Corroborating Text Evaluation Results with Heterogeneous Measures

Enrique Amigó † Julio Gonzalo † Jesús Giménez ‡ Felisa Verdejo†

† UNED, Madrid
{enrique,julio,felisa}@lsi.uned.es

‡ UPC, Barcelona
{jgimenez}@lsi.upc.edu

Abstract

Automatically produced texts (e.g. translations or summaries) are usually evaluated with n-gram based measures such as BLEU or ROUGE, while the wide set of more sophisticated measures that have been proposed in the last years remains largely ignored for practical purposes. In this paper we first present an in-depth analysis of the state of the art in order to clarify this issue. After this, we formalize and verify empirically a set of properties that every text evaluation measure based on similarity to human-produced references satisfies. These properties imply that corroborating system improvements with additional measures always increases the overall reliability of the evaluation process. In addition, the greater the *heterogeneity* of the measures (which is measurable) the higher their combined reliability. These results support the use of heterogeneous measures in order to consolidate text evaluation results.

1 Introduction

The automatic evaluation of textual outputs is a core issue in many Natural Language Processing (NLP) tasks such as Natural Language Generation, Machine Translation (MT) and Automatic Summarization (AS). State-of-the-art automatic evaluation methods all operate by rewarding similarities between automatically-produced candidate outputs and manually-produced reference solutions, so-called human references or models.

Over the last decade, a wide variety of measures, based on different quality assumptions, have been proposed. Recent work suggests exploiting external knowledge sources and/or deep linguistic annotation, and measure combination (see Section 2). However, original measures based on lexical matching, such as BLEU (Papineni et al., 2001a) and ROUGE (Lin, 2004) are still preferred as de facto standards in MT and AS, respectively. There are, in our opinion, two main reasons behind this fact. First, the use of a common measure certainly allows researchers to carry out objective comparisons between their work and other published results. Second, the advantages of novel measures are not easy to demonstrate in terms of correlation with human judgements.

Our goal is not to answer which is the most reliable metric or to propose yet another novel measure. Rather than this, we first analyze in depth the state of the art, concluding that it is not easy to determine the reliability of a measure. In absence of a clear proof of the advantages of novel measures, system developers naturally tend to prefer well-known standard measures. Second, we formalize and check empirically two intrinsic properties that any evaluation measure based on similarity to human-produced references satisfies. Assuming that a measure satisfies a set of basic formal constraints, these properties imply that corroborating a system comparison with additional measures always increases the overall reliability of the evaluation process, even when the added measures have a low correlation with human judgements. In most papers, evaluation results are corroborated with similar n-gram based measures (eg. BLEU and ROUGE). However, according to our second property, the greater the *heterogeneity* of

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 455–466,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

the measures (which is measurable) the higher their reliability. The practical implication is that, corroborating evaluation results with measures based on higher linguistic levels increases the heterogeneity, and therefore, the reliability of evaluation results.

2 State of the Art

2.1 Individual measures

Among NLP disciplines, MT probably has the widest set of automatic evaluation measures. The dominant approach to automatic MT evaluation is, today, based on lexical metrics (also called n-gram based metrics). These metrics work by rewarding lexical similarity between candidate translations and a set of manually-produced reference translations. Lexical metrics can be classified according to how they compute similarity. Some are based on edit distance, e.g., WER (Nießen et al., 2000), PER (Tillmann et al., 1997), and TER (Snover et al., 2006). Other metrics are based on computing lexical precision, e.g., BLEU (Papineni et al., 2001b) and NIST (Doddington, 2002), lexical recall, e.g., ROUGE (Lin and Och, 2004a) and CDER (Leusch et al., 2006), or a balance between the two, e.g., GTM (Melamed et al., 2003; Turian et al., 2003b), METEOR (Banerjee and Lavie, 2005), BLANC (Lita et al., 2005), SIA (Liu and Gildea, 2006), MAXSIM (Chan and Ng, 2008), and O_l (Giménez, 2008).

The lexical measure BLEU has been criticized in many ways. Some drawbacks of BLEU are the lack of interpretability (Turian et al., 2003a), the fact that it is not necessary to increase BLEU to improve systems (Callison-burch and Osborne, 2006), the overscoring of statistical MT systems (Le and Przybocki, 2005), the low reliability over rich morphology languages (Homola et al., 2009), or even the fact that a poor system translation of a book can obtain higher BLEU results than a manually produced translation (Culy and Riehemann, 2003).

