
Article:
Fomicheva, M. and Specia, L. (2019) Taking MT evaluation metrics to extremes : beyond correlation with human judgments. Computational Linguistics, 45 (3). pp. 515-558. ISSN: 0891-2017

https://doi.org/10.1162/coli_a_00356

Taking MT Evaluation Metrics to Extremes: Beyond Correlation with Human Judgments

Marina Fomicheva
University of Sheffield
m.fomicheva@sheffield.ac.uk

Lucia Specia
Imperial College London
University of Sheffield
l.specia@imperial.ac.uk

Automatic Machine Translation (MT) evaluation is an active field of research, with a handful of new metrics devised every year. Evaluation metrics are generally benchmarked against manual assessment of translation quality, with performance measured in terms of overall correlation with human scores. Much work has been dedicated to the improvement of evaluation metrics to achieve a higher correlation with human judgments. However, little insight has been provided regarding the weaknesses and strengths of existing approaches and their behavior in different settings. In this work we conduct a broad meta-evaluation study of the performance of a wide range of evaluation metrics focusing on three major aspects. First, we analyze the performance of the metrics when faced with different levels of translation quality, proposing a local dependency measure as an alternative to the standard, global correlation coefficient. We show that metric performance varies significantly across different levels of MT quality: Metrics perform poorly when faced with low-quality translations and are not able to capture nuanced quality distinctions. Interestingly, we show that evaluating low-quality translations is also more challenging for humans. Second, we show that metrics are more reliable when evaluating neural MT than the traditional statistical MT systems. Finally, we show that the difference in the evaluation accuracy for different metrics is maintained even if the gold standard scores are based on different criteria.

1. Introduction

The use of automatic evaluation is a common practice in the field of Machine Translation (MT). It allows for cost-effective quality assessment, making it possible to compare different approaches to MT, optimize parameters of statistical MT systems, and select models in neural MT systems. The most common approach to evaluation is based on the assumption that the closer the MT output is to a human reference translation, the higher its quality. For example, the well-known metric BLEU (Papineni et al. 2002) follows a simple strategy of counting the proportion of word *n*-grams in the MT output that are also found in one or more references.

Submission received: 5 October 2018; revised version received: 19 March 2019; accepted for publication: 12 June 2019.

https://doi.org/10.1162/COLI_a_00356

BLEU has been severely criticized for several of its limitations, such as its poor performance at sentence-level and inadequate handling of recall (Callison-Burch, Osborne, and Koehn 2006). Significant work has been dedicated to developing more advanced metrics, primarily by integrating different sources of information (synonyms, paraphrases, and syntactic and semantic analysis) and using learning techniques to appropriately combine them into a single score.

The performance of evaluation metrics is typically assessed in terms of system- and sentence-level correlation with human judgments, mostly for the task of ranking alternative MT system translations for the same source segment. Existing work on meta-evaluation has extensively discussed the limitations of n-gram–based metrics (Coughlin 2003; Culy and Riehemann 2003; Koehn and Monz 2006), whereas the studies examining the contributions of more advanced strategies, for example, the integration of linguistic information (Amigó et al. 2009), are more rare. Influential evaluation campaigns such as the WMT Metrics Task receive new metric submissions every year with many recent metrics reported to outperform standard metrics like BLEU (Machácek and Bojar 2014; Stanojević et al. 2015; Bojar et al. 2016c; Bojar, Graham, and Kamran 2017). However, very little insight has been provided regarding where existing metrics succeed and where they fail, and why.

Furthermore, the performance of evaluation metrics is known to be unstable across evaluation settings. Metrics can be more or less reliable, depending on the target language (and resources available for such a language), text type and genre, type of MT system under evaluation, properties of human translation, and the quality aspect being measured (e.g., adequacy vs. fluency). A closer look at the impact of such factors on the behavior of different metrics will lead to a better understanding of existing approaches and what still needs to be improved. We conduct a broad meta-evaluation study of a wide range of metrics in varying evaluation settings focusing on three of such factors: level of MT quality, type of MT system, and type of human judgment (i.e., the criterion used to generate gold quality assessments manually).

Our first contribution is to demonstrate the effect of translation quality, as reflected in human judgments, on the performance of automatic evaluation. It has been generally assumed that the main reason for low correlation between metric scores and human judgments is a poor performance of the metrics when evaluating high-quality translations, since the metrics tend to underestimate acceptable MT outputs that differ from the available reference(s) but express the same meaning (Amigó et al. 2009; Padó et al. 2009). Via an in-depth analysis of the behavior of state-of-the-art evaluation metrics on MT outputs with varying levels of quality, we show that this is not the case. On average, metrics do a better job at evaluating high-quality MT, whereas low-quality MT evaluation appears to be more challenging. We suggest that the reason for this is two-fold. On the one hand, in the case of low-quality outputs the lack of information resulting from the absence of candidate-reference matches is much more severe. On the other hand, low-quality translation contains a higher number and variety of translation errors whose impact is difficult to measure. We show that this latter factor also affects the consistency of manual evaluation. In order to carry out this study, we borrow methods from finance and econometrics that—to the best of our knowledge—have not been applied in natural language processing. Specifically, we use a local dependency measure recently proposed by Tjøstheim and Hufthammer (2013) to describe the relation between metric scores and human judgments at different levels of translation quality.

Our second contribution is to investigate how the performance of evaluation metrics is affected by the type of MT systems. Previous work has shown, for instance,

that metrics such as BLEU heavily penalize translations from rule-based MT systems compared with translations from statistical MT systems (Callison-Burch, Osborne, and Koehn 2006). Given the recent success of neural MT, which has been leading to this becoming the de facto approach to MT, the question arises of how reliable automatic evaluation is for such systems. In this work we examine how existing evaluation metrics perform on the output of neural MT, as compared with the conventional statistical MT. We show that automatic evaluation results are more accurate for neural MT than for statistical MT because of the difference in the distribution of different types of translation errors.

Our third contribution is to show that the relative performance of different evaluation metrics is maintained across varying types of human judgments. That is to say, for example, that metrics that perform the best for evaluating adequacy also perform the best for evaluating fluency. Thus, the results from meta-evaluation with a particular type of manual assessment can be more often than not extrapolated to other quality aspects.

The rest of this article is organized as follows. Section 2 introduces related work on analyzing MT metrics from different perspectives. Sections 3 and 4 present the evaluation metrics and the data sets used in our experiments. Section 5 describes the analysis of automatic and manual evaluation in relation to MT quality levels. Section 6 analyzes the differences in metric performance for statistical and neural MT. Finally, Section 7 compares the results of meta-evaluation using different types of human judgments.

2. Background and Related Work

Automatic MT evaluation has been at the core of MT development for decades. In addition to comparing MT systems and measuring progress over time, with the advent of statistical approaches in the early 1990s it became evident that cost-effective automatic metrics with reproducible outcomes were also needed for the building of such systems (i.e., parameter tuning). A number of metrics were proposed to measure distance or similarity against one or more human (reference) translations. Simplistic metrics borrowed from speech recognition such as word error rate (WER) and its position-independent variant (PER) were soon replaced by more elaborate metrics that reward similarity beyond word-level, notably BLEU (Papineni et al. 2002), or perform comparisons at stem and synonymy levels, rather than exact match only, namely, Meteor (Banerjee and Lavie 2005). Nearly three decades on, automatic metrics still play a critical role in MT research and development and, despite a handful of metrics proposed every year, the problem is far from solved. Evidence of that is the annual campaign run by the Conference on Machine Translation (WMT), which—among other tasks—invites researchers to submit new evaluation metrics that are benchmarked against human judgments in a Metrics Task (see Bojar, Graham, and Kamran [2017], Bojar et al. [2016c], Stanojević et al. [2015], and Machácek and Bojar [2014] for the most recent task results).

Starting in 2005, WMT has conducted yearly evaluations of machine translation quality using human judgments, as well as meta-evaluation of automatic evaluation metric performance based on such human judgments. The nature of the human judgments varied over the years, from 1- to 5-point scale scores for fluency and adequacy for entire sentences or sentence constituents, to rankings of up to 5 translations from different MT systems, to a 1–100 score per sentence according to its fluency or adequacy. Different types of correlation with human judgments are computed (Pearson r, Kendal τ, etc.), depending on the time of judgment and evaluation level (corpus or segment).

For a recent summary over the various years of the meta-evaluation campaigns, we refer the reader to Bojar et al. (2016b).

Initially, meta-evaluation focused on system-level analysis (Doddington 2002; Melamed, Green, and Turian 2003; Lin and Och 2004a). In this scenario, a single measurement is provided for a set of sentences generated by an MT system. For manual evaluation, this is usually an average of sentence-level scores, whereas for automatic evaluation system-level score is computed differently by different metrics. The correlation is then computed over such average measurements collected for multiple MT systems. System-level evaluation is useful for comparing the performance of different MT systems and is generally an easy task for MT evaluation metrics. In fact, according to the meta-evaluation shared tasks, such as the Metrics Task at WMT, the vast majority of current metrics perform extremely well at ranking systems (system-level evaluation), with correlations above 0.9 with human rankings. However, MT system ranking is only one of the applications of such metrics and is not indicative of the advantages and limitations of different MT systems. It has long been shown that in order to ensure the reliability of evaluation metrics over different situations, correlation at the sentence level is necessary (Banerjee and Lavie 2005). For the purposes of assessing the performance of automatic evaluation metrics in this article, we thus concentrate on sentence-level evaluation, which allows us to observe significant differences among metrics.

Despite these major efforts to evaluate new and existing metrics, the results of the annual Metrics Tasks are limited to the correlation between metric scores and human judgments, providing no insight regarding the actual advantages and disadvantages of the participating metrics, nor their performance on different types of translation or translation systems. In fact, even the papers about the metrics themselves rarely attempt to provide a more detailed account of their performance.

The first aspect of meta-evaluation discussed in this paper is how the level of translation quality affects the performance of evaluation metrics. The difficulties faced by the metrics change, depending on the quality of MT output. High-quality translation presents the problem of acceptable variation between the MT output and human reference (Giménez and Màrquez 2010b). Low-quality translation, on the other hand, requires an ability to assess the impact of different types of MT errors (Liu and Gildea 2005). However, hardly any rigorous meta-evaluation analysis has been performed that would indicate which problem is more damaging for the overall metrics performance. Besides very few exceptions, the analysis is limited to computing the correlation with human judgments.

One notable exception is the work by Amigó et al. (2009). Following substantial research dedicated to the use of linguistic information in automatic MT evaluation, Amigó et al. (2009) analyze the benefits of introducing linguistic features into evaluation metrics. They introduce various meta-evaluation criteria to provide a better understanding of the reliability of different evaluation methods, focusing on the comparison between linguistically informed metrics and the traditional *n*-gram based approaches. First, they test the metrics capability to accurately reveal improvements between two systems. Second, they analyze to what extent a metric can be trusted if it predicts that the translation is very good or very bad. Finally, they test whether the metrics are able to identify good MT if it is different from the reference provided.

For the second criterion, Amigó et al. (2009) count the number of cases where the quality predicted by the metric is very low, while the quality predicted by the human is very high, and vice versa. This analysis is the closest to ours. However, as will be shown in what follows, we ask a different, more general question: How well

can the metrics evaluate low-quality translation and high-quality translation, without limiting ourselves to the cases where the results are contradictory between the metrics and human assessments. To answer this question, we propose a principled, formally grounded analysis of human–metric correlation at different levels of translation quality.

Another direction for meta-evaluation analysis is the consistency of human evaluation. Manual assessment is taken to be the gold standard for the performance of automatic evaluation metrics. However, translation evaluation is a challenging task not only for automatic metrics, but also for human annotators. The perception of translation quality is subjective and depends on individual background and expectations of the participants. No clear guidance is typically provided regarding what should be considered acceptable, what should not be, and to what extent. The levels of inter- and intra-annotator agreement for the MT evaluation task has been fairly low (Denkowski and Lavie 2010; Graham et al. 2017). An alternative view is to accept the possibility of multiple correct assessments and take their average as the ground truth. This method has been successfully used in the last two years of the WMT campaign. In this work, we use the data obtained using this method to determine whether it is more difficult for human annotators to deal with low-quality translations.

A final aspect of meta-evaluation analysis explored in this work is the relation between metric performance and MT approaches. It has been previously shown that the reliability of automatic evaluation varies, depending on the type of MT system being evaluated. Callison-Burch, Osborne, and Koehn (2006) found that *n*-gram–based metrics tend to favor statistical systems over rule-based ones, as they are more likely to match the sub-language (e.g., lexical choice and order) represented by reference translations. With the advent of neural MT (Bahdanau, Cho, and Bengio 2014; Sutskever, Vinyals, and Le 2014), it becomes important to test whether existing metrics perform differently on the outputs of these systems. Some recent studies have analyzed the differences between statistical MT and neural MT, concluding that neural MT reduces the number of word order errors and, in general, improves fluency, sometimes at the cost of adequacy (Junczys-Dowmunt, Dwojak, and Hoang 2016; Castilho et al. 2017; Toral and Sanchez-Cartagena 2017). However, no previous work has evaluated the performance of automatic evaluation metrics on the output of neural versus other MT approaches.

3. MT Evaluation Metrics

In this section we present the metrics used in the experiments. Although many more metrics exist, we considered those for which either implementation or results for a given data set are available, and which are less reliant on external resources. For each metric in the following description we indicate the specific implementation used. To facilitate reproducibility, we used a set of metrics from the Asiya toolkit (Giménez and Màrquez 2010a), which can be run all at once, and additionally the most recent developments in the MT evaluation, including the top metrics that participated in the WMT14–WMT17 evaluation campaigns.

Lexical Similarity. Most of the metrics used are based on the lexical similarity between the MT output and the reference translation. These are:

BLEU. (Bilingual Evaluation Understudy) (Papineni et al. 2002). Measures the similarity between MT and the reference translation based on the number of matching word *n*-grams. Specifically, BLEU score is a product between *n*-gram precision and a brevity penalty that down-scales the score for the MT outputs that are shorter in length than the

reference translation. In all the experiments with this metric, we use a smoothed version of BLEU as described by Lin and Och (2004b) with $N = 4$.

Meteor. (Denkowski and Lavie 2014). Meteor aligns MT output to the reference translation using stems, synonyms, and paraphrases, besides exact word matching, and computes candidate-reference similarity based on the proportion of aligned words in the candidate and in the reference. Different weights are assigned to the word matches, depending on the type of lexical similarity, and to function and content words. Additionally, Meteor integrates a fragmentation penalty that penalizes the differences in word order. It is based on the number of chunks (sequential word matches) in candidate-reference alignment. The final Meteor score is a parametrized combination of F-measure and fragmentation penalty.

