Are Automatic Metrics Robust and Reliable in Specific Machine Translation Tasks?

Mara Chinea-Ríos **Álvaro Peris** **Francisco Casacuberta**
Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València, València, Spain
{machirio, lvapeab, fcn}@prhlt.upv.es

Abstract

We present a comparison of automatic metrics against human evaluations of translation quality in several scenarios which were unexplored up to now. Our experimentation was conducted on translation hypotheses that were problematic for the automatic metrics, as the results greatly diverged from one metric to another. We also compared three different translation technologies.

Our evaluation shows that in most cases, the metrics capture the human criteria. However, we face failures of the automatic metrics when applied to some domains and systems. Interestingly, we find that automatic metrics applied to the neural machine translation hypotheses provide the most reliable results. Finally, we provide some advice when dealing with these problematic domains.

1 Introduction

Machine translation (MT) assessment is an open research question. The most accurate methods require a manual evaluation of the MT system. Unfortunately, this is a difficult and costly process, being unaffordable while developing new MT engines. Therefore, protocols for automatic evaluation of MT are required. The most common approach for evaluating the MT quality is to compare the system hypotheses with one or more reference sentences and compute a quality score.

A significant research effort has been spent on enhancing the automatic metrics. For instance, a shared task is running since 2008, as part of the Conference on Machine Translation (WMT). Although several metrics have been proposed, the literature is nowadays dominated by BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002) and, to a lesser extent, by TER (Translation Edit Rate) (Snover et al., 2006).

Despite their usefulness, those metrics may diverge, sometimes leading to deceiving conclusions. This is the case of unconventional tasks or domains (e.g Chinea-Rios et al. (2017)).

This work aims to shed some light on these behaviors, by conducting a human evaluation of MT outputs that produce inconsistencies in the metrics. More precisely, we study the correlation between human judgment and automatic evaluation on three problematic domains and for three different MT systems. We analyze the strengths and flaws that each automatic metric conveys, giving some advice for future research. The main contributions of this paper are the following:

- We deepen into an unexplored field: the evaluation of MT outputs which present inconsistencies between automatic metrics.
- We conduct a human evaluation of MT hypothesis which produced inconsistent automatic evaluations, following the direct assessment (DA) methodology.
- We compare a large number of state-of-the-art automatic metrics for our tasks at hand.
- We study the correlation of all metrics with