represents the ratio of errors over the set of high scored translations. The first observation is that all metrics are less reliable when they assign low scores (which corresponds with the situation A described in Section 4.2). For instance, the best metric erroneously assigns a low score in more than 20% of the cases. In general, the linguistic metrics do not improve the ability to capture wrong translations (horizontal axis in the figure). However, again, the combining metric ULC achieves the same reliability as the best n-gram based metric.

[3] Notice that we just have 75 significant improvement samples, so small differences in SIP or SIR have no relevance

In order to check the robustness of these results, we computed the correlation of individual metric failures between test beds, obtaining 0.67 Pearson for the lowest correlated test bed pair (AE$_{\mathbf{2004}}$ and CE$_{\mathbf{2005}}$) and 0.88 for the highest correlated pair (AE$_{\mathbf{2004}}$ and CE$_{\mathbf{2004}}$).

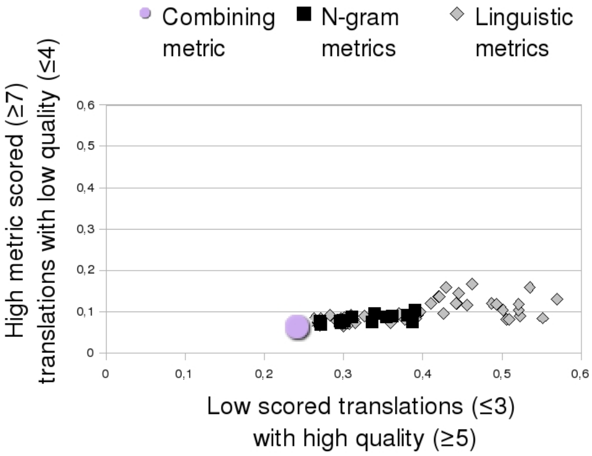

Figure 4: Counter sample ratio for high vs low metric scored translations

5.2.1 Analysis of Evaluation Samples

In order to shed some light on the reasons for the automatic evaluation failures when assigning low scores, we have manually analyzed cases in which a metric score is low but the quality according to humans is high ($Q_{N_m} \leq 3$ and $Q_h \geq 7$). We have studied 100 sentence evaluation cases from representatives of each metric family including: 1-PER, BLEU, DP-O_r-$\star$, GTM ($e = 2$), METEOR and ROUGE$_L$. The evaluation cases have been extracted from the four test beds. We have identified four main (non exclusive) failure causes:

Format issues, e.g. *"US* " vs "*United States*"). Elements such as abbreviations, acronyms or numbers which do not match the manual translation.

Pseudo-synonym terms, e.g. "*US* **Scheduled** *the Release*" vs. "*US* **set** *to Release*").) In most of these cases, synonymy can only be identified from the discourse context. Therefore, terminological resources (e.g., WordNet) are not enough to tackle this problem.

Non relevant information omissions, e.g. *"Thank you"* vs. *"Thank you very much"* or *"dollar"* vs. *"US dollar"*)). The translation system obviates some information which, in context, is not considered crucial by the human assessors. This effect is specially important in short sentences.

Incorrect structures that change the meaning while maintaining the same idea (e.g., "*Bush Praises NASA 's Mars Mission*" vs " *Bush praises nasa of Mars mission*").

Note that all of these kinds of failure - except formatting issues - require deep linguistic processing while n-gram overlap or even synonyms extracted from a standard ontology are not enough to deal with them. This conclusion motivates the incorporation of linguistic processing into automatic evaluation metrics.

5.3 Ability to Deal with Translations that are Dissimilar to References.

The results presented in Section 5.2 indicate that a high score in metrics tends to be highly related to truly good translations. This is due to the fact that a high word overlapping with human references is a reliable evidence of quality. However, in some cases the translations to be evaluated are not so similar to human references.

An example of this appears in the test bed NIST05AE which includes a human-aided system, LinearB (Callison-Burch, 2005). This system produces correct translations whose words do not necessarily overlap with references. On the other hand, a statistics based system tends to produce incorrect translations with a high level of lexical overlapping with the set of human references. This case was reported by Callison-Burch et al. (2006) and later studied by Giménez and Màrquez (2007). They found out that lexical metrics fail to produce reliable evaluation scores. They favor systems which share the expected reference sublanguage (e.g., statistical) and penalize those which do not (e.g., LinearB).