MPEDA. (Zhang et al. 2016). MPEDA is based on Meteor but uses a domain-specific paraphrase database instead of a general one to reduce noisy paraphrase matches. To extract domain-specific paraphrases, Zhang et al. (2016) first filter the large scale general monolingual corpus into a domain-specific sub-corpus using the M-L approach (Moore and Lewis 2010), and then exploit the Markov Network model to extract paraphrase tables from that sub-corpus.

-WER. (Word Error Rate) (Nießen et al. 2000). WER is based on the edit distance defined as the minimum number of word substitutions, deletions, and insertions that need to be performed to convert MT output into the reference translation.[1]

-PER. (Position-independent Word Error Rate) (Tillmann et al. 1997). WER may be considered excessively strict for automatic MT evaluation as it does not allow any differences in word order. PER addressed this limitation by comparing MT and reference words without taking the word order into account.

-TER. (Translation Edit Rate) (Snover et al. 2006). This metric is also based on edit distance. However, in contrast to WER and PER, in TER possible edits include shifts of words and word sequences.

-TERp-A. (Snover et al. 2009). This metric enriches TER with stemming, synonyms, lookup, and paraphrase support. The metric is optimized for the adequacy criterion.

NIST. (Doddington 2002). NIST differs from BLEU in two aspects. First, to handle the low co-occurrences for larger values of N, an arithmetic mean is used instead of a geometric mean when combining the precisions of n-gram matches. Second, the n-grams are weighted depending on their frequency in a reference corpus, assuming that high frequency n-grams are less informative. We use the standard cumulative 5-gram NIST score.

ROUGE. (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Och 2004a). ROUGE computes lexical recall among n-grams up to length 4. It also allows for

1 For the metrics based on edit distance WER, PER, TER, and TERp-A, we will use -WER, -PER, -TER, and -TERp-A to make the results more easily comparable with the rest of the metrics.

considering stemming and discontinuous matchings (skip bigrams). We used five different variants from Lin and Och (2004a) implemented in Asiya toolkit:[2]

- ROUGE-n: for several N-gram lengths $N \in [1, 4]$
- ROUGE-L: longest common subsequence
- ROUGE-S: skip bigrams with no max-gap-length
- ROUGE-SU: skip bigrams with no max-gap-length, including unigrams
- ROUGE-W: weighted longest common subsequence with weighting factor $w = 1.2$

ChrF. ChrF$_\beta$ (Popovic 2015, 2016) is a recently proposed evaluation metric that calculates the F-score of character n-grams of maximal length 6. The β parameter gives β times weight to recall. Using characters instead of words helps ameliorate the sparcity of word n-gram matches and better handle morphological differences. Throughout this article we use ChrF3 with $\beta = 3$, as suggested by Popovic (2015).

Linguistic Representations. To overcome the limitations of the metrics based on lexical similarity, another family of evaluation metrics explores the use of different **linguistic representations** (morphological, syntactic, semantic, and discourse) for comparing the MT output against the reference translation. The motivation behind these metrics is, on the one hand, to abstract away from surface word forms and, on the other hand, to try to better assess the grammaticality of the MT output.

UPF-Cobalt. UPF-Cobalt (Fomicheva et al. 2015; Fomicheva and Bel 2016) is an alignment-based metric that incorporates a syntactically informed context penalty to penalize the matches of lexically similar words that play different roles in the candidate and reference sentences. The sentence-level score combines the information on lexical similarity with the average context penalty. Word similarity is detected in various ways, including cosine similarity over distributed word representations.

SP-$*$. SP metrics (Giménez and Màrquez 2010a, 2010b) measure the similarities at the level of parts of speech, word lemmas, and base phrase chunks. Sentences are automatically annotated using the SVMTool (Giménez and Marquez 2004) linguistic processors. Specifically, the following metrics are defined:

- SP-Op(*): Average overlap between words belonging to the same part of speech
- SP-Oc(*): Average overlap between words belonging to chunks of the same type
- SP-NIST$l|p|c|iob$: NIST score over sequences of lemmas, parts of speech, phrase chunks, and chunk IOB labels[3]

2 For this and the following Asiya metrics with multiple variants, we report only the best-performing variant in our experiments.

3 IOB labels indicate the position (Inside, Outside, or Beginning) and type of chunk.

DP-∗. DP metrics (Giménez and Màrquez 2010a, 2010b) capture similarities between dependency trees associated with MT outputs and reference translations. Dependency trees are obtained using MINIPAR (Lin 2003). Specifically, the following metrics are defined:

- DP-HWCM: Head-Word Chain Matching (Liu and Gildea 2005). Only chains up to length 4 are considered. Three different variants according to the item type are available:
  - DP-HWCM(w) word forms
  - DP-HWCM(c) grammatical categories
  - DP-HWCM(r) grammatical relations
- DP-Ol(*) Average lexical overlap between items according to their tree level
- DP-Oc(*) Average lexical overlap between terminal nodes according to their grammatical category
- DP-Or(*) Average lexical overlap between items according to their grammatical relationship

CP-∗. CP metrics (Giménez and Màrquez 2010a, 2010b) analyze similarities between constituent parse trees associated with MT outputs and reference translations. Constituent trees are obtained using the Charniak-Johnson's Max-Ent reranking parser (Charniak and Johnson 2005). The following measures are defined:

- CP-Op(*) Average overlap between words belonging to the same part of speech.
- CP-Oc(*) Average overlap between words belonging to constituents of the same type
- CP-STMd This measure corresponds to the Syntactic Tree Matching defined by Liu and Gildea (2005), except that overlap is used instead of precision. Subtrees up to different d depths ($d \in 4, 5, 6$) are considered.

SR∗. SR metrics (Giménez and Màrquez 2010a, 2010b) analyze similarities between MT outputs and reference translations by comparing the semantic roles (SRs) (i.e., arguments and adjuncts) that occur in them. Sentences are automatically annotated using the SwiRL package (Surdeanu and Turmo 2005). The following measures are defined:

- SR-Or(*): Average lexical overlap over semantic roles
- SR-Mr(*): Average lexical matching over semantic roles
- SR-Or: Average role overlap, i.e., overlap between semantic roles independently of their lexical realization

Feature Combination. The most recent improvements in the performance of evaluation metrics is related to the use of machine learning techniques in order to combine a wide variety of features describing different aspects of MT quality. To be able to train on WMT ranking data and produce absolute scores at test time, most of the metrics described here

(unless stated otherwise) use the learn-to-rank approach (Burges et al. 2005) for tuning the feature weights.

BEER. BEER (Stanojević and Sima'an 2014) is a trained evaluation metric with a linear model that combines lexical similarity features (precision, recall, and F-score over word and character *n*-gram matches) and features based on Permutation Trees (Zhang and Gildea 2007) to account for differences in word order.

DPMFComb. DPMF (Yu et al. 2015) is a syntax-based metric that parses the reference translation with a standard parser and trains a new parser on the tree of the reference translation. This new parser is then used for scoring the MT output. DPMF uses an F-score of unigrams in combination with the syntactic score. DPMF performs quite poorly as an individual metric. To boost performance DPMFComb (Yu et al. 2015) combines DPMF and the lexical, syntactic, and semantic metrics from the Asiya evaluation toolkit (Giménez and Màrquez 2010a) in a learning framework.

Cobalt-F-comp and Metrics-F. Cobalt-F-comb and Metrics-F (Fomicheva et al. 2016) combine features extracted from UPF-Cobalt with reference-free features that capture translation fluency. Cobalt-F-comb combines various components of UPF-Cobalt with a series of fine-grained features intended to capture the number and scale of disfluent fragments contained in the MT outputs. Metrics-F is a combination of three evaluation metrics, BLEU, Meteor and UPF-Cobalt, with the fluency-oriented features.

UoW-ReVal. UoW-ReVal (Gupta, Orasan, and van Genabith 2015) uses a dependency-tree Long Short-Term Memory (LSTM) network to represent both the MT output and the reference with a dense vector. The segment level scores are obtained from a neural network that takes into account both the distance and the Hadamard product of the two representations. Training is performed on WMT ranking judgments converted to similarity scores.

QualityEstimation. Different from reference-based evaluation metrics, Quality Estimation (QE) metrics (Blatz et al. 2004; Specia et al. 2009) aim to predict the quality of a machine translated segment solely from information about the segment itself and its corresponding source segment (and optionally) information about the MT system that produced the translation. The problem is framed as a supervised machine learning task, where, given source-MT pairs annotated with a quality label, a number of features can be extracted and used to train a machine learning algorithm. Sentence-level QE has been covered as a shared task in the last six editions of the Conference on Machine Translation (WMT) (see Bojar et al. (2017) for the state-of-the-art approaches and latest results). In our experiments we use the best performing system from the WMT17 QE estimation task, the POSTECH system (Kim, Lee, and Na 2017). This is a neural prediction system that relies on two components (each based on a bidirectional long short-term memory unit): a *predictor*, which extracts in unsupervised ways "quality" vector representations from good examples of translations (i.e., large parallel corpora of human translations) and an *estimator*, which use the quality vectors and human quality labels to build prediction models.

4. Data Sets

This section presents the various data sets used in our meta-evaluation study. Each of them contains a number of original and reference sentences, as well MT outputs provided by one or various MT systems. They also contain manual assessments provided based on different quality criteria (adequacy, fluency, post-editing effort) collected using several different methods (interval-level scale, continuous scale, pairwise preference). These particular data sets were selected with the following criteria in mind: wide use in the community, availability of automatic evaluation results for the most recent evaluation metrics, availability of neural MT, and variability in the type of human judgments. With the exception of the MTSummit17 English-Latvian data set (see Section 4.5), all of the selected data sets contain into-English translations, as many evaluation metrics are available only for English.

4.1 WMT16 Direct Assessment Data Set

This data set (Bojar et al. 2016c) consists of source sentences, reference translations, and the outputs of the MT systems that participated in the WMT16 News Translation Task (Bojar et al. 2016a). Human quality judgments were collected according to the adequacy criterion following the **Direct Assessment** (DA) procedure described in Graham, Mathur, and Baldwin (2015) for all available into-English language pairs. More specifically, human assessors were asked how much of the meaning of the reference translation was preserved in the MT output. The evaluation was performed using a 0–100 rating scale. Raw human scores were converted into z-scores, that is, standardized according to an individual annotator's overall mean and standard deviation. Up to 15 assessments were collected for each MT output from different assessors and the results were averaged to obtain the final score. A total of 560 MT segments sampled randomly from the data were annotated by humans for each language pair, resulting in a total of 3,360 segments of into-English translations. The distribution of the normalized and averaged DA scores is shown in Figure 1.[4]

4.2 WMT16 Ranking Data Set

This data set (Bojar et al. 2016c) includes source texts, human reference translations, and the outputs from the MT systems participating in the WMT16 News Translation Task, for into-English and out-of-English translation for six languages (Czech, German, Finnish, Romanian, Russian, and Turkish). For manual evaluation, annotators were presented with the source sentence, its human translation, and the outputs of different MT systems and asked to rank the MT outputs from best to worst. Annotations were collected from volunteers from the participating research teams. For efficiency reasons, annotators were asked to compare the outputs of five MT systems (randomly sampled from the data set) for each sentence at once and rank them from best to worst. From this compact annotation, 10 pairwise ranking judgments can be extracted for each sentence in a straightforward way. For example, if a judge ranked the outputs of the systems A, B, C, D, E as A $>$ B $>$ C $>$ D $>$ E, then A $>$ B, A $>$ C, A $>$ D, A $>$ E, and so forth. It should be

4 Because automatic evaluation metrics compute the scores based exclusively on the MT output and the reference translation, for our experiments we combine all available assessments for into-English translations, ignoring the source language in order to have more data for the analysis.

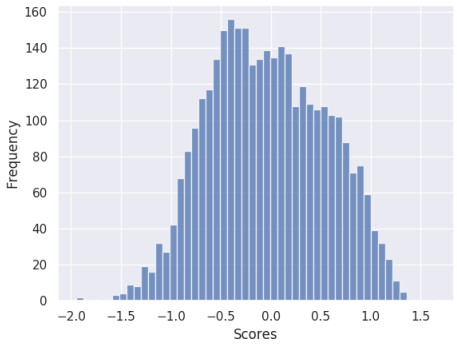


Figure 1
Distribution of adequacy scores in the WMT16 Direct Assessment data set.

Table 1
Number of pairwise preference judgments for the WMT16 data set.

Language pair	# judgments
Czech–English	70,000
German–English	15,000
Finnish–English	19,000
Romanian–English	11,000
Russian–English	18,000
Turkish–English	7,000

noted that neither the absolute value of the ranking, nor the degree of the difference, is taken into consideration. Table 1 shows the number of pairwise judgments per language pair for the into-English part of this data set that we use in our experiments.

4.3 Multiple-Translation Chinese Data Set

Multiple-Translation Chinese Data set (MTC) is a Chinese–English data set (LDC2006T04) that contains 919 source sentences from the news domain, 4 reference translations,[5] and MT outputs generated by 10 translation systems. Human judgments were collected for two criteria, both on a 5-point scale: adequacy and fluency, based on the following questions:

How much of the meaning expressed in the gold-standard translation is also expressed in the target translation?

- 5 = All
- 4 = Most
- 3 = Much

5 In our experiments we randomly selected one of the four human references to be used for evaluation, as many of the evaluation metrics considered are not equipped to be used with multiple references.

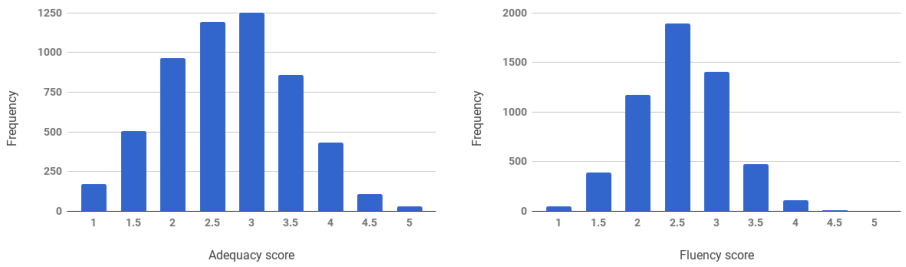


Figure 2
Distribution of adequacy and fluency scores in the MTC data set.

- 2 = Little
- 1 = None

How do you judge the fluency of this translation? It is:

- 5 = Flawless English
- 4 = Good English
- 3 = Non-native English
- 2 = Disfluent English
- 1 = Incomprehensible

In both cases, the assessments were provided by two annotators. We use their average as the final score. The distribution of adequacy and fluency scores is shown in Figure 2.