The reaction to these criticisms has been focused on the development of more sophisticated measures in which candidate and reference translations are automatically annotated and compared at different linguistic levels. Some of the features employed include parts of speech (Popovic and Ney, 2007; Giménez and Màrquez, 2007), syntactic dependencies (Liu and Gildea, 2005; Giménez and Màrquez, 2007; Owczarzak et al., 2007a; Owczarzak et al., 2007b; Owczarzak et al., 2008; Chan and Ng, 2008; Kahn et al., 2009), CCG parsing (Mehay and Brew, 2007), syntactic constituents (Liu and Gildea, 2005; Giménez and Màrquez, 2007), named entities (Reeder et al., 2001; Giménez and Màrquez, 2007), semantic roles (Giménez and Màrquez, 2007), discourse representations (Giménez, 2008), and textual entailment features (Padó et al., 2009). In general, when a higher linguistic level is incorporated, linguistic features at lower levels are preserved.

The proposals for summarization evaluation are less numerous. Some proposals for AS tasks are based on syntactic units (Tratz and Hovy, 2008), dependency triples (Owczarzak, 2009) or convolution kernels (Hirao et al., 2005) which reported some reliability improvement over ROUGE in terms of correlation with human judgements.

In general, however, it is not easy to determine clearly the contribution of deeper linguistic knowledge in those proposals. In the case of MT, improvements versus BLEU have been reported (Liu and Gildea, 2005; Kahn et al., 2009), but not over a more elaborated metric such as METEOR (Mehay and Brew, 2007; Chan and Ng, 2008). Besides, controversial results on their performance at sentence vs system level have been reported in shared evaluation tasks (Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010).

2.2 Combined measures

Several researchers have suggested integrating heterogeneous measures. Some of them optimize the measure combination function according to the metric's ability to emulate the behavior of human assessors (i.e., correlation with human assessments). For instance, using linear combinations (Padó et al., 2009; Liu and Gildea, 2007; Giménez and Màrquez, 2008), Decision Trees (Akiba et al., 2001; Quirk, 2004), regression based algorithms (Paul et al., 2007; Albrecht and Hwa, 2007a; Albrecht and Hwa, 2007b) or a variety of supervised machine learning algorithms(Quirk et al., 2005; Corston-Oliver et al., 2001; Kulesza and Shieber, 2004; Gamon et al., 2005; Amigó et al., 2005).

Some of these works report evidence on the contribution of combining heterogeneous measures. For instance, Albrecht and Hwa included syntax-based measures together with lexical measures, outperforming other combination schemes (Albrecht and Hwa, 2007a; Albrecht and Hwa, 2007b). Liu and Gildea, after examining the contribution of each component metric, found that *"metrics showing different properties of a sentence are more likely to make a good combined metric"*(Liu and Gildea, 2007). Akiba et al., which combined multiple edit-distance features based on lexical, morphosyntactic and lexical semantic information, observed that their approach improved single editing distance for several data sets (Akiba et al., 2001). More evidence was provided by Corston and Oliver. They showed that results on the task of discriminating between manual and automatic translations improve when combining linguistic and n-gram based features. In addition, they showed that this mixed combination improved over the combination of linguistic or n-gram based measures alone (Corston-Oliver et al., 2001). (Padó et al., 2009) reported a reliability improvement by including measures based on textual entailment in the set. In (Giménez and Màrquez, 2008), a simple arithmetic mean of scores for combining measures at different linguistic levels was applied with remarkable results in recent shared evaluation tasks (Callison-Burch et al., 2010).

2.3 Meta-evaluation criteria

Meta-evaluation methods have been gradually introduced together with evaluation measures. For instance, Papineni et al. (2001b) evaluated the reliability of the BLEU metric according to its ability to emulate human assessors, as measured in terms of Pearson correlation with human assessments of adequacy and fluency at the document level. The measure NIST (Doddington, 2002) was meta-evaluated also 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. They suggested using the non-parametric Spearman correlation coefficients instead. Lin and Och meta-evaluated ROUGE over both Pearson and Spearman correlation over a wide set of metrics, including NIST, WER, PER, and variants of ROUGE, BLEU and GTM. They obtained similar results in both cases (Lin and Och, 2004a). Banerjee and Lavie (2005) argued that the reliability of metrics at the document level can be due to averaging effects but might not be robust across sentence translations. In order to address this issue, they computed the translation-by-translation correlation with human assessments (i.e., correlation at the sentence level).

However, correlation with human judgements is not enough to determine the reliability of measures. First, correlation at sentence level (unlike correlation at system level) tends to be low and difficult to interpret. Second, correlation at system and segment levels can produce contradictory results. In (Amigó et al., 2009) it is observed that higher linguistic levels in measures increases the correlation with human judgements at the system level at the cost of correlation at the segment level. As far as we know, a clear explanation for these phenomena has not been provided yet.