We can find in our test bed many instances in which the statistical systems obtain a metric score similar to the assisted system while achieving a lower mark according to human assessors. For instance, for the following translations, ROUGE$_L$ assigns a slightly higher score to the output of a statistical system which contains a lot of grammatical and syntactical failures.

Human assisted system: *The Chinese President made unprecedented criticism of the leaders of Hong Kong after political failings in the former British colony on Monday* . Human assessment=8.5.

Statistical system: *Chinese President Hu Jintao today unprecedented criticism to the leaders of Hong Kong wake political and financial failure in the former British colony.* Human assessment=3.

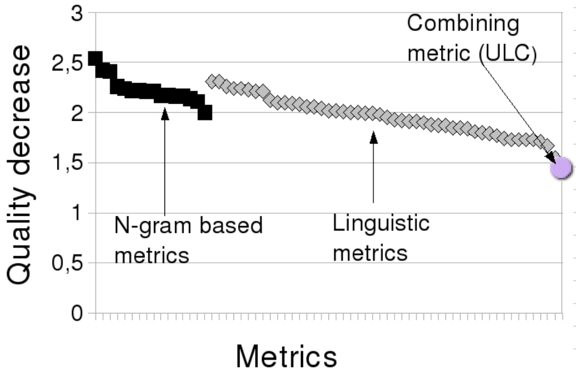

Figure 5: Maximum translation quality decreasing over similarly scored translation pairs.

In order to check the metric resistance to be cheated by translations with high lexical overlapping, we estimate the quality decrease that we could cause if we optimized the human-aided translations according to the automatic metric. For this, we consider in each translation case c, the worse automatic translation t that equals or improves the human-aided translation t_h according to the automatic metric m. Formally the averaged quality decrease is:

$$\text{Quality decrease}(m) =$$

$$\text{Avg}_c(\text{max}_t(Q_h(t_h) - Q_h(t) | Q_m(t_h) \leq Q_m(t)))$$

Figure 5 illustrates the results obtained. All metrics are suitable to be cheated, assigning similar or higher scores to worse translations. However, linguistic metrics are more resistant. In addition, the combined metric ULC obtains the best results, better than both linguistic and n-gram based metrics. Our conclusion is that including higher linguistic levels in metrics is relevant to prevent ungrammatical n-gram matching to achieve similar scores than grammatical constructions.

5.4 The Oracle System Test

In order to obtain additional evidence about the usefulness of combining evaluation metrics at different processing levels, let us consider the following situation: given a set of reference translations we want to train a combined system that takes the most appropriate translation approach for each text segment. We consider the set of translations system presented in each competition as the translation approaches pool. Then, the upper bound on the quality of the combined system is given by the predictive power of the employed automatic evaluation metric. This upper bound is obtained by selecting the highest scored translation t according to a specific metric m for each translation case c. The Oracle System Test (OST) consists of computing the averaged human assessed quality Q_h of the selected translations according to human assessors across all cases. Formally:

$$\text{OST}(m) = \text{Avg}_c(Q_h(\text{Argmax}_t(Q_m(t))|t \in c))$$

Metric	OST
maxOST	**6.72**
ULC	5.79
ROUGE$_W$	5.71
DP-O_r-$\star$	5.70
CP-O_c-$\star$	5.70
NIST	5.70
randOST	5.20
minOST	3.67

Table 4: Metrics ranked according to the Oracle System Test

We use the sum of adequacy and fluency, both in a 1-5 scale, as a global quality measure. Thus, OST scores are in a 2-10 range. In summary, the OST represents the best combined system that could be trained according to a specific automatic evaluation metric.

Table 4 shows OST values obtained for the best metrics. In the table we have also included a random, a maximum (always pick the best translation according to humans) and a minimum (always pick the worse translation according to human) OST for all [4]. The most remarkable result in Table 4 is that metrics are closer to the random baseline than to the upperbound (maximum OST). This result confirms the idea that an improvement on metric reliability could contribute considerably to the systems optimization process. However, the key point is that the combined metric, ULC, improves all the others (5.79 vs. 5.71), indicating the importance of combining n-gram and linguistic features.

6 Conclusions

Our experiments show that, on one hand, traditional n-gram based metrics are more or equally

[4] In all our experiments, the meta-metric values are computed over each test bed independently before averaging in order to assign equal relevance to the four possible contexts (test beds)

reliable for estimating the translation quality at the segment level, for predicting significant improvement between systems and for detecting poor and excellent translations.

On the other hand, linguistically motivated metrics improve n-gram metrics in two ways: (i) they achieve higher correlation with human judgements at system level and (ii) they are more resistant to reward poor translations with high word overlapping with references.

The underlying phenomenon is that, rather than managing the linguistics variability, linguistic based metrics introduce additional restrictions for assigning high scores. This effect decreases the recall over significant system improvements achieved by n-gram based metrics and does not solve the problem of detecting wrong translations. Linguistic metrics, however, are more difficult to cheat.

In general, the greatest pitfall of metrics is the low reliability of low metric values. Our qualitative analysis of evaluated sentences has shown that deeper linguistic techniques are necessary to overcome the important surface differences between acceptable automatic translations and human references.

But our key finding is that combining both kinds of metrics gives top performance according to every meta-evaluation criteria. In addition, our Combined System Test shows that, when training a combined translation system, using metrics at several linguistic processing levels improves substantially the use of individual metrics.

In summary, our results motivate: (i) working on new linguistic metrics for overcoming the barrier of linguistic variability and (ii) performing new metric combining schemes based on linear regression over human judgements (Kulesza and Shieber, 2004), training models over human/machine discrimination (Albrecht and Hwa, 2007) or non parametric methods based on reference to reference distances (Amigó et al., 2005).

Acknowledgments

This work has been partially supported by the Spanish Government, project INES/Text-Mess. We are indebted to the three ACL anonymous reviewers which provided detailed suggestions to improve our work.

References

Joshua Albrecht and Rebecca Hwa. 2007. Regression for Sentence-Level MT Evaluation with Pseudo References. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 296–303.

Enrique Amigó, Julio Gonzalo, Anselmo Pe nas, and Felisa Verdejo. 2005. QARLA: a Framework for the Evaluation of Automatic Summarization. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 280–289.

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. 2006. MT Evaluation: Human-Like vs. Human Acceptable. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 17–24.

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 MT 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 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Chris Callison-Burch. 2005. Linear B system description for the 2005 NIST MT evaluation exercise. In *Proceedings of the NIST 2005 Machine Translation Evaluation Workshop*.

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 International Conference on Human Language Technology*, pages 138–145.

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*, pages 256–264.

Jesús Giménez and Lluís Màrquez. 2008a. Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, pages 319–326.

Jesús Giménez and Lluís Màrquez. 2008b. On the Robustness of Linguistic Features for Automatic MT Evaluation. (Under submission).

Jesús Giménez and Lluís Màrquez. 2009. On the Robustness of Syntactic and Semantic Features for Automatic MT Evaluation. In *Proceedings of the 4th Workshop on Statistical Machine Translation (EACL 2009)*.

Jesús Giménez. 2007. IQMT v 2.0. Technical Manual (LSI-07-29-R). Technical report, TALP Research Center. LSI Department. `http://www.lsi.upc.edu/~nlp/IQMT/IQMT.v2.1.pdf`.

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 (TMI)*, pages 75–84.

Audrey Le and Mark Przybocki. 2005. NIST 2005 machine translation evaluation official results. In *Official release of automatic evaluation scores for all submissions, August*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

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 MT and/or Summarization*, pages 25–32.

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 the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (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. Labelled Dependencies in Machine Translation Evaluation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 104–111.

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

Maja Popovic and Hermann Ney. 2007. Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 48–55, Prague, Czech Republic, June. Association for Computational Linguistics.

Florence Reeder, Keith Miller, Jennifer Doyon, and John White. 2001. The Naming of Things and the Confusion of Tongues: an MT Metric. In *Proceedings of the Workshop on MT Evaluation "Who did what to whom?" at Machine Translation Summit VIII*, pages 55–59.

Christoph Tillmann, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP based Search for Statistical Translation. In *Proceedings of European Conference on Speech Communication and Technology*.