4.4 WMT17 Quality Estimation German–English Data Set

This data set belongs to the pharmaceutical domain and provides translations from German into English. It was originally used for the WMT17 shared task on QE (Bojar et al. 2017). Automatic translations were generated with a statistical phrase-based MT system (corpus-level BLEU score = 0.534) and post-edited by professional translators. After post-editing each segment, translators rated them using a 1- to 4-point scale, according to the effort they needed to fix the translation. More specifically, the following question was asked of the translators: *How good was the machine translation?*, with the following possible answers:

- 1 = Perfect or near perfect (typographical errors only)
- 2 = Very good, could be post-edited quickly
- 3 = Poor, required significant post-editing
- 4 = Very poor, required retranslation

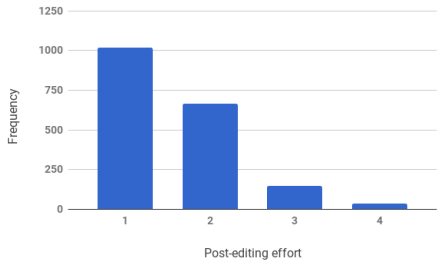


Figure 3
Distribution of post-editing effort scores in the WMT17 Quality Estimation data set.

For our experiments we only used the test partition of the corpus, which has 2,000 sentences, since we needed quality predictions in addition to human labels. Figure 3 shows the distribution of the human scores for the test partition of the data set. As we can see, most sentences are judged as near-perfect translations, which makes prediction particularly challenging for quality estimation metrics.

4.5 MTSummit17 English–Latvian Data Set

This data set (Specia et al. 2017) contains segments in the IT domain, from English into Latvian. Two models trained on exactly the same parallel data were built using statistical (corpus-level BLEU score = 0.465) and neural (corpus-level BLEU score = 0.384) MT approaches. Translations for the same 20,738 source segments produced by both neural and statistical MT (in total, 41,476 segments) were post-edited by professional translators and the quality of the raw MT assessed using the same 4-point scale scheme as for the WMT17 Quality Estimation German–English data set. The distribution of the scores for statistical and neural MT systems is shown in Figure 4. The average human scores are 1.64 for the statistical MT system and 1.84 for the neural MT system.

This data set also has a sample of 2,000 segments for each MT system type annotated for errors at the word level. More specifically, a subset of sentences scored as 2 (very good) were annotated such that all issues resolved during the PE phase were classified using the Multidimensional Quality Metrics (MQM) error annotation framework (Lommel, Burchardt, and Uszkoreit 2014). The list of errors is divided into the main

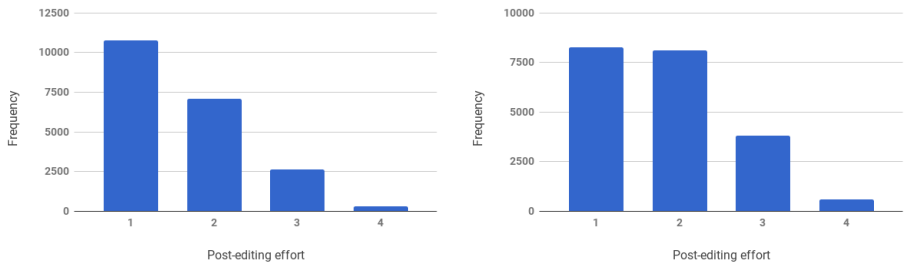


Figure 4
Distribution of post-editing effort scores for statistical and neural MT systems, respectively, in the MTSummit17 data set.

Table 2
Percentage of translation errors in the output of statistical and neural MT (NMT) systems, respectively, for the MTSummit17 data set.

	SMT		NMT	
	#	%	#	%
Fluency	156	8.7	146	8.1
Grammar	6	0.3	0	0.0
Function words	0	0.0	0	0.0
Extraneous	21	1.2	23	1.3
Incorrect	25	1.4	24	1.3
Missing	31	1.7	12	0.7
Word form	128	7.1	134	7.4
Tense/aspect/mood	21	1.2	23	1.3
Part of speech	16	0.9	14	0.8
Agreement	159	8.8	125	6.9
Word order	180	10.0	73	4.1
Spelling	105	5.8	133	7.4
Typography	343	19.1	179	9.9
Unintelligible	5	0.3	7	0.4
Accuracy	18	1.0	22	1.2
Addition	151	8.4	128	7.1
Mistranslation	150	8.3	382	21.2
Omission	221	12.3	327	18.2
Untranslated	46	2.6	29	1.6
Terminology	17	0.9	19	1.1
Total: Fluency	1,196	66.4	893	49.6
Total: Accuracy	586	32.6	888	49.3

issue categories *accuracy*, *fluency*, and *terminology*, which fold into a selection of more detailed categories from the MQM hierarchy. The set of all 20 error categories used in the annotation are shown in Table 2. Annotators were instructed to use the subcategories whenever possible and to resort to the more general category level only in case of doubt. We note that very often annotators backed off to the most general category in this particular data set.

4.6 GALE Arabic–English Data Set

Three Arabic newswire data sets produced as part of the DARPA GALE project are used: MT08, GALE09, and GALE10, containing 813, 683, and 1,089 sentences, respectively. Each data set was translated into English by two in-domain phrase-based SMT systems, system 1 and system 2, and annotated for adequacy in previous work for quality estimation (Specia et al. 2011).

Translation adequacy annotations were provided by two Arabic–English professional translators, who judged the translations along with the source sentences. Each translation was annotated once (for each translation, one translator was randomly selected). A 4-point scale was used to answer the question *To which degree does the translation convey the meaning of the original text*?":

- 4 = Highly adequate
- 3 = Fairly adequate

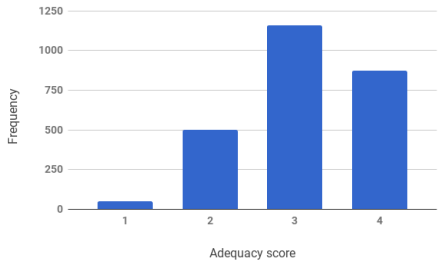


Figure 5
Distribution of adequacy scores in the GALE data set.

- 2 = Poorly adequate
- 1 = Completely inadequate.

For the purposes of this work, we combined the three data sets together and used the MT outputs from system 1. The distribution of human scores is shown in Figure 5.

4.7 EAMT11 French–English Data Set

The EAMT11 data set (Specia 2011) contains 2,525 French source sentences in the news domain and their translation into English produced by a phrase-based statistical MT system (corpus-level BLEU score = 0.25), as well as a human reference translation. The source segments come from the WMT news-test2009 data set (Callison-Burch et al. 2010). In addition, the EAMT11 data set contains the post-edition of the MT output generated by professional translators, as well as an absolute score on the post-editing effort required to fix the translations, as given by the human translator on a 4-point scale:

- 1 = Requires complete retranslation
- 2 = Requires some retranslation, but post-editing still quicker than retranslation
- 3 = Very little post-editing needed
- 4 = Fit for purpose

The distribution of the scores for this data set is shown in Figure 6.

5. Metrics across Translation Quality Levels

Correlation with human judgments is by far the most widely used measure for assessing the accuracy of automatic evaluation. Although it is a good indicator of the overall performance of evaluation metrics and allows us to compare different approaches, correlation alone provides no indication regarding the weaknesses of a given evaluation method. The use of correlation as the only measure for assessing the metrics performance has received some criticism, but alternative methods have hardly been discussed (see Section 2).

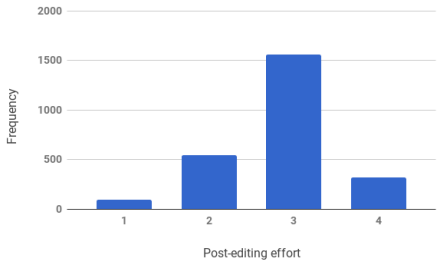


Figure 6
Distribution of post-editing effort scores for the EAMT11 data set.

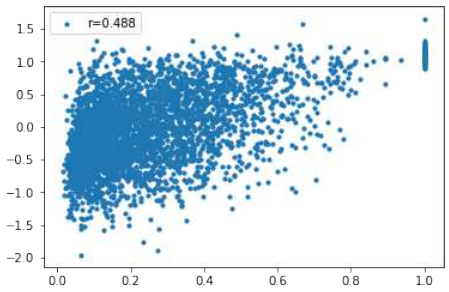
Automatic evaluation can go wrong in different ways. Given that there are multiple correct translations for the same source sentence, estimating quality based on the similarity to one reference translation may lead to penalizing perfectly acceptable MT outputs. Accurately evaluating an imperfect translation is also difficult, since translations that contain the same number of reference words can be incorrect to a varying extent (e.g., lack of agreement vs. omitting content-bearing words). In this section we investigate what is more challenging or causes problems more often: evaluating low-quality or high-quality translation. We will refer to this intuitively general distinction as **quality levels**, whereas more specific definitions can only be provided based on criteria established in a given evaluation setting.

In what follows we divide our analysis according to the general type of quality label used: continuous (Section 5.1) versus discrete (Section 5.2). We do so because different types of labels require different methods to assess the performance of metrics. Most of the analysis concentrates on continuous human scores, as these seem to be the most widely accepted nowadays. In addition, we examine human consistency on quality judgments for different levels of quality, as the lack of consistency can undermine the results of meta-evaluation (Section 5.3).

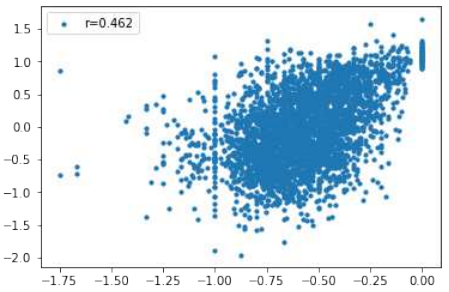
5.1 Continuous Human Scores

For continuous scores analysis, we use the WMT16 DA data set. Recall from Section 4.1 that this data set contains human assessments based on the adequacy criterion collected using a continuous 0–100 scale. The assessments of each annotator were converted to z-scores in order to avoid individual bias. Each MT output was evaluated by up to 15 annotators and the average of individual scores was used as the final segment-level score.

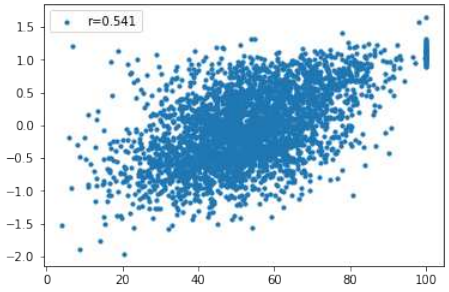
We start by examining the scatter plots for various evaluation metrics (Figure 7). Metric and human scores are plotted on the x-axis and y-axis, respectively. A good metric would be expected to have the points close to a straight line, indicating a high correlation with human scores. The plots for the different metrics look fairly similar, with the exception of BLEU. In contrast to the rest of the metrics, BLEU assigns very low scores to the majority of translations due to the sparseness of n-gram matches with the references. The advanced feature-based metrics that incorporate linguistic information, Metrics-F and DPMFcomb, have the data points closer to the diagonal and fewer outliers than lexical matching-based approaches. Another difference between these metrics and the ones based on lexical matching is that the MT outputs exactly



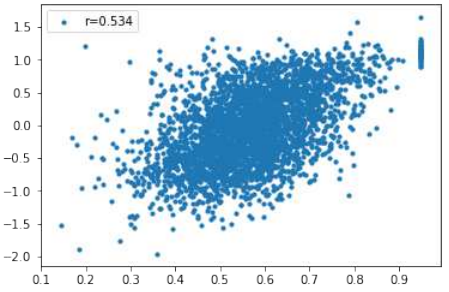
(a) BLEU



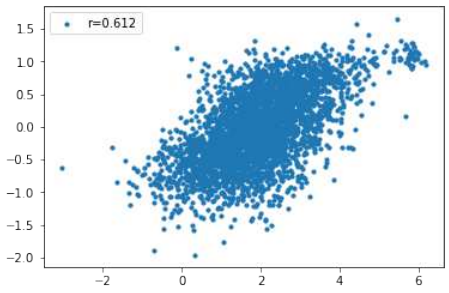
(b) -TER



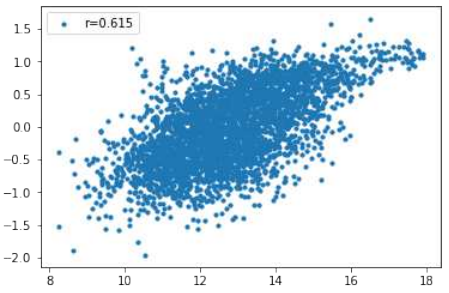
(c) ChrF3



(d) BEER



(e) Metrics-F



(f) DPMFcomb

Figure 7
Scatter plots illustrating the correlation (*r*) between metric and human scores for the WMT16 DA data set. Metric and human scores are plotted on the *x*-axis and *y*-axis, respectively.

matching the reference is less clearly separated from the rest of the data.[6] Thus, a simple visual inspection of the scatter plots already provides some insights into the metrics' behavior. It does not suffice, however, to describe the performance of evaluation metrics in relation to MT quality.

If the accuracy of automatic evaluation indeed changes depending on the level of translation quality, the strength of the relationship between metric and human scores should be different in the subsets of data corresponding to different (human) quality levels. Computing the correlation coefficient for two random variables conditioned on

6 The spread of human direct assessment scores for the MT outputs that have maximum automatic evaluation score (i.e., exactly match the reference) is due to the fact that direct assessment scores were obtained by averaging *z*-scores from individual annotators (Section 4.1).

the realizations of one of them is a well-known problem in the domain of finance, where it needs to be handled in order to study financial contagion (i.e., to determine whether financial markets become more interdependent during financial crises [Longin and Solnik 2001; Hong, Tu, and Zhou 2006]). A naive approach would consist of measuring the ordinary Pearson correlation coefficient in various sub-samples of data. This method is referred to as conditional correlation or "correlation breakdown" (Bertero and Mayer 1990; Baig and Goldfajn 1999). It has been shown, however, that tests for changes in correlation that do not take into account conditional heteroskedasticity (the fact that the variability of a random variable can be unequal across the range of values) may be severely biased (Forbes and Rigobon 2002). Although aware of its limitations, we first look at this naive approach and then explore a more complex strategy that avoids such limitations.

Conditional Correlation. Formally, as defined in Tjøstheim and Hufthammer (2013), given two variables X_1 and X_2 with observed values (X_{1i}, X_{2i}), $i = 1, ..., n$ the correlation between X_1 and X_2 conditional on being in a region of values A is given by:

$$\hat{p}_c(A) = \frac{\sum\limits_{(X_{1i},X_{2i})\in A} (X_{1i} - \hat{\mu}_{X_1,c})(X_{2i} - \hat{\mu}_{X_2,c})}{\left(\sum\limits_{(X_{1i},X_{2i})\in A} (X_{1i} - \hat{\mu}_{X_1,c})^2\right)^{1/2} \left(\sum\limits_{(X_{1i},X_{2i})\in A} (X_{2i} - \hat{\mu}_{X_2,c})^2\right)^{1/2}} \quad (1)$$

where $\hat{\mu}_{X_1,c} = \frac{1}{n_A} \sum\limits_{(X_{1i},X_{2i})\in A} X_{1i}$ and $\hat{\mu}_{X_2,c} = \frac{1}{n_A} \sum\limits_{(X_{1i},X_{2i})\in A} X_{2i}$, with n_A being the number of pairs with $(X_{1i}, X_{2i}) \in A$.

In the context of MT evaluation, the variables X_1 and X_2 correspond to metric and human scores, whereas A refers to the regions of values with different levels of translation quality. The regions of values representing different quality levels can be defined in various ways. If human scores are provided on a continuous scale one can either use the absolute value of the scores to split the data, or use quantiles as cutoff points. For the WMT16 DA data set the first option would imply dividing the data based on the scores from the 0–100 scale that was used for collecting human judgments. However, as discussed in Graham, Mathur, and Baldwin (2015), standardized human scores are more reliable as the gold standard for assessing metric performance, as they neutralize the bias of individual annotators. Therefore, we use z-scores for the analysis presented in this section and split the data based on quantiles as cutoff points. In addition, splitting the data based on quantiles allows us to have the same number of data points in each quality band and thus facilitates the comparison of the behavior of the metrics. [7]

Furthermore, the data can be split with different levels of granularity. For instance, a two-way split can be done resulting in two levels that would correspond to low- and high-quality translation. To check how accurately the metrics can assess translations in the medium quality range in addition to the extremes, the data can be further split into more levels.

Tables 3 and 4 show the conditional correlation between metric scores and human judgments for two granularity levels: two quantiles and four quantiles. We start with a binary split (Table 3) using the median of the human scores as the cutoff point, resulting

7 Note, however, that quality levels defined in this way should be interpreted in relative terms, namely, lower quality vs. higher quality for this data set.

Table 3
Conditional Pearson correlation with direct assessment scores for popular and top scoring metrics from the WMT16 Metrics for high-quality and low-quality data partitions. † indicates that the correlation for high-quality data samples (Q_{high} and Q^*_{high}) is significantly different from the correlation for the low-quality data sample (Q_{low}). In each column, results for the metrics that are not significantly outperformed by any other metric are marked in **bold**.

	Q_{low}	Q_{high}	Q^*_{high}	*All*
Meteor	**0.313**	**0.514**†	0.420†	0.570
-TERp-A	0.265	0.459†	0.394†	0.570
MPEDA	**0.313**	0.512†	0.417†	0.568
ROUGE-SU*	0.274	0.453†	0.373†	0.551
ChrF3	**0.321**	0.425†	0.336	0.541
NIST-4	0.258	0.415†	0.327	0.508
BLEU-4	0.159	0.462†	0.360†	0.488
-TER	0.129	0.433†	0.358†	0.462
-WER	0.090	0.458†	0.387†	0.456
-PER	0.175	0.361†	0.281†	0.422
UPF-Cobalt	0.256	0.467†	0.394†	0.566
CP-Oc(*)	0.225	0.453†	0.359†	0.527
SP-lNIST	0.272	0.416†	0.328	0.512
DP-Oc(*)	0.112	0.395†	0.322†	0.424
SR-Or(*)	0.137	0.273†	0.244†	0.371
DPMFcomb	**0.314**	0.512†	0.438†	**0.615**
Metrics-F	0.271	**0.528**†	**0.447**†	**0.612**
Cobalt-F-comp	0.231	**0.530**†	**0.463**†	0.599
BEER	**0.315**	0.422†	0.328	0.534
UoW-ReVal	0.217	0.441†	0.375†	0.525

in two samples: top 50% and bottom 50% of the data, each containing 1,680 sentences and corresponding to "lower" (Q_{low}) and "higher" (Q_{high}) quality translations. In order to avoid obvious biases, we computed the correlation for the high-quality sample, eliminating the sentences where MT output exactly matches the reference, that is, the cases where the metric scores are guaranteed to be correct (column Q^*_{high} in Table 3). Column *All* shows the correlation results for the full data set.[8]

In Table 4 we test how accurately the metrics can assess translations in the middle of the quality range. We split the WMT16 DA data set using four quantiles based on human scores as the cut-points. This results in four samples: top 25%, mid-high 25%, mid-low 25%, and bottom 25% of the data (which we call Q_1-Q_4), each containing 840 sentences. As before, column Q^*_4 shows the correlation for the high-quality sample, eliminating the sentences where MT output exactly matches the reference, that is, the cases where the metric scores are guaranteed to be correct.

The metrics in Tables 3 and 4 are divided into three groups, as discussed in Section 3. The first group corresponds to the metrics based on lexical similarity. The second

8 The results presented here differ from the official WMT16 results, as the latter were computed separately per language pair (see Section 4.1).

Table 4
Conditional Pearson correlation with direct assessment scores for popular and top scoring metrics from the WMT16 Metrics Task for a four-way split of the data set resulting in data samples corresponding to four quality levels (Q_1–Q_4). † and ‡ indicate, for each column, if the correlation is significantly different from the correlation in Q_1 and Q_4, respectively. In each column, results for the metrics that are not significantly outperformed by any other metric are marked in **bold**.

	Q_1	Q_2	Q_3	Q_4	Q_4^*	*All*
Meteor	**0.198**‡	**0.151**‡	0.163‡	**0.514**†	**0.347**†	0.570
-TERp-A	0.168‡	0.113‡	**0.180**‡	0.404†	0.287†	0.570
MPEDA	**0.200**‡	**0.150**‡	0.166‡	0.512†	0.343†	0.568
ROUGE-SU*	**0.199**‡	0.118‡	**0.193**‡	0.398†	0.252	0.551
ChrF3	**0.218**‡	0.119†‡	0.139‡	0.375†	0.219	0.541
NIST-4	0.189‡	0.109‡	0.137‡	0.397†	0.246	0.508
BLEU-4	0.051‡	0.084‡	0.136‡	0.453†	0.282†	0.488
-TER	0.051‡	0.056‡	**0.172**†‡	0.388†	0.254†	0.462
-WER	0.031‡	0.048‡	**0.189**†‡	0.404†	0.276†	0.456
-PER	0.115‡	0.072‡	0.133‡	0.351†	0.212†	0.422
UPF-Cobalt	0.150‡	0.122‡	0.170‡	0.403†	0.275†	0.566
CP-Oc(*)	0.164‡	0.078‡	**0.172**‡	0.431†	0.270†	0.527
SP-lNIST	0.198‡	0.109‡	0.141‡	0.392†	0.241	0.512
DP-Oc(*)	0.055‡	0.072‡	0.153†‡	0.349†	0.224†	0.424
SR-Or(*)	0.083‡	0.085‡	0.062‡	0.215†	0.174	0.371
DPMFcomb	**0.204**‡	**0.146**‡	**0.193**‡	0.443†	0.303†	**0.615**
Metrics-F	0.127‡	**0.172**‡	**0.199**‡	0.480†	**0.327**†	**0.612**
Cobalt-F-comp	0.092‡	**0.160**‡	**0.216**†‡	0.469†	**0.344**†	0.599
BEER	**0.228**‡	0.119†‡	0.143‡	0.384†	0.218	0.534
UoW-ReVal	0.096‡	0.092‡	0.163‡	0.376†	0.257†	0.525

group includes the metrics that use different kinds of information about sentence structure (constituency parsing, dependency parsing, semantic parsing, and named entity recognition). Finally, the third group contains trained, feature-based metrics. For brevity, for each type of linguistic metrics from the Asiya toolkit discussed in Section 3 (SP-*, DP-*, CP-*, and SR*), we only present the metric that obtained the best overall correlation on this data set.

To compute the significance of the difference in correlation for different quality levels, we used Fisher's *z*-transformation. In Table 3, † indicates whether the correlation for high-quality data (Q_{high} and Q^*_{high}) is significantly different from the correlation for Q_{low}. In Table 4, † and ‡ indicate, for each column, whether the correlation is significantly different from the correlation in Q_1 and Q_4, respectively.[9] To compare the results of different metrics against each other, we used the Hotteling-Williams test for dependent correlations (Williams 1959). In each column, results for metrics that are not significantly outperformed by any other metric are marked in bold.

9 The correlation for Q_4^* was compared to Q_1 and not to Q_4.

We observe that the correlation of all evaluation metrics is substantially lower for low-quality MT output. The difference in correlation between the low- and high-quality samples is preserved even when the sentences where MT output exactly matches the reference are eliminated (Q^*_{high} and Q^*_4). One possible explanation for this is that low-quality translations contain a higher number and variety of translation errors. Determining consistently to what extent a particular type of error should be reflected in a translation quality measurement is difficult even for human annotators, and is nearly impossible for current similarity-based evaluation metrics that do not consider the type of translation errors explicitly. Consider the following two examples from the WMT16 DA data set.[10]

Example 1
Src: Ve Washington'a da kızgınlar.
MT: And Washington also angry.
Ref: And they are angry at Washington, too.

Example 2
Src: Bunun için sindirim sisteminizin sağlıklı çalışması gerekir.
MT: Sindirim sisteminizin healthy for it to work.
Ref: To do this, your digestive system should work healthily.

In Example 1 all the content words are translated, but the grammatical link between them is missing, making it difficult to understand the sentence. In Example 2 several source words are left untranslated, which makes the sentence completely incomprehensible. Human direct assessment scores for these two MT outputs are -0.35 and -0.78, respectively, indicating that the MT output from Example 2 is perceived as considerably lower quality.[11] For comparison, the corresponding BLEU scores are 0.07 and 0.05, failing to indicate any difference in quality, as the percentage of matching n-grams is similarly low in both examples.

Another reason for lower correlation in the low-quality partition of the data is that high-quality outputs tend to contain a higher number of matches with the reference, and thus evaluation metrics naturally have more information to measure translation quality. By contrast, low-quality MT outputs contain very few matches and thus metric scores simply indicate that the MT output is different from the available reference. Human judges, on the other hand, assign different scores to low-quality translations, depending on how bad they are.

For the high-quality sample, the best results are achieved by the metrics with the overall highest correlation, namely, the feature-based metrics from the third group. For the low-quality sample, however, the best metrics are: Meteor (MPEDA is a variant of Meteor), ChrF, and BEER. For the latter two the correlations for Q_{low} and Q^*_{high} in Table 3 are not significantly different. All these metrics have in common the fact that they are less affected by the sparseness of reference matches. Both ChrF and BEER use character n-grams, which is a more robust representation than word n-grams, whereas Meteor allows for stem, synonym, and paraphrase matches. Another common feature of these metrics is that they put a higher weight on content word matches. Meteor explicitly

10 Example 1: language pair – Turkish-English, MT system – online-B, segment – 143. Example 2: language pair – Turkish-English, MT system – jhu-syntax, segment – 2842.
11 See Section 4.1 for the definition of manual assessment scores in the WMT16 Direct Assessment data set.

assigns different weights to content vs. function words. In the case of character-based metrics this distinction stems from the fact that content words tend to be longer, thus containing a higher number of n-grams and therefore matches between such words increase the score to a larger extent than matches between stopwords. This feature may be particularly important for the evaluation of low-quality translation as it allows us to better discriminate between translation errors. The correlation for intermediate quality levels follows the pattern we observe for Q_1 and Q_4: Metrics perform better when dealing with higher- (Q_3) than lower-quality (Q_2) translations. On average, the correlation is higher for Q_1 and Q_4, showing that more nuanced quality distinctions (Q_2 and Q_3) are more difficult to capture.

So far we have looked at the conditional correlation approach for describing local dependency between two random variables. We computed the Pearson correlation between metric and human scores on various ranges of values defined based on the human scores. Although straightforward and simple to use, this approach has certain limitations (Tjøstheim and Hufthammer 2013). First, there is no evident formal way of defining the ranges of values. Second, the granularity of an analysis based on conditional correlation is limited by the amount of variation in human scores inside each range of values. If human scores inside each level do not indicate any meaningful differences in quality, the correlation between metric and human scores would not be informative. In general, restricting the range of values reduces the variation and, therefore, will restrict the correlation to less than would be observed in the full range of values. In this sense, note that the correlation for different quality levels (columns Q_1–Q_4) in Table 4 is always lower than the correlation for the full data set (column *All*). Finally, the relation between metric and human scores in local regions of values may be nonlinear, which would make the use of the Pearson correlation coefficient inappropriate.

Local Gaussian Correlation. To address the limitations of the conditional correlation discussed above, various strategies have been proposed in the literature on statistics (Doksum et al. 1994; Jones and Koch 2003; Delicado and Smrekar 2009). Here, we explore the local dependence measure recently designed by Tjøstheim and Hufthammer (2013) for measuring financial contagion, which they call local Gaussian correlation. The idea behind this measure is to fit a Gaussian bivariate density in a neighborhood of each data point using local likelihood. Thus, at each specific neighborhood, the local dependence properties will be described by the local covariance matrix, fully characterizing the dependence relationship in that neighborhood.

Formally, for a general bivariate density f for the variables (X_1, X_2), Tjøstheim and Hufthammer (2013) define the local Gaussian bivariate density in a neighborhood of each point $x = (x_1, x_2)$ as follows:[12]

$$\psi(v, \theta(x)) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-p^2}} \exp\left[-\frac{1}{2(1-p^2)} \left(\frac{(v_1-\mu_1)^2}{\sigma_1^2} - 2p\frac{(v_1-\mu_1)(v_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(v_2-\mu_2)^2}{\sigma_2^2} \right) \right] \quad (2)$$

where $v = (v_1, v_2)$ is the running variable of the Gaussian distribution *in the neighborhood of the point* x and $\theta(x)$ is the 5-dimensional vector $[\mu_1(x), \mu_2(x), \sigma_1^2(x), \sigma_2^2(x), p(x)]$ in which $\mu_i(x), i = 1, 2$ are the local means, $\sigma_i(x), i = 1, 2$ are the local standard deviations, and

12 In Equation (2) we drop the argument (x) for μ_i, σ_i, and p for simplicity.

$p(x)$ is the local correlation at the point x.[13] It is the last parameter, local correlation $p(x)$, that we ultimately desire to find to discover local dependence properties.

In order to fit $\psi(v, \theta(x))$ locally, the parameters $\theta(x)$ need to be such that $\psi(v, \theta(x))$ equals general density $f(x)$ at $v = x$ and is close to f in a neighborhood of x. To estimate $\theta(x)$, Tjøstheim and Hufthammer (2013) follow the local likelihood approach proposed by Hjort and Jones (1996). Specifically, they define the parameters $\theta(x)$ to be the minimizer of the following local penalty function:

$$q = \int K_b(v - x)[\psi(v, \theta(x)) - \log \psi(v, \theta(x)) f(v)]dv \tag{3}$$

$$K_b(v - x) = \frac{K(\frac{(v_1 - x_1)}{b_1})K(\frac{(v_2 - x_2)}{b_2})}{b_1 b_2} \tag{4}$$

where K_b is a product kernel function with bandwidth $b = (b_1, b_2)$.[14] As mentioned in Tjøstheim and Hufthammer (2013), the penalty function q can be interpreted as a locally weighted Kullback-Leibler distance from the general density f to its parametric local approximation $\psi(v, \theta(x))$.

This function is used to fit a Gaussian density in the neighborhood of each point x. Thus, the general bivariate density f is represented by a family of local densities and, because ψ is Gaussian by definition, the local dependence relationship in the neighborhood of each estimation point is fully described by $p(x)$. For more details regarding the mathematical formulation and the relevant theory, the reader is referred to Tjøstheim and Hufthammer (2013).

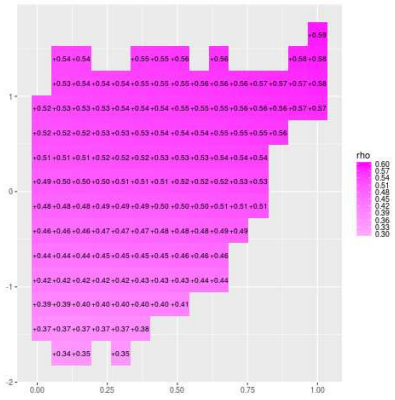
We use an existing R package **localgauss** (Berentsen et al. 2014) to compute and visualize the local Gaussian correlation between metric and human scores. Figure 8 displays the local Gaussian correlation plots for various evaluation metrics from the WMT16 data set. Similar to the scatter plots in Figure 7, metric and human scores are represented on the x and y axes, respectively. The tiles show the estimated local Gaussian correlation in the neighborhood of a set of estimation points.[15] The color scale is relative, with darker colors representing higher correlation values and lighter colors representing lower correlation values for a given plot. For all of the metrics, the plots clearly show a stronger correlation for higher-quality data. In accordance with the analysis presented earlier in this section, the difference appears less pronounced for the character-based metrics, ChrF and BEER.

The ordinary correlation coefficient does not capture the subtleties of the relation between quality scores predicted by the automatic evaluation metrics and actual translation quality as reflected in manual quality assessment. We hypothesized that the strength of this relation in fact changes with the position on the translation quality scale, as the performance of evaluation metrics is affected by the level of MT quality. We have looked into various ways to analyze the relation between metric and human scores for

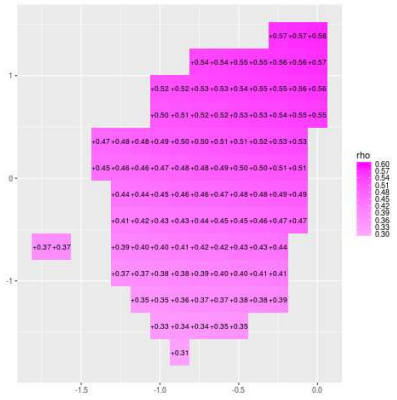
13 Note that the definition in Equation (2) is analogous to the general definition of bivariate Gaussian density (see, for example, Tong 1990), but the observations are sampled from the neighborhood of a chosen data point.

14 Following Støve, Tjøstheim, and Hufthammer (2014), for the choice of bandwidth we use a simple rule of thumb—the global standard deviation times a constant close to one.

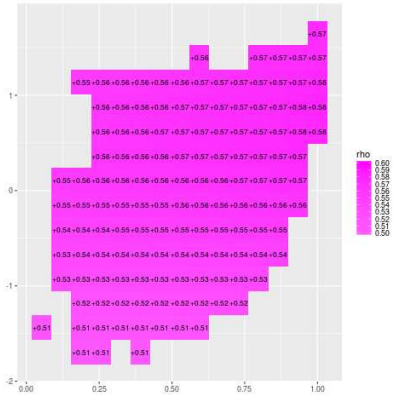
15 The **localgauss** visualization package allows us to select the estimation points manually, or set them automatically using the method described in Berentsen et al. (2014). We followed the latter option.



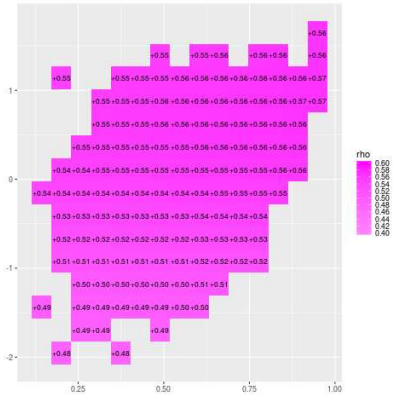
(a) BLEU



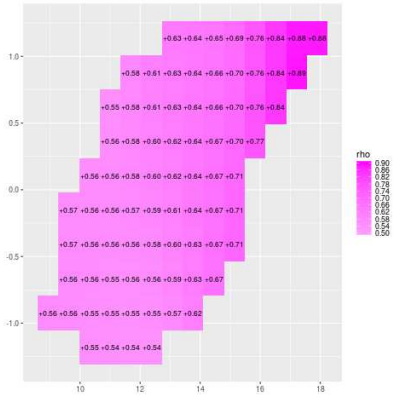
(b) -TER



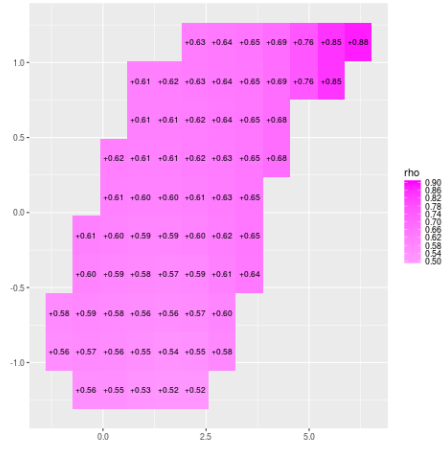
(c) ChrF3



(d) BEER



(e) DPMFcomb



(f) Metrics-F

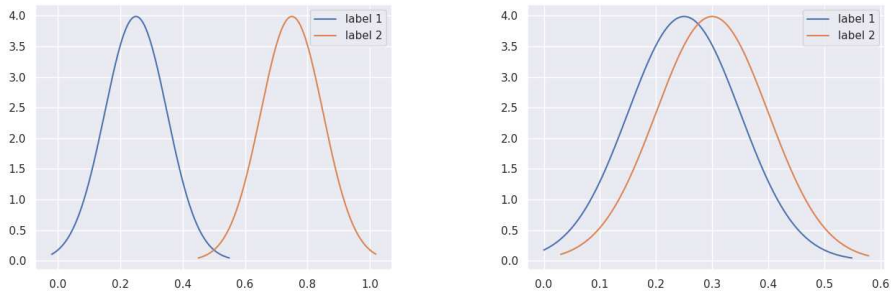
Figure 8
Local Gaussian correlation for evaluation metrics in the WMT16 DA data set. x and y axes correspond to metric and human scores. The tiles show the estimated local Gaussian correlation in the neighborhood of a set of estimation points.

the case of continuous variables, including a naive conditional correlation approach that amounts to computing ordinary Pearson correlation on different ranges of values, and a more complex method that models local dependence between the variables by fitting a Gaussian density function at each data point. Both methods reveal the fact that correlation with human judgments is lower for lower-quality translation. Subsequently, we investigate the relation between quality levels and the performance of automatic evaluation metrics for the scenario where human scores are provided on an interval-level scale.

5.2 Discrete Human Scores

The methods discussed earlier only apply when the manual evaluation data can be split into various slices of values with some variation within each slice. This was the case with the direct assessment approach to manual evaluation used at the WMT Conference in the last few years where manual assessment is collected on a continuous scale, as was discussed in the previous section. A different but not less common manual evaluation scenario involves collecting human assessments using discrete quality labels, for example 1–4 or 1–5 adequacy or fluency scales (see Section 2). In such a setting, quality levels can be defined in a straightforward way based on the discrete manual evaluation scores. However, the techniques discussed in the previous section cannot be applied, as a certain amount of variation inside each level would be required in order to measure the correlation between metric and human scores. We suggest that in order to describe the difference in metric performance in relation to MT quality for data sets with discrete human scores, the distribution of metric scores corresponding to each quality level can be examined. A good evaluation metric would have non-overlapping distributions of scores with equally distant means for the outputs assigned the same quality label in manual evaluation. Conversely, an unsuccessful metric would have overlapping distributions for the MT outputs assigned different scores by humans. For illustration, see the plots in Figure 9.

In this section we use the WMT17 Quality Estimation data set, which contains human assessments in the form of discrete scores in the range [1, 4] (from best to worst), indicating the effort required for post-editing the MT outputs (see Section 4.4 for a



(a) Metric A (b) Metric B

Figure 9
A hypothetical Metric A discriminates well between translations assigned different labels by human judges, whereas Metric B generates similar scores for MT outputs belonging to different quality levels.

Table 5
Overall absolute Pearson correlation of metrics with human judgments at sentence level in the WMT17 Quality Estimation data set.

Metric	Pearson *r*
ROUGE-SU*	0.379
TERp-A	0.370
Meteor	0.358
ChrF3	0.351
BLEU-4	0.351
NIST-4	0.335
WER	0.272
TER	0.260
PER	0.226
UPF-Cobalt	0.384
DP-Oc(*)	0.365
CP-Oc(*)	0.362
SP-lNIST	0.334
SR-Or(*)	0.259
BEER	0.330
POSTECH	0.502
HTER	0.638

detailed description of the data set). We define the quality levels accordingly as Q_1–Q_4, Q_1, indicating the highest translation quality in this case. Table 5 shows the overall Pearson correlation results for various evaluation metrics.[16] As before, the metrics (described in Section 3) are divided into three groups. The first group corresponds to the metrics based on lexical similarity, the second group includes the metrics that use different kinds of information on sentence structure, and the third group contains feature-based metrics.[17]

The third group of metrics in Table 5 also includes the results from the POSTECH system (Kim, Lee, and Na 2017), the best performing QE system that participated in the WMT17 QE task (see Section 3). As discussed in Section 3, QE (Blatz et al. 2004; Specia et al. 2009) is a different approach to MT evaluation that aims to predict the quality of a machine translated segment in the absence of a gold standard human translation. The task is typically addressed in a supervised machine learning framework, where given source sentences and the corresponding MT outputs annotated with some quality labels, a number of features can be extracted and used to train a machine learning

16 Recall that Pearson correlation coefficient values range from −1 to 1. If both variables tend to increase or decrease together, the coefficient is positive. If one variable tends to increase as the other decreases, the coefficient is negative. The stronger the association between the two variables, the closer the Pearson correlation coefficient will be to either −1 or 1. For for WMT17 Quality Estimation data set (Section 4.4) and for MTSummit17 data set (Section 4.5), manual assessments were collected on a 4-point scale from best to worst. Thus, lower human scores indicate *higher* quality, which results in negative correlations with evaluation metrics. To avoid confusion, for these data sets we report absolute correlation values in Tables 5, 9, and 10.

17 The metric set is different from the one presented in the previous section, as some of the metrics from WMT16 Metrics Task are not publicly available.

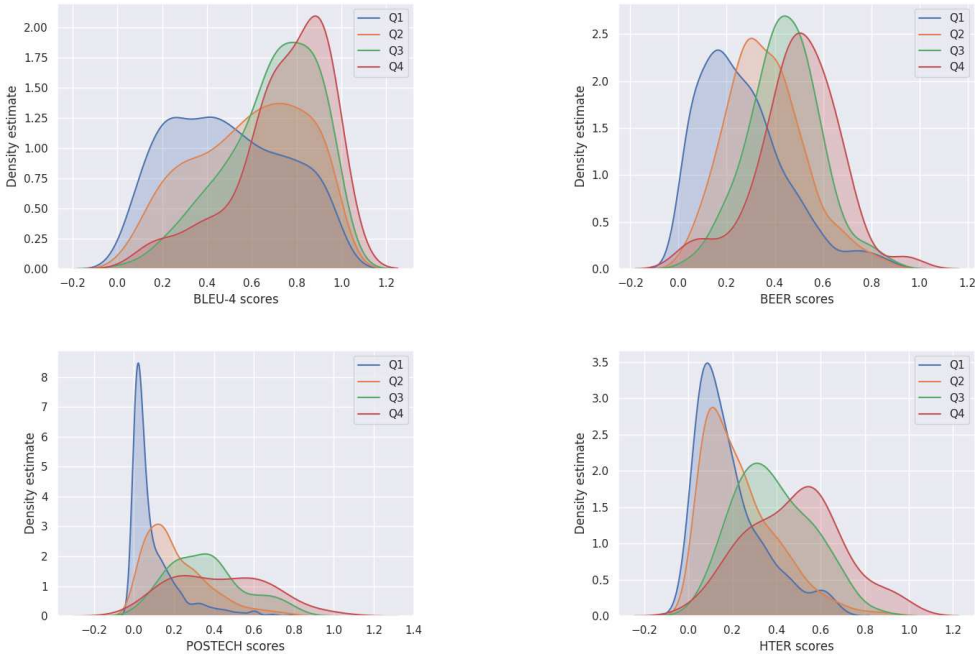


Figure 10
Kernel density estimation plots for the scores generated by BLEU, BEER, POSTECH QE system, and HTER scores in the WMT17 QE data set. For BLEU and BEER the scores are inverted (i.e., $score = 1 - score$), so that they are comparable with QE system scores and discrete human scores, where the lower the score the higher the quality.

algorithm to predict such labels for unseen data. The QE systems participating in the WMT17 QE task were trained and evaluated using the HTER scores[18] as quality labels. As an upper bound for the performance of the QE systems, the last row in Table 5 shows the correlation between true HTER scores and human scores. The results in each group are ordered from best to worst.

Overall, the highest correlation for this data set is obtained by the QE system. Judging from the correlation alone it is very difficult to know what the advantages are of QE systems in this particular setting. Figure 10 shows the kernel density estimation plots for BLEU, BEER, and POSTECH generated using the metric scores for the MT outputs assigned to each of the four different quality levels (Q_1–Q_4).[19] The plots for the HTER scores are also provided for comparison. As mentioned before, the quality levels here correspond directly to the discrete human scores provided on the scale 1–4 from best to worst. As illustrated in Figure 9, the plot for an ideal evaluation metric would have non-overlapping curves for the distribution of metric scores corresponding to the different human quality levels. In Figure 10, however, a considerable overlap can be observed. This points toward the limitations of the metrics and potentially of manual evaluation. On the one hand, some overlap may be due to the noise in manual

18 HTER (human-targeted edit rate) is a widely used measure of post-editing effort obtained by computing the TER metric between the MT output and its post-edited version (Snover et al. 2006).

19 For BLEU and BEER scores we use $score = 1 - score$ so that they are comparable with QE system scores and discrete human scores, where the lower the score, the higher the quality.

Table 6
Average overlap between the estimated distributions of the metric scores for each quality level in the WMT17 Quality Estimation data set.

Metric	Q_1	Q_2	Q_3	Q_4
ROUGE-SU*	0.533	0.685	0.686	0.645
-TERp-A	0.526	0.661	0.660	0.575
Meteor	0.561	0.704	0.698	0.605
ChrF3	0.562	0.705	0.705	0.621
BLEU-4	0.580	0.728	0.715	0.670
NIST-4	0.594	0.716	0.696	0.627
-WER	0.561	0.725	0.735	0.689
-TER	0.560	0.723	0.731	0.686
-PER	0.570	0.698	0.713	0.629
UPF-Cobalt	0.533	0.686	0.683	0.595
DP-Oc(*)	0.546	0.719	0.702	0.689
CP-Oc(*)	0.557	0.697	0.684	0.610
SP-lNIST	0.605	0.722	0.702	0.641
BEER	0.582	0.712	0.709	0.636
POSTECH	0.405	0.586	0.586	0.560
HTER	0.251	0.503	0.529	0.490

assessments. General quality is continuous and it is difficult to establish clear-cut limits when classifying MT outputs into a few quality categories, particularly given that the definition of such categories is quite vague. On the other hand, the metrics are not able to reliably discriminate between outputs with different quality.

More specifically, in the case of BLEU there is a substantial overlap between the scores of the sentences belonging to all the four quality levels. The form of the curve for the high-quality translations (Q_1) resembles a uniform distribution highlighting a well-known problem of BLEU metric when used for segment-level evaluation, which consists in harshly penalizing any kind of differences between MT output and the reference. Translations belonging to different levels are much better separated by BEER, which shows more robust performance with high-quality outputs and also a better separation of low-quality translations (Q_3 and Q_4).

The behavior of the QE system is quite different from that of the metrics. Unlike the metrics, which use reference translation to generate the scores, QE is based on reference-independent features. The system was trained using HTER scores and behaves in a similar way. The plots clearly show that for this data set the QE system outperforms the other metrics because it is able to better identify high-quality translations. Although the system does not seem to distinguish among low-quality translations very well, in this data set the number of translations assigned low scores is very small (see Section 4.4) and, therefore, this has a small impact on the overall correlation.

A possible way to quantify the discriminative power of the metrics illustrated by the plots in Figure 10 for different quality levels is to measure the overlap between the distribution of metric scores for each of them. The overlap between two density functions can be defined as the integral of $min(f(x), g(x))$, where f and g are the estimated density functions. For each quality level, we computed the average overlap with the other levels for the corresponding distributions of the metric scores. The results are shown in Table 6. The higher the values, the worse the metrics' ability to distinguish

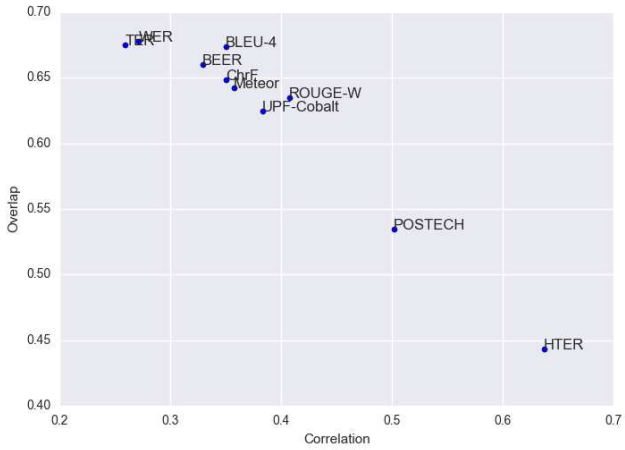


Figure 11
Comparison between the overlap of density functions of the metric scores and the overall Pearson correlation between the metric and human scores in the WMT17 Quality Estimation data set.

between translations belonging to different levels. The overlap tends to be higher for the intermediate levels. Furthermore, the overlap is always higher for Q_4 (lowest MT quality) than for Q_1 (highest quality), which agrees with the observations presented in the previous section. Figure 11 shows the relation between this overlap measure and the overall correlation of the metric and human scores computed using all available data. As expected, the lowest overlap corresponds to the highest correlation coefficient.

So far we have looked at the difference in the accuracy of automatic evaluation for different levels of translation quality using various instruments and observed a consistent degradation in metric performance for lower-quality translation. We have suggested that aside from the limitations of the metrics, such degradation could be due to a possible decrease in the reliability of manual evaluation. We test this hypothesis in Section 5.3. Finally, we have seen the impact of the distribution of the scores in the data set on the overall correlation results. We return to this issue in Section 7.

5.3 Quality Levels and Evaluation Consistency

Besides the inherent limitations of the evaluation metrics, a possible reason for lower correlation for low-quality MT outputs is a lack of consistency in manual evaluation. Manual evaluation is typically treated as the gold standard for assessing the performance of automatic evaluation metrics. However, MT quality assessment is known to be a complex task with low levels of agreement between annotators (Graham et al. 2013). One could hypothesize that evaluating poor quality translations is more demanding and harder for human annotators. They contain a higher number of errors of different types whose impact on quality can be difficult to determine (Denkowski and Lavie 2010).

We test this hypothesis using the WMT16 DA data set (Section 4.1). The annotations in this data set were collected using Amazon Mechanical Turk, which often raises questions about the reliability of the data. However, as described in Graham, Mathur, and Baldwin (2015), a rigorous quality control was conducted in order to filter out

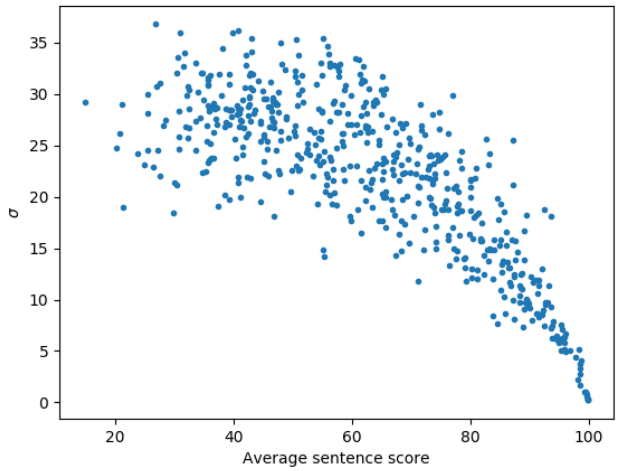


Figure 12
Average quality score and standard deviation of the scores assigned to the same sentences by different annotators.

unreliable workers. For the analysis presented in this section, only the data from the workers who passed quality control was used. The WMT16 Metrics Tasks followed different procedures for collecting evaluation data for segment-level and system-level tasks. For segment-level analysis, up to 15 assessments from different judges were collected per MT output before combining them into a mean adequacy score in order to increase the reliability of the evaluation. For system-level analysis, up to two assessments were collected per MT output, but this was done for all the systems participating in the WMT16 News Translation Task. We take advantage of both types of data in the following analysis.

We start with a very simple analysis of the segment-level data. First, we included sentences with the average score lower/higher than 50 in low- and high-quality partitions, respectively. Next, we calculated the average standard deviation of the scores assigned to each sentence in the low- and high-quality partitions.[20] The resulting values are 27.82 and 21.07, respectively. The difference between the variances in the low-quality sample and in the high-quality sample was found to be significant with $p < .05$, according to the Levene test, which tests the null hypothesis that the population variances are equal (Levene 1961). The variability in sentence scores provided by different annotators reflects the uncertainty involved in the evaluation process. Higher variability indicates that the sentence is more difficult to assess. This is the case for lower-quality translations. As an illustration, consider the plot in Figure 12. The x axis represents the means of the scores assigned by different annotators to the same sentence, and the y axis shows the standard deviation for each sentence. At the higher end of the quality scale the variability between the scores tends to be smaller.

20 For automatic evaluation analysis in Section 5.1 we used standardized segment-level scores, as they neutralize the differences in scoring strategies of individual annotators and constitute a more reliable gold standard for assessing metric performance. Here the goal is to assess the consistency of human assessments themselves and, therefore, raw human scores are used.

Table 7
Average score difference for the WMT16 Direct Assessment data set for a given judge and across two different judges.

	Inter-AA	Intra-AA
Q_1	33.721	17.721
Q_2	23.541	11.936
Q_2^*	23.739	12.087
All	25.279	13.450

We further compare the levels of inter- and intra-annotator agreement for lower and higher-quality translations. The kappa coefficient commonly used to calculate the consistency of human judgments in the context of MT evaluation (Callison-Burch et al. 2007) is not suitable for a continuous measurement scale. Instead, we used the method described in Graham et al. (2013) to compare evaluation consistency.

Specifically, using system-level data from the WMT16 Metrics Task, we computed the average difference between the scores assigned to the same MT output by different judges and by the same judge. This was done separately for the segments, with the average score higher than or equal to 50 and with the average score lower than 50 (thus corresponding to high- and low-quality partitions, respectively). We also computed the average difference for the high-quality partition, excluding the cases where MT output exactly matches the reference translation.

Results are reported in Table 7. Q_1 corresponds to the low-quality partition, Q_2 corresponds to the high-quality partition, and Q_2^* corresponds to the high-quality data, excluding the MT outputs exactly matching the reference. "Inter-AA" and "Intra-AA" refer to inter- and intra-annotator agreement, respectively. For both intra- and inter-annotator scenarios, the difference between Q_1 and Q_2/Q_2^* samples was found to be statistically significant.[21] Thus, we can see that the annotators are indeed less consistent when they assign lower scores. As mentioned before, a probable reason for that is the fact that low-quality MT outputs contain a higher number of errors, and the perceived impact of different translation errors on the overall translation quality can vary greatly, depending on individual annotators' ideas about the purpose of translation, translation priorities, and so on.

In this section, we have seen that meta-evaluation based on Pearson correlation alone can hide very different behaviors of evaluation metrics. The main outcome of the analysis presented here is that discriminating between lower-quality translations appears to be more challenging in both automatic and manual evaluation scenarios. Finding meaningful distinctions between low-quality MT outputs is difficult for reference-based metrics, as there is less information available in terms of the relation with the reference translation. Furthermore, low-quality translations contain a higher number of difficult-to-compare errors, which makes evaluation difficult even for human annotators. In the face of such an outcome, we suggest that a possible way for further development of evaluation metrics is to resume the work on error-based MT evaluation (Popović and Ney 2011; Toral et al. 2012), searching for a better way of automatically

21 Following Graham et al. (2013) we use the non-parametric Mann-Whitney test to test the null hypothesis that the differences between the scores assigned to the same MT output by different annotators in the low-quality sample is the same as in the high-quality sample.

detecting MT errors of different types and predicting their impact on the overall MT quality. For manual evaluation, error-based methods have already been proposed and successfully used, such as the MQM error annotation framework (Lommel, Burchardt, and Uszkoreit 2014).

6. Metrics across MT Approaches

In this section we analyze another meta-evaluation aspect: the impact of the type of MT system under evaluation on the performance of evaluation metrics. As mentioned in Section 2, some work has been done comparing MT evaluation metrics for statistical systems versus rule-based systems. However, the behavior of metrics when it comes to neural MT (NMT), the new paradigm in MT research and development, has not yet been inspected. In this section we compare how well the metrics correlate with human assessments when evaluating neural vs. statistical MT and provide an explanation for the results obtained. The analysis is performed using two data sets: the WMT16 DA data set (Section 4.1) and the MTSummit17 English–Latvian data set (Section 4.5). As in the previous sections, the results are reported for three groups of evaluation metrics: metrics based on lexical similarity, metrics that use different kinds of information on sentence structure, and feature-based metrics. As Latvian is a low-resourced language, only the metrics based on simple text similarity are available for the MTSummit17 English–Latvian data set.

WMT16 Direct Assessment Data Set. In the first experiment, we use the data from the WMT16 Direct Assessment Data Set (see Section 4.1) to compare the overall correlation between metric scores and human judgments for different MT paradigms. We selected all available data for three MT systems, corresponding to three different approaches: a phrase-based statistical MT system (PBMT), a syntax-based statistical MT (SYNTAX), and a neural MT system (NMT)—all these are University of Edinburgh's systems, as described in Williams et al. (2016) and Sennrich, Haddow, and Birch (2016). The number of sentences with available direct assessment judgments for these systems is as follows, respectively: 231, 238, and 342 sentences. As shown in Table 8, all the metrics achieve consistently higher correlation on NMT outputs, although the difference is not significant for all the metrics due to the small size of the data set.[22]

According to the results from the previous section, our initial hypothesis was that the difference in correlation can be attributed to the fact that the quality of translations produced by NMT is generally higher (this system topped the shared task [Bojar et al. 2016a]). However, further investigation with a larger data set that we present subsequently disproves this finding, suggesting that the difference in the performance of the evaluation metrics is due to inherent properties of the outputs generated by statistical versus neural MT systems.

MTSummit17 English–Latvian Data Set. In Table 9 we show results for a subset of metrics that could be computed for the MTSummit17 English–Latvian data set (Section 4.5). Overall results are similar to those in the WMT16 DA data set. As before, the correlation between metric scores and human judgments is significantly higher in the case of the NMT system. The error-based metrics (TER, PER, and WER) seem to be an exception

22 As in the previous section, we use Fisher's z-transformation to compute the significance of the difference between independent correlations.

Table 8
Pearson correlation with human judgments from WMT16 Metrics Task on the outputs of different MT systems in the WMT16 Direct Assessment data set. † indicates that the correlations for the neural MT system (NMT) is significantly different from the correlation for the statistical system (PBMT).

	PBMT	SYNTAX	NMT
-TERp-A	0.535	0.538	0.627†
Meteor	0.519	0.530	0.568†
MPEDA	0.515	0.527	0.563†
ROUGE-SU*	0.518	0.486	0.597
ChrF3	0.509	0.442	0.579
NIST-4	0.477	0.455	0.557
BLEU-4	0.430	0.415	0.507
-WER	0.406	0.388	0.525†
-TER	0.368	0.388	0.525†
-PER	0.372	0.367	0.507†
UPF-Cobalt	0.532	0.540	0.557
CP-Oc(*)	0.452	0.480	0.584†
SP-lNIST	0.490	0.465	0.559
DP-Oc(*)	0.443	0.374	0.417
SR-Or(*)	0.333	0.434	0.340
DPMFcomb	0.574	0.590	0.628
Metrics-F	0.564	0.582	0.622
Cobalt-F-comp	0.544	0.595	0.606
UoW-ReVal	0.534	0.552	0.556
BEER	0.510	0.438	0.599

and also have much smaller correlation values. A manual inspection of the data shows that this is due to the presence of outliers, namely, segments with an extremely high error rate. If the data points with metric scores lying more than four standard deviations away from the mean (less than 1% of the data) are removed, the correlation values (shown as TER*, PER*, and WER* in Table 9) are in line with other evaluation metrics.[23]

Unlike in the previous section, the quality of the translations produced by the neural MT system here is not higher—the average human scores being 1.64 for the statistical MT system and 1.84 for the neural MT system (lower scores indicate better quality for this data set). To further investigate what could be the reason why the metrics consistently show a higher correlation when evaluating the output of the neural MT system, we computed the percentage of different types of errors annotated following the MQM guidelines on a 2,000-sentence data set for both MT system type outputs (see Section 4.5). Table 2 shows the relative frequencies of each error category for statistical MT and neural MT systems. NMT contains a lower number of fluency errors (e.g. word order errors) and a much higher number of adequacy errors (in particular, mistranslation errors), as was also observed in Toral and Sanchez-Cartagena (2017). This is a plausible explanation of the difference in evaluation accuracy. Reference-based

23 Compare also the results reported for this data set in Appendix A in terms of Spearman correlation coefficient, which is less sensitive to strong outliers.

Table 9
Absolute Pearson correlation between automatic evaluation scores and human judgments in the MTSummit17 data set. † indicates that the correlations for the neural MT system is significantly different from the correlation for the statistical system. WER*, TER*, and PER* show correlations for the corresponding metrics when eliminating the outliers.

	PBMT	NMT
ROUGE-SU*	0.411	0.506†
ChrF3	0.400	0.478†
BLEU-4	0.403	0.461†
NIST-4	0.379	0.464†
WER	0.285	0.267
TER	0.270	0.260
PER	0.226	0.236
WER*	0.405	0.461†
TER*	0.400	0.463†
PER*	0.369	0.451†
BEER	0.416	0.511†

metrics are much better suited for detecting mistranslation errors than fluency errors, as they compute the similarity to a human translation (adequacy) and do not explicitly consider the appropriateness of MT output in the target language (fluency).

In this section we have compared the segment-level correlation between metric and human scores for the outputs of statistical and neural MT systems, showing that evaluation metrics consistently achieve better performance for neural MT. We suggest that this is an encouraging outcome for further work on leveraging reference-based metrics for the optimization of NMT model hyperparameters (i.e., model selection).

7. Metrics across Human Judgments

As a final part of our analysis, we compare the performance of a wide variety of metrics discussed in this article across different types of human judgments in order to investigate whether the conclusions drawn based on one evaluation setting can be extrapolated to other types of evaluation. As discussed in the previous sections, besides the type of human judgments, metric performance can be affected by various factors, such as domain, language pair, type of MT system, level of MT quality, and so forth. In order to isolate the impact of the type of human judgments on the meta-evaluation results, in the following analysis we use the data sets that belong to the news domain (with the exception of the WMT17 Quality Estimation data set), have English as the target language, and have the outputs of MT systems based on statistical approach (with the exception of the WMT16 data sets that contain different types of MT systems). However, other factors, such as the average level of MT quality or the reliability of manual evaluation scores, are more difficult to control.

Table 10 shows the correlation results for various evaluation metrics[24] discussed in previous sections for the data sets presented in Section 4. WMT16-DA refers to

24 Only the metrics available for all the data sets are included here.

Table 10
Correlation with human judgments at sentence level across various data sets with different types of human judgments. Kendall tau is reported for WMT16 Relative Ranking data set, and absolute Pearson correlation coefficient is reported for the rest of the data sets.

	WMT16 -DA	WMT16 -RA	WMT17	MTC -A	MTC -F	GALE	EAMT	EAMT -PE
Meteor	.570	.362	.358	.450	.262	.442	.280	.742
TERp-A	.570	.356	.370	.435	.268	.468	.266	.716
ROUGE-SU*	.551	.335	.379	.398	.249	.400	.254	.717
ChrF3	.541	.366	.351	.429	.222	.392	.253	.722
NIST-4	.508	.301	.335	.391	.212	.396	.224	.709
BLEU-4	.488	.289	.351	.298	.201	.353	.260	.733
TER	.462	.293	.260	.275	.195	.360	.214	.728
WER	.456	.290	.272	.246	.192	.353	.218	.737
PER	.422	.266	.226	.307	.180	.305	.167	.664
UPF-Cobalt	.566	.375	.384	.460	.281	.467	.320	.717
CP-Oc(*)	.527	.323	.362	.365	.229	.409	.263	.723
SP-lNIST	.512	.315	.334	.409	.221	.391	.226	.708
DP-Oc(*)	.424	.235	.365	.223	.164	.323	.232	.559
SR-Or(*)	.371	.208	.259	.278	.211	.295	.149	.408
BEER	.534	.366	.330	.415	.225	.386	.210	.737

the WMT16 Direct Assessment data set (Section 4.1), where human assessments were collected on a continuous adequacy scale. WMT16-RA refers to the WMT16 Ranking data set, with manual assessments having the form of preference judgments (see Section 4.2). WMT17 refers to the WMT17 Quality Estimation data set (Section 4.4), with human judgments collected using a discrete scale following the post-editing effort criterion. GALE is the GALE Arabic–English data set presented in Section 4.6, where manual evaluation was conducted following the adequacy criterion on a discrete scale from 1 to 5. MTC-A and MTC-F refer to the sets of adequacy and fluency judgments, respectively, collected using an interval scale for the Chinese–English MTC data set (see Section 4.3). EAMT11 stands for the EAMT11 French–English data set, presented in Section 4.7. The judgments were collected based on the post-editing effort. Finally, the last column in Table 10 (EAMT11-PE) shows the correlation obtained when instead of using the independently collected reference for evaluation, the MT is compared against its post-edited version. This scenario is equivalent to HTER, which has been shown to correlate very well with human judgments (Snover et al. 2006), but here we also report the performance of other metrics using this type of reference. The Pearson correlation coefficient is used for all the data sets except WMT16-RA, where the Kendall tau coefficient for pairwise ranking is reported.

In absolute terms, the correlation is different between different data sets. This is not surprising because, as mentioned before, multiple confounding factors are involved, including the difficulty of predicting a particular type of judgment, the overall level of MT quality, the reliability of manual assessments, and so on. Thus, the results in Table 10 should not be compared directly but rather in relative terms, that is, by seeing whether the ordering of the metrics based on the correlation with human assessments is similar for different data sets, annotated with different labels. Figure 13 shows a scatterplot matrix where metric correlation coefficients for each data set are plotted against the other data sets.

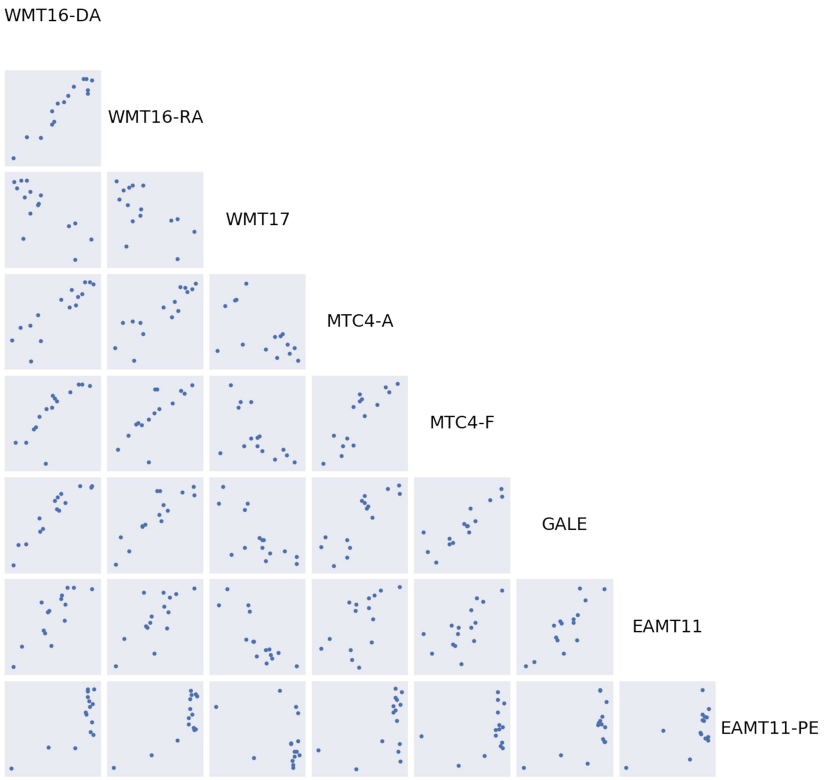


Figure 13
Scatterplot matrix for the correlation coefficients obtained by the metrics for the different data sets presented Table 10.

Although different criteria are used in manual evaluation, the results for different data sets are very closely related. In particular, WMT16, GALE, and MTC-A—where human evaluation is based on adequacy—behave very similarly. The results for MTC-F, with human scores based on the fluency criterion, are also very closely related to those for MTC-A. This is to be expected, since it has been often reported that it is difficult for human evaluators to completely isolate fluency from adequacy (Callison-Burch et al. 2007). WMT16 and EAMT11 also correlate well, although manual evaluation in EAMT11 is based on the post-editing effort criterion. Thus, the metrics that best predict adequacy seem to be the ones that also best predict fluency and post-editing effort. The explanation for such an outcome can be two-fold. On the one hand, even though the gold-standard scores are based on different criteria, the main challenges affecting metric performance are the same. On the other hand, even if the task is formulated differently, human annotators may focus on some general idea of quality, resulting in a substantial overlap between criteria for manual evaluation.

The exceptions are the EAMT11-PE and WMT17 data sets. EAMT11-PE is very different from other data sets, because it was generated using the post-edited version of the MT as reference, rather than independently created references. Unlike with a reference translation, the differences between MT output and its post-edited version are guaranteed to be related to translation errors and the challenge consists in estimating to what extent the errors affect translation quality. On average, the correlation is therefore

considerably higher for EAMT11-PE than for the other data sets. In the case of WMT17, the reason for the difference may consist of the difference in the distribution of manual evaluation scores with a very high number of near-perfect translations and a small number of low-quality outputs.

In this section we have examined how stable the meta-evaluation results are across different types of human judgments. We have seen that even though metric scores may vary considerably depending on the domain, type of human reference translation, language pair, and so forth, the relative difference in evaluation accuracy is maintained across different data sets and types of manual assessments. Thus, meta-evaluation results obtained using a particular type of assessment can generally be extrapolated to other evaluation scenarios.

8. Conclusions

Automatic MT evaluation remains a prominent field of research, with various evaluation methods proposed every year. However, studies describing the weaknesses and strengths of the existing approaches are very rare. The performance of evaluation metrics is typically assessed by computing correlation between metric scores and manual quality assessments. However, this approach has limitations. First, a correlation coefficient by itself is hardly interpretable, as metrics with very different behaviors and limitations can have the same correlation coefficients. Second, the behavior of the metrics and their correlation with human judgments may be affected by a variety of factors (language pair, domain, type of MT system, MT quality, type of manual evaluation and its reliability, etc.). We conducted a large-scale meta-evaluation study involving a set of state-of-the-art evaluation methods covering the most influential approaches to automatic MT evaluation developed in recent years. We proposed novel meta-evaluation techniques beyond overall correlation with human judgments and analyzed the influence of some of the above-mentioned factors on the performance of the evaluation metrics: MT quality, MT system types, and manual evaluation type.

First, we examined the behavior of a variety of reference-based MT evaluation metrics on MT outputs with different levels of translation quality as reflected in human judgments. We analyzed the local dependence between metric scores and human judgments. We showed that the accuracy of automatic evaluation varies depending on the overall MT quality. All the metrics examined obtain higher correlation for good quality MT. Reference-based metrics are not reliable for discriminating between low-quality translations, because with very few candidate-reference matches they lack information to draw any meaningful conclusions regarding how bad the MT output is. The metrics that use a less sparse representation for candidate-reference comparison (e.g., character n-grams) achieve the best correlation on low-quality data, suggesting that this may be a good back-off strategy for more complex evaluation systems. Besides the actual effectiveness of evaluation metrics, the correlation is also affected by the amount of noise in the results of manual evaluation used as the gold standard. We compared the consistency of manual evaluation on low-quality and high-quality translation and showed that evaluating low-quality translations is also more challenging for humans. Thus, the results indicate that both evaluation metrics and human annotators are less reliable when working with low-quality translation. In light of these findings we suggest that the focus of future research on MT evaluation should move from handling acceptable variation between MT output and reference translations to estimating the impact of translation errors on MT quality.

Second, we examined the influence of the type of MT system on automatic evaluation. We compared the performance of a wide variety of evaluation metrics for the state-of-the-art statistical MT and the recent neural MT approach on two different data sets. The metrics tested in this work achieve a higher correlation with human assessments (and therefore, are more reliable) when evaluating the outputs of neural MT systems. In order to understand the reasons for that, we compared the number of different types of errors in the translations generated by statistical MT as opposed to the ones produced by neural MT. Neural MT contains more adequacy errors, which are more easily detected by evaluation metrics than the ones affecting translation fluency. This outcome encourages further work on using evaluation metrics for direct optimization of NMT model hyperparameters (Shen et al. 2016).

Finally, we investigated to what extent the results obtained using different data sets vary in terms of how well metrics do. The performance of evaluation methods was tested on six data sets with different types of manual quality assessments. The rankings of the metrics in terms of their correlation with human judgments for the different data sets were compared. Testing the metrics using adequacy judgments generated using either discrete or continuous scale, as well as fluency judgments, produced very similar results, whereas using the judgments based on a post-editing effort criterion generated different metric orderings. This can be because some of the errors that strongly affect translation adequacy can be easy to correct—negation being a well known example of this. Overall, we have seen that the results from meta-evaluation with a particular type of manual assessment can be more often than not extrapolated to other quality aspects.

Appendix A. Spearman Correlation

For completeness, we provide the Spearman correlation coefficient values for Tables 3, 4, 8, 9, and 10.

Table A1
Conditional Spearman correlation with direct assessment scores for popular and top scoring metrics from the WMT16 Metrics for high-quality and low-quality data partitions. This table corresponds to Table 3 in the main body of the article.

	Q_{low}	Q_{high}	Q^*_{high}	*All*
Meteor	0.297	0.448	0.403	0.565
-TERp-A	0.255	0.436	0.391	0.554
MPEDA	0.295	0.444	0.399	0.563
ROUGE-SU*	0.262	0.407	0.363	0.521
ChrF3	0.311	0.388	0.339	0.515
NIST-4	0.242	0.373	0.322	0.478
BLEU-4	0.173	0.381	0.332	0.456
-TER	0.133	0.413	0.366	0.455
-WER	0.095	0.437	0.391	0.443
-PER	0.177	0.351	0.299	0.429
UPF-Cobalt	0.245	0.437	0.392	0.544
CP-Oc(*)	0.212	0.391	0.343	0.491
SP-lNIST	0.258	0.375	0.325	0.483
DP-Oc(*)	0.108	0.346	0.304	0.385
SR-Or(*)	0.148	0.196	0.198	0.307
DPMFcomb	0.298	0.471	0.428	0.589
Metrics-F	0.267	0.477	0.435	0.590
Cobalt-F-comp	0.234	0.493	0.452	0.586
BEER	0.302	0.377	0.328	0.505
UoW-ReVal	0.224	0.423	0.377	0.507

Table A2
Conditional Spearman correlation with direct assessment scores for popular and top scoring metrics from the WMT16 Metrics Task for a four-way split of the data set resulting in data samples corresponding to four quality levels (Q_1–Q_4). This table corresponds to Table 4 in the main body of the article.

	Q_1	Q_2	Q_3	Q_4	Q_4^*	*All*
Meteor	0.176	0.154	0.142	0.399	0.303	0.565
-TERp-A	0.157	0.107	0.180	0.384	0.286	0.554
MPEDA	0.177	0.153	0.146	0.393	0.297	0.563
ROUGE-SU*	0.178	0.109	0.186	0.354	0.258	0.521
ChrF3	0.185	0.128	0.139	0.345	0.243	0.515
NIST-4	0.158	0.105	0.141	0.361	0.263	0.478
BLEU-4	0.051	0.065	0.133	0.364	0.264	0.456
-TER	0.044	0.062	0.183	0.371	0.273	0.455
-WER	0.016	0.056	0.194	0.389	0.293	0.443
-PER	0.110	0.091	0.140	0.350	0.248	0.429
UPF-Cobalt	0.139	0.118	0.161	0.388	0.293	0.544
CP-Oc(*)	0.147	0.071	0.171	0.358	0.258	0.491
SP-lNIST	0.166	0.105	0.142	0.359	0.260	0.483
DP-Oc(*)	0.036	0.044	0.144	0.302	0.213	0.385
SR-Or(*)	0.098	0.081	0.042	0.171	0.173	0.307
DPMFcomb	0.178	0.139	0.186	0.405	0.312	0.589
Metrics-F	0.124	0.164	0.192	0.413	0.321	0.590
Cobalt-F-comp	0.098	0.150	0.203	0.423	0.332	0.586
BEER	0.194	0.123	0.147	0.340	0.238	0.505
UoW-ReVal	0.111	0.097	0.165	0.380	0.282	0.507

Table A3
Spearman correlation with human judgments from WMT16 Metrics Task on the outputs of different MT systems in the WMT16 Direct Assessment data set. This table corresponds to Table 8 in the main body of the article.

	PBMT	SYNTAX	NMT
-TERp-A	0.504	0.509	0.623
Meteor	0.486	0.500	0.613
MPEDA	0.478	0.493	0.608
ROUGE-SU*	0.494	0.458	0.579
ChrF3	0.478	0.405	0.568
NIST-4	0.431	0.413	0.548
BLEU-4	0.388	0.392	0.497
-WER	0.403	0.379	0.547
-TER	0.378	0.382	0.549
-PER	0.382	0.396	0.523
UPF-Cobalt	0.500	0.511	0.551
CP-Oc(*)	0.407	0.450	0.572
SP-lNIST	0.453	0.423	0.546
DP-Oc(*)	0.414	0.363	0.404
SR-Or(*)	0.316	0.366	0.292
DPMFcomb	0.521	0.555	0.624
Metrics-F	0.526	0.550	0.620
Cobalt-F-comp	0.531	0.580	0.592
UoW-ReVal	0.521	0.535	0.542
BEER	0.465	0.400	0.579

Table A4
Absolute Spearman correlation between automatic evaluation scores and human judgments in the MTSummit17 data set. This table corresponds to Table 9 in the main body of the article.

	PBMT	NMT
ROUGE-SU*	0.411	0.506
ChrF3	0.400	0.478
BLEU-4	0.403	0.461
NIST-4	0.379	0.464
WER	0.428	0.487
TER	0.424	0.492
PER	0.402	0.487
BEER	0.417	0.511

Table A5
Correlation with human judgments at sentence level across various data sets with different types of human judgments. Kendall tau is reported for WMT16 Relative Ranking data set, and absolute Spearman correlation coefficient is reported for the rest of the data sets. This table corresponds to Table 10 in the main body of the article.

	WMT16 -DA	WMT16 -RA	WMT17	MTC -A	MTC -F	GALE	EAMT	EAMT -PE
Meteor	.565	.362	.380	.431	.237	.461	.261	.719
TERp-A	.554	.356	.392	.421	.243	.441	.259	.717
ROUGE-SU*	.521	.335	.391	.390	.229	.379	.223	.712
ChrF3	.515	.366	.362	.414	.209	.376	.213	.698
NIST-4	.478	.301	.350	.380	.192	.372	.201	.696
BLEU-4	.456	.289	.357	.308	.185	.346	.228	.707
TER	.455	.293	.373	.301	.207	.367	.220	.737
WER	.443	.290	.377	.245	.184	.360	.223	.748
PER	.429	.266	.356	.336	.192	.325	.185	.693
UPF-Cobalt	.544	.375	.411	.436	.251	.458	.311	.710
CP-Oc(*)	.491	.323	.370	.360	.206	.383	.230	.704
SP-lNIST	.483	.315	.348	.396	.205	.368	.201	.697
DP-Oc(*)	.385	.235	.378	.214	.139	.307	.213	.611
SR-Or(*)	.307	.208	.202	.256	.165	.276	.083	.380
BEER	.505	.366	.358	.405	.215	.369	.194	.719

References

Amigó, Enrique, 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*, pages 306–314, Singapore.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*, pages 1–15, San Diego, CA.

Baig, Taimur and Ilan Goldfajn. 1999. Financial market contagion in the Asian crisis. *IMF Staff Papers*, 46(2):167–195.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT

evaluation with improved correlation with human judgments. In *Proceedings of the Association for Computational Linguistics (ACL) Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, volume 29, pages 65–72, Ann Arbor, MI.

Berentsen, Geir Drage, Tore Selland Kleppe, Dag Tjøstheim, et al. 2014. Introducing localgauss, an R package for estimating and visualizing local Gaussian correlation. *Journal of Statistical Software*, 56(1):1–18.

Bertero, Elisabetta and Colin Mayer. 1990. Structure and performance: Global interdependence of stock markets around the crash of October 1987. *European Economic Review*, 34(6):1155–1180.

Blatz, John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th Conference on Computational Linguistics*, pages 315–321, Geneva.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor and Marcos Zampieri. 2016a. Findings of the 2016 Conference on Machine Translation. In *First Conference on Machine Translation, Volume 2: Shared Task Papers*, WMT, pages 131–198, Berlin.

Bojar, Ondřej, Christian Federmann, Barry Haddow, Philipp Koehn, Lucia Specia, and Matt Post. 2016b. Ten years of WMT evaluation campaigns: Lessons learnt. In *Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–36, Portoroz.

Bojar, Ondřej, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513, Copenhagen.

Bojar, Ondřej, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016c. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 199–231, Berlin.

Burges, Chris, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96, Bonn.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. 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, Uppsala.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. *EACL*, 6:249–256, Trento.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120.

Charniak, Eugene and Mark Johnson. 2005. Coarse-to-fine *n*-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 173–180, Ann Arbor, MI.

Coughlin, Deborah. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*, pages 63–70, New Orleans, LA.

Culy, Christopher and Susanne Z. Riehemann. 2003. The limits of N-gram translation evaluation metrics. In *Proceedings of MT Summit IX*, pages 71–78, New Orleans, LA.

Delicado, Pedro and Marcelo Smrekar. 2009. Measuring non-linear dependence for two random variables distributed along a curve. *Statistics and Computing*, 19(3):255–269.

Denkowski, Michael and Alon Lavie. 2010. Choosing the right evaluation for machine translation: An examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the Ninth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, CO.

Denkowski, Michael and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, MD.

Doddington, George. 2002. Automatic evaluation of machine translation quality using *n*-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Diego, CA.

Doksum, Kjell, Stephen Blyth, Eric Bradlow, Xiao-Li Meng, and Hongyu Zhao. 1994. Correlation curves as local measures of variance explained by regression. *Journal of the American Statistical Association*, 89(426):571–582.

Fomicheva, Marina and Núria Bel. 2016. Using contextual information for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2755–2761, Portorož.

Fomicheva, Marina, Núria Bel, Iria da Cunha, and Anton Malinovskiy. 2015. UPF-Cobalt Submission to WMT15 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 373–379, Lisbon.

Fomicheva, Marina, Núria Bel, Lucia Specia, Iria da Cunha, and Anton Malinovskiy. 2016. CobaltF: A Fluent Metric for MT Evaluation. In *Proceedings of the First Conference on Machine Translation*, pages 483–490, Berlin.

Forbes, Kristin J. and Roberto Rigobon. 2002. No contagion, only interdependence: Measuring stock market comovements. *The Journal of Finance*, 57(5):2223–2261.

Giménez, Jesús and Lluis Marquez. 2004. Fast and accurate part-of-speech tagging: The SVM approach revisited. *Recent Advances in Natural Language Processing III*, pages 153–162.

Giménez, Jesús and Lluís Màrquez. 2010a. Asiya: An open toolkit for automatic machine translation (meta-)evaluation. *Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Giménez, Jesús and Lluís Màrquez. 2010b. Linguistic measures for automatic machine translation Evaluation. *Machine Translation*, 24(3-4):209–240.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 33–41, Sofia.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone? *Natural Language Engineering*, 23(1):3–30.

Graham, Yvette, Nitika Mathur, and Timothy Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of NAACL-HLT*, pages 1183–1191, Denver, CO.

Gupta, Rohit, Constantin Orasan, and Josef van Genabith. 2015. ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon.

Hjort, Nils Lid and M. Chris Jones. 1996. Locally parametric nonparametric density estimation. *Annals of Statistics*, 48:1619–1647.

Hong, Yongmiao, Jun Tu, and Guofu Zhou. 2006. Asymmetries in stock returns: Statistical tests and economic evaluation. *Review of Financial Studies*, 20(5):1547–1581.

Jones, M. C. and I. Koch. 2003. Dependence maps: Local dependence in practice. *Statistics and Computing*, 13(3):241–255.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. In *Proceedings of 13th International Workshop on Spoken Language Translation*, volume 1. Seattle, WA.

Kim, Hyun, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 562–568, Copenhagen.

Koehn, Philipp 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, New York, NY.

Levene, Howard. 1961. Robust tests for equality of variances. I. Olkin, S. G.

Ghurge, W. Hoeffding, W. G. Madow, and H. B. Mann, editors, *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*, pages 279–292.

Lin, Chin-Yew and Franz Josef Och. 2004a. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 605–612.

Lin, Chin-Yew 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*, pages 501–507.

Lin, Dekang. 2003. Dependency-based evaluation of minipar, In A. Abeillé, editor, *Treebanks*. Springer, pages 317–329.

Liu, Ding and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Ann Arbor, MI.

Lommel, Arle Richard, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 12:455–463.

Longin, Francois and Bruno Solnik. 2001. Extreme correlation of international equity markets. *Journal of Finance*, 56(2):649–676.

Machácek, Matouš and Ondrej Bojar. 2014. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, MD.

Melamed, Dan, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 61–63, Edmonton.

Moore, Robert C. and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala.

Nießen, Sonja, Franz Josef Och, Gregor Leusch, Hermann Ney et al. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. pages 39–45, Athens.

Padó, Sebastian, Michel Galley, Dan Jurafsky, and Chris 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: Volume 1-Volume 1*, pages 297–305, Singapore.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, PA.

Popovic, Maja. 2015. chrF: character *n*-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon.

Popovic, Maja. 2016. CHRF deconstructed: β parameters and *n*-gram weights. In *Proceedings of the First Conference on Machine Translation. Association for Computational Linguistics*, 2, pages 499–504, Berlin.

Popović, Maja and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin.

Shen, Shiqi, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1683–1692, Berlin.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, volume 200, pages 223–231, Cambridge, MA.

Snover, Matthew G, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. TER-Plus: Paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2-3):117–127.

Specia, Lucia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven.

Specia, Lucia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz. 2011. Predicting machine translation adequacy. In *Machine Translation Summit XIII*, pages 513–520, Xiamen.

Specia, Lucia, Kim Harris, Aljoscha Burchardt, Marco Turchi, Matteo Negri, and Inguna Skadina. 2017. Translation quality and productivity: A study on rich morphology languages. In *Proceedings of the 16th Machine Translation Summit*, pages 55–71, Nagoya.

Specia, Lucia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *13th Annual Conference of the European Association for Machine Translation*, pages 28–37, Barcelona.

Stanojević, Miloš and Khalil Sima'an. 2014. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, MD.

Stanojević, Miloš, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon.

Støve, Bård, Dag Tjøstheim, and Karl Ove Hufthammer. 2014. Using local Gaussian correlation in a nonlinear re-examination of financial contagion. *Journal of Empirical Finance*, 25:62–82.

Surdeanu, Mihai and Jordi Turmo. 2005. Semantic role labeling using complete syntactic analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 221–224, Ann Arbor, MI.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Tillmann, Christoph, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *Fifth European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes.

Tjøstheim, Dag and Karl Ove Hufthammer. 2013. Local Gaussian correlation: A new measure of dependence. *Journal of Econometrics*, 172(1):33–48.

Tong, Yung Liang. 1990. *The Multivariate Normal Distribution*. Springer-Verlag, New York.

Toral, Antonio, Sudip Naskar, Federico Gaspari, and Declan Groves. 2012. DELiC4MT: A tool for diagnostic MT evaluation over user-defined linguistic phenomena. *Prague Bulletin of Mathematical Linguistics*, 98:121–131.

Toral, Antonio and Víctor M. Sanchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia.

Williams, Evan James. 1959. *Regression Analysis*, volume 14. Wiley, New York.

Williams, Philip, Rico Sennrich, Maria Nadejde, Matthias Huck, Barry Haddow, and Ondrej Bojar. 2016. Edinburgh's statistical machine translation systems for WMT16. In *Proceedings of the First Conference on Machine Translation*, pages 399–410, Berlin.

Yu, Hui, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015. CASICT-DCU participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 417–421, Lisbon.

Zhang, Hao and Daniel Gildea. 2007. Factorization of synchronous context-free grammars in linear time. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 25–32, Rochester, NY.

Zhang, Lilin, Zhen Weng, Wenyan Xiao, Jianyi Wan, Zhiming Chen, Yiming Tan, Maoxi Li, and Mingwen Wang. 2016. Extract domain-specific paraphrase from monolingual corpus for automatic evaluation of machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 511–517, Berlin.