Third, a high correlation at system level does not ensure a high reliability. Culy and Rieheman observed that, although BLEU can achieve a high correlation at system level in some test suites, it over-scores a poor automatic translation of "Tom Sawyer" against a human produced translation (Culy and Riehemann, 2003). This meta-evaluation criterion based on the ability to discern between manual and automatic translations have been referred to as *human likeness* (Amigó et al., 2006), in contrast to correlation with human judgements which is referred to as *human acceptability*. Examples of meta-measures based on this criterion are ORANGE (Lin and Och, 2004b) and KING (Amigó et al., 2005). In addition, many of the approaches to metric combination described in Section 2.2 take human likeness as the optimization criterion (Corston-Oliver et al., 2001; Kulesza and Shieber, 2004; Gamon et al., 2005). The main advantage of meta-evaluation based on human likeness is that, since human assessments are not required, metrics can be evaluated over larger test beds. However, the meta-evaluation in terms of *human likeness* is difficult to interpret.

2.4 The use of evaluation measures

In general, the state of the art includes a wide set of results that show the drawbacks of n-gram based measures as BLEU, and a wide set of proposals for new single and combined measures which are meta-evaluated in terms of human acceptability (i.e., their ability to emulate human judges, typically measured in terms of correlation with human judgements) or human-likeness (i.e., their ability to discern between automatic and human translations) (Amigó et al., 2006). However, the original measures BLEU and ROUGE are still preferred.

We believe that one of the reasons is the lack of an in-depth study on to what extent providing additional evaluation results with other metrics contributes to the reliability of such results. The state of the art suggests that the use of heterogeneous measures can improve the evaluation reliability. However, as far as we know, there is no comprehensive analysis on the contribution of novel measures when corroborating evaluation results with additional measures.

3 Similarity Based Evaluation Measures

In general, automatic evaluation measures applied in tasks like MT or AS are similarity measures between system outputs and human references. These measures are related with precision, recall or overlap over specific types of linguistic units. For instance, ROUGE measures n-gram recall. Other measures that work at higher linguistic levels apply precision, recall or overlap of linguistic components such as dependency relations, grammatical categories, semantic roles, etc.

In order to delimit our hypothesis, let us first define what is a similarity measure in this context. Unfortunately, as far as we know, there is no formal concept covering the properties of current evaluation similarity measures. A close concept is that of *"metric"* or *"distance function"*. But, actually, measures such as ROUGE or BLEU are not proper *"metrics"*, because they do not satisfy the *symmetry* and the *triangle inequality* properties. Therefore, we need a new definition.

Being Ω the universe of system outputs s and gold-standards g, we assume that a *similarity measure*, in our context, is a function $x : \Omega^2 \longrightarrow \Re$ such that there exists a decomposition function $f : \Omega \longrightarrow \{e_1..e_n\}$ (e.g., words or other linguistic units or relationships) satisfying the following constraints: (i) maximum similarity is achieved only when then the decomposition of the system output resembles exactly the gold-standard decomposition; and (ii) growing overlap or removing non overlapped elements implies growing x. Formally, if x ranges from 0 to 1:

$$f(s) = f(g) \leftrightarrow x(s,g) = 1$$
$$(f(s) = f(s') \cup \{e \in f(g) \setminus f(s')\}) \rightarrow x(s,g) > x(s',g)$$
$$(f(s) = f(s') - \{e \in f(s') \setminus f(g)\}) \rightarrow x(s,g) > x(s',g)$$

For instance, a random function and the reversal of a similarity funtion ($f'(s) = \frac{1}{f(s)}$) do not satisfy these constraints. While the F measure over *Precision* and *Recall* satisfies these constraints[1], precision and recall in isolation do not satisfy all of them: maximum recall can be achieved without resembling the goldstandard text decomposition; and maximum precision can be achieved with only a few overlapped elements.

BLEU (Papineni et al., 2001a) computes the n-gram precision while the metric ROUGE (Lin and Och, 2004a) computes the n-gram recall. However, in general, both metrics satisfy all the constraints, given that BLEU includes a brevity penalty and ROUGE penalizes or limits the system output length. The measure METEOR creates an alignment between the two strings (Banerjee and Lavie, 2005). This overlap-based measure satisfies also the previous constraints. Measures based on edit distance over n-grams (Tillmann et al., 1997; Nießen et al., 2000) or other linguistic units (Akiba et al., 2001; Popovic and Ney, 2007) match also our definition of similarity measure. The editing distance is minimum when the two compared text are equal. The more the evaluated text contains elements from the gold-standard the more the editing distance is reduced (higher similarity). The word ordering can be also expressed in terms of a decomposition function. A similar reasoning applies to every relevant measure in the state-of-the art.

4 Data Sets and Measures

4.1 Data sets

In this paper, we provide empirical results for MT and AS. For MT, we use the data sets from the Arabic-to-English (AE) and Chinese-to-English (CE) NIST MT Evaluation campaigns in 2004 and

[1] There is an exception. In an extreme case, when recall is zero, removing non overlapped elements does not modify the F measure.

	$\mathbf{AE_{2004}}$	$\mathbf{CE_{2004}}$	$\mathbf{AE_{2005}}$	$\mathbf{CE_{2005}}$
#human-references	5	5	5	4
#systems	5	10	7	10
#system-outputs-assessed	5	10	6	5
#system-outputs	1,353	1,788	1,056	1,082
#outputs-assessed per-system	347	447	266	272

Table 1: Description of the test beds from 2004 and 2005 NIST MT evaluation campaigns used in the experiments throughout the paper.

	DUC 2005	DUC 2006
#human-references	3-4	3-4
#systems	32	35
#system-outputs-assessed	32	35
#system-outputs	50	50
#outputs-assessed per-system	50	50

Table 2: Description of the test beds from 2005 and 2006 DUC evaluation campaigns used in the experiments throughout the paper.

2005[2]. Both include two translations exercises: for the 2005 campaign we contacted each participant individually and asked for permission to use their data[3]. In our experiments, we take the sum of adequacy and fluency, both in a 1-5 scale, as a global measure of quality (LDC, 2005). Thus, human assessments are in a 2-10 scale. For AS, we have used the AS test suites developed in the DUC 2005 and DUC 2006 evaluation campaigns[4]. This AS task was to generate a question focused summary of 250 words from a set of 25-50 documents to a complex question. Summaries were evaluated according to several criteria. Here, we will consider the responsiveness judgements, in which the quality score was an integer between 1 and 5. See Tables 1 and 2 for a brief quantitative description of these test beds.

4.2 Measures

As for evaluation measures, for MT we have used a rich set of 64 measures provided within the ASIYA Toolkit (Giménez and Màrquez, 2010)[5]. This includes measures operating at different linguistic levels: lexical, syntactic, and semantic. At the lexical level this set includes variants of 8 measures employed in the state of the art: BLEU, NIST, GTM, METEOR, ROUGE, WER, PER and TER. In addition, we have included a basic measure O_l that computes the lexical overlap without considering word ordering. All these measures have similar granularity. They use n-grams of a varying length as the basic unit with additional information provided by linguistic tools. The underlying similarity criteria include precision, recall, overlap, or edit rate, and the decomposition functions include words, dependency tree nodes (DP_HWC, DP-Or, etc.), constituency parsing (CP-STM), discourse roles (DR-Or), semantic roles (SR-Or), named entities, etc. Further details on the measure set may be found in the ASIYA technical manual (Giménez and Màrquez, 2010).

According to our computations, our measures cover high and low correlations at both levels. Correlation at system level spans between 0.63 and 0.95. Correlations at sentence level ranges from 0.18 up to 0.54. We will discriminate between two subsets of

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

[3] We are grateful to a number of groups and companies who responded positively: University of Southern California Information Sciences Institute (ISI), University of Maryland (UMD), Johns Hopkins University & University of Cambridge (JHU-CU), IBM, University of Edinburgh, University of Aachen (RWTH), National Research Council of Canada (NRC), Chinese Academy of Sciences Institute of Computing Technology (ICT), Instituto Trentino di Cultura - Centro per la Ricerca Scientifica e Tecnologica(ITC-IRST), MITRE.

[4] http://duc.nist.gov/

[5] http://www.lsi.upc.edu/~nlp/Asiya

measures. The first one includes those that decompose the text into words, n-grams, stems or lexical semantic tags. This set includes BLEU, ROUGE, NIST, GTM, PER and WER families. We will refer to them as "lexical" measures. The second set are those that consider deeper linguistic levels such as parts of speech, syntactic dependencies, syntactic constituents, etc. We will refer to them as "linguistic" measures.

In the case of automatic summarization (AS), we have employed the standard variants of ROUGE (Lin, 2004). These 7 measures are ROUGE-$\{1..4\}$, ROUGE-SU, ROUGE-L and ROUGE-W. In addition we have included the reversed precision version for each variant and the F measure of both. Notice that the original ROUGE measures are oriented to recall. In total, we have 21 measures for the summarization task. All of them are based on n-gram overlap.

5 Additive reliability

As discussed in Section 2, a number of recent publications address the problem of measure combination with successful results, specially when heterogeneous measures are combined. The following property clarifies this issue and justifies the use of heterogeneous measures when corroborating evaluation results. It asserts that *the reliability of system improvements always increases when the evaluation result is corroborated by an additional similarity measure, regardless of the correlation achieved by the additional measure in isolation*.

For the sake of clarity, in the rest of the paper, we will denote the similarity $x(s, g)$ between system output s and human reference g by $x(s)$. The quality of a system output s will be referred to as $Q(s)$. Let us define the *reliability* $R(X)$ of a measure set as the probability of a real improvement (as measured by human judges) when a score improvement is observed simultaneously for all measures in the set X. :

$$R(X) \equiv P(Q(s) \geq Q(s')|x(s) \geq x(s')\ \forall x \in X)$$

According to this definition, we may not be able to predict the quality of any system output (i.e. a translation) with a highly *reliable* measure set, but we can ensure a system improvement when all measures corroborate the result. Then the *additive reliability* property can be stated as:

$$R(X \cup \{x\}) \geq R(X)$$

We could think of violating this property by adding, for instance, a measure consisting of a random function ($x'(s) = rand(0..1)$) or a reversal of the original measure ($x'(s) = 1/x(s)$). These kind of measures, however, would not satisfy the constraints defined in Section 3.

This property is based on the idea that similarity with human references according to any aspect should not imply statistically a quality decrease. Although our test suites includes measures with low correlation at segment and system level, we can confirm empirically that all of them satisfy this property.

We have developed the following experiment: taking all possible measure pairs in the test suites, we have compared their reliability as a set versus the maximal reliability of any of them (by computing the difference $R(X) - max(R(x_1), R(x_2))$). Figure 1 shows the obtained distribution of this difference for our MT and AS test suites. Remarkably, in almost every case this difference is positive.

This result has a key implication: Corroborating evaluation results with a new measure, even when it has lower correlation with human judgements, increases the reliability of results. Therefore, if the correlation with judgements is not determinant, the question is now what factor determines the contribution of the new measures. According to the following property, this factor is the heterogeneity of measures.

6 Heterogeneity

This property states that *the reliability of any measure combination is lower bounded by the heterogeneity of the measure set.* In other words, a single measure can be more or less reliable, but a system improvement according to all measures in an heterogeneous set is reliable.

Let us define the *heterogeneity* $H(X)$ of a set of measures X as, given two system outputs s and s' such that $g \neq s \neq s' \neq g$ (g is the reference text), the probability that there exist two measures that contradict each other. That is:

$$H(X) \equiv P(\exists x, x' \in X . x(s) > x(s') \wedge x'(s) < x'(s'))$$

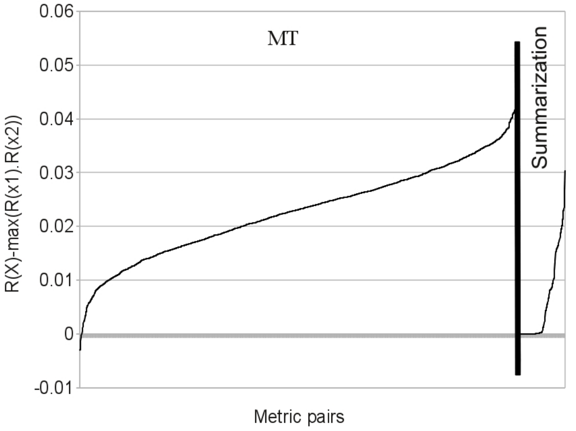

Figure 1: Additive reliability for metric pairs.

Thus, given a set X of measures, the property states that there exists a strict growing function F such that:

$$R(X) \geq F(H(X)) \text{ and } H(X) = 1 \rightarrow R(X) = 1$$

In other words, the more the similarity measures tend to contradict each other, the more a unanimous improvement over all similarity measures is reliable. Clearly, the harder it is that measures agree, the more meaningful it is when they do.

The first part is derived from the *Additive Reliability* property. Intuitively, any individual measure has zero heterogeneity. Increasing the heterogeneity implies joining measures or measure sets progressively. According to the *Additive Reliability* property, this joining implies a reliability increase. Therefore, the higher the heterogeneity, the higher the minimum Reliability achieved by the corresponding measure sets.

The second part is derived from the *Heterogeneity* definition. If $H(X) = 1$ then, for any distinct pair of outputs that differ from the reference, there exist at least two measures in the set contradicting each other. That is, $H(X) = 1$ implies that:

$$\forall s \neq s' \neq g(\exists x, x' \in X.x(s) > x(s') \wedge x'(s) < x'(s'))$$

Therefore, if one output improves the other according to all measures, then the output must be equal than the reference.

$$\neg(\exists x, x' \in X.x(s) > x(s') \wedge x'(s) < x'(s')) \rightarrow$$

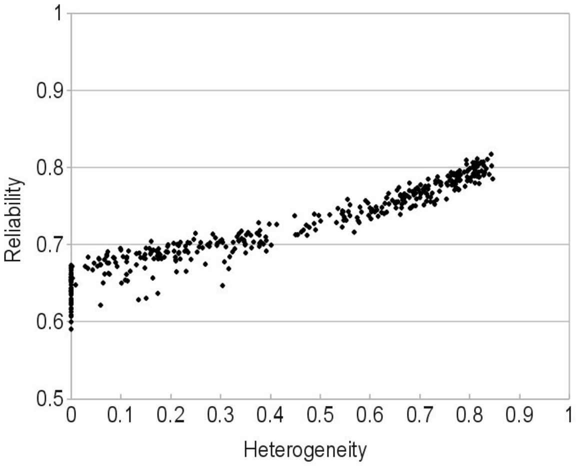

Figure 2: Heterogeneity vs. reliability in MT test suites.

$$\neg(g \neq s \neq s' \neq g) \rightarrow g = s \vee g = s'$$

According to the first constraint of similarity measures, a text that is equal to the reference achieves the maximum score:

$$g = s \rightarrow f(g) = g(s) \rightarrow \forall x.x(s) \geq x(s')$$

Finally, if we assume that the reference (human produced texts) has a maximum quality, then it will have equal or higher quality than the other output.

$$g = s \rightarrow Q(s) \geq Q(s')$$

Therefore, the reliability of the measure set is maximal. In summary, if $H(X) = 1$ then:

$$R(X) = P(Q(s) \geq Q(s') | x(s) \geq x(s') \, \forall x \in X) =$$

$$= P(Q(s) \geq Q(s') | s = g) = 1$$

Figures 2 and 3 show the relationship between the heterogeneity of randomly selected measure sets and their reliability for the MT and summarization test suites. As the figures show, the higher the heterogeneity, the higher the reliability of the measure set. The results in AS are less pronounced due to the redundancy in ROUGE measure.

Notice that the heterogeneity property does not necessarily imply a high correlation between reliability and heterogeneity. For instance, an ideal single measure would have zero heterogeneity and

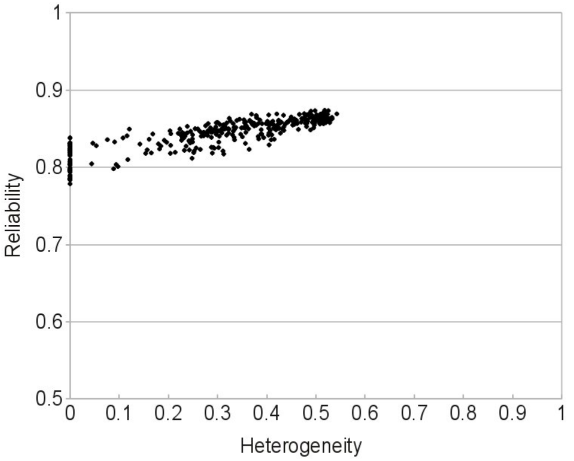

Figure 3: Heterogeneity vs. reliability in summarization test suites.

achieve maximum reliability, appearing in the top left area. The property rather brings us to the following situation: let us suppose that we have a set of single measures available which achieve a certain range of reliability. We can improve our system according to any of these measures. Without human assessments, we do not know what is the most reliable measure. But if we combine them, increasing the heterogeneity, the minimal reliability of the selected measures will be higher. This implies that combining heterogeneous measures (e.g. at high linguistic levels) that do not achieve high correlation in isolation, is better than corroborating results with any individual measure alone, such as ROUGE and BLEU, which is the common practice in the state of the art.

The main drawback of this property is that increasing the heterogeneity implies a sensitivity reduction. For instance, if $H(X) = 0.9$, then only for 10% of output pairs in the corpus there exists an improvement according to all measures. In other words, unanimous evaluation results from heterogeneous measures are reliable but harder to achieve for the system developer. The next section investigates on this issue.

Finally, Figure 4 shows that linguistic measures increase the heterogeneity of measure sets. We have generated sets of metrics of size 1 to 10 made up by lexical or lexical and linguistic metrics. As the figure shows, in the second case, the measure sets achieve a higher heterogeneity.

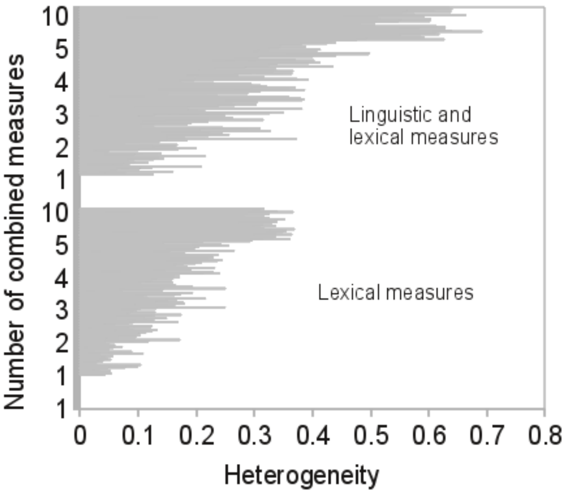

Figure 4: Heterogeneity of lexical measures vs. lexical and linguistic measures.

7 Score thresholds vs. Additive Reliability

According to the previous properties, corroborating evaluation results with several measures increases the reliability of evaluation results at the cost of sensitivity. On the other hand, increasing the score threshold of a single measure should have a similar effect. Which is then the best methodology to improve reliability? In this section we provide experimental evidence on the relationship between both ways of increasing reliability: we have found that, corroborating evaluation results over single texts with additional measures is more reliable than requiring higher score differences according to any individual measure in the set. More specifically, we have found that *the reliability of a measure set is higher than the reliability of each of the individual measures at a similar level of sensitivity.*

Formally, we define the sensitivity $S(X)$ of a metric set X as the probability of finding a score improvement within text pairs with a real (i.e. human assessed) quality improvement:

$$S(X) = P(x(s) \geq x(s')\forall x \in X|Q(s) \geq Q(s'))$$

Being $R_{th}(x)$ and $S_{th}(x)$ the reliability and sensitivity of a single measure x for a certain increase score threshold th:

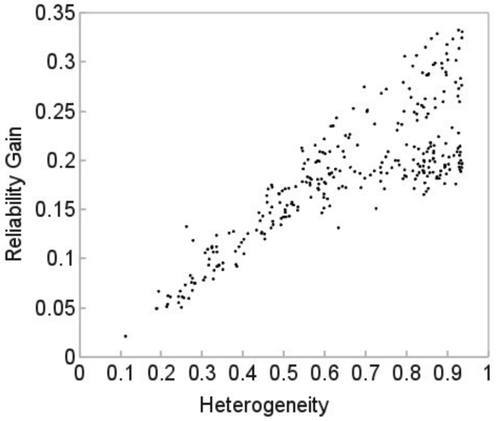

Figure 5: Heterogeneity vs. reliability Gain for MT test suites.

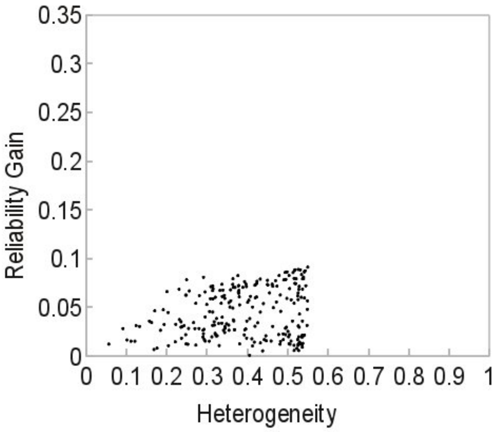

Figure 6: Heterogeneity vs. reliability Gain for MT test suites.

$$R_{th}(x) = P(Q(s) \geq Q(s')|x(s) - x(s') \geq th)$$
$$S_{th}(x) = P(x(s) - x(s') \geq th|Q(s) \geq Q(s'))$$

The property that we want to check is that, at the same sensitivity level, combining measures is more reliable than increasing the score threshold of single measures:

$$S(X) = S_{th}(x).x \in X \longrightarrow R(X) \geq R_{th}(x)$$

Note that if we had a perfect measure x_p such that $R(x_p) = S(x_p) = 1$, then combining this measure with a low reliability measure x_l would produce a lower sensitivity, but the maximal reliability would be preserved.

In order to confirm empirically this property, we have developed the following experiment: (i) We compute the reliability and sensitivity of randomly chosen measure sets over single text pairs. We have generated sets of 2,3,5,10,20 and 40 measures. In the case of summarization corpora we have combined up to 20 measures. In addition, we compute also the heterogeneity $H(X)$ of each measure set; (ii) Experimenting with different values for the threshold th, we compute the reliability of single measures for all potential sensitivity levels; (iii) For each measure set, we compare the reliability of the measure set versus the reliability of single measures at the same sensitivity level. We will refer to this as the *Reliability Gain*:

$$\text{Reliability Gain} =$$
$$R(X) - max\{R_{th}(x)/x \in X \wedge S_{th}(x) = S(X)\}$$

If there are several reliability values with the same sensitivity for a given single measures, we choose the highest reliability value for the single measure.

Figures 5 and 6 illustrate the results for the MT and AS corpora. The horizontal axis represents the Heterogeneity of measure sets, while the vertical axis represents the reliability gain. Remarkably, the reliability gain is positive for all cases in our test suites. The maximum reliability gain is 0.34 in the case of MT and 0.08 for AS (note that summarization measures are more redundant in our corpora). In both test suites, the largest information gains are obtained with highly heterogeneous measure sets.

In summary, given comparable measures in terms of reliability, corroborating evaluation results with several measures is more effective than optimizing systems according to the best measure in the set. This empirical property provides an additional evidence in favour of the use of heterogeneous measures and, in particular, of the use of linguistic measures in combination with standard lexical measures.

8 Conclusions

In this paper, we have analyzed the state of the art in order to clarify why novel text evaluation measures are not exploited by the community. Our first conclusion is that it is not easy to determine the reliability of measures, which is highly corpus-dependent and often contradictory when comparing correlation with human judgements at segment vs. system levels.

In order to tackle this issue, we have studied a number of properties that suggest the convenience of using heterogeneous measures to corroborate evaluation results. According to these properties, we can ensure that, even when if we can not determine the reliability of individual measures, corroborating a system improvement with additional measures always increases the reliability of the results. In addition, the more heterogeneous the measures employed (which is measurable), the higher the reliability of the results. But perhaps the most important practical finding is that the reliability at similar sensitivity levels by corroborating evaluation results with several measures is always higher than improving systems according to any of the combined measures in isolation.

These properties point to the practical advantages of considering linguistic knowledge (beyond lexical information) in measures, even if they do not achieve a high correlation with human judgements. Our experiments show that linguistic knowledge increases the heterogeneity of measure sets, which in turn increases the reliability of evaluation results when corroborating system comparisons with several measures.

Acknowledgements

This work has been partially funded by the Spanish Government (Holopedia, TIN2010-21128-C02 and OpenMT-2, TIN2009-14675-C03) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247762 (FAUST project, FP7-ICT-2009-4-247762).

References

Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. 2001. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. In *Proceedings of Machine Translation Summit VIII*, pages 15–20.

Joshua Albrecht and Rebecca Hwa. 2007a. A Reexamination of Machine Learning Approaches for Sentence-Level MT Evaluation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 880–887.

Joshua Albrecht and Rebecca Hwa. 2007b. 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.

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Felisa Verdejo. 2009. The contribution of linguistic features to automatic machine translation evaluation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 306–314, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 and Miles Osborne. 2006. Reevaluating the role of bleu in machine translation research. In *In EACL*, pages 249–256.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53. Revised August 2010.

Yee Seng Chan and Hwee Tou Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62.

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (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 International Conference on Human Language Technology*, pages 138–145.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-Level MT evaluation without reference translations: beyond language modeling. In *Proceedings of EAMT*, pages 103–111.

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. 2008. 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. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 1(94):77–86.

Jesús Giménez. 2008. *Empirical Machine Translation and its Evaluation*. Ph.D. thesis, Universitat Politècnica de Catalunya.

Tsutomu Hirao, Manabu Okumura, and Hideki Isozaki. 2005. Kernel-based approach for automatic evaluation of natural language generation technologies: Application to automatic summarization. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 145–152, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Petr Homola, Vladislav Kuboň, and Pavel Pecina. 2009. A simple automatic mt evaluation metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeremy G. Kahn, Matthew Snover, and Mari Ostendorf. 2009. Expected Dependency Pair Match: Predicting translation quality with expected syntactic structure. *Machine Translation*.

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.

LDC. 2005. Linguistic Data Annotation Specification: Assessment of Adequacy and Fluency in Translations. Revision 1.5. Technical report, Linguistic Data Consortium. `http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess04.pdf`.

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

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 241–248.

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 the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chin-Yew Lin and Franz Josef Och. 2004b. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

Chin-Yew Lin. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Lucian Vlad Lita, Monica Rogati, and Alon Lavie. 2005. BLANC: Learning Evaluation Metrics for MT. In *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 740–747.

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.

Ding Liu and Daniel Gildea. 2006. Stochastic Iterative Alignment for Machine Translation Evaluation. 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 539–546.

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)*, pages 41–48.

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, Josef van Genabith, and Andy Way. 2007a. Dependency-Based Automatic Evaluation for Machine Translation. In *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 80–87.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007b. Labelled Dependencies in Machine Translation Evaluation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 104–111.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2008. Evaluating machine translation with lfg dependencies. *Machine Translation*, 21(2):95–119.

Karolina Owczarzak. 2009. Depeval(summ): dependency-based evaluation for automatic summaries. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 190–198, Morristown, NJ, USA. Association for Computational Linguistics.

Sebastian Padó, Michael Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 297–305.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, jul.

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

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2007. Reducing Human Assessments of Machine Translation Quality to Binary Classifiers. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

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.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 271–279.

Chris Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Metric. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 825–828.

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.

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 the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

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*.

Stephen Tratz and Eduard Hovy. 2008. Summarization evaluation using transformed basic elements. In *In Proceedings of TAC-08. Gaithersburg, Maryland*.

Joseph Turian, Luke Shen, and I. Dan Melamed. 2003a. Evaluation of machine translation and its evaluation. In *In Proceedings of MT Summit IX*, pages 386–393.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003b. Evaluation of Machine Translation and its Evaluation. In *Proceedings of MT SUMMIT IX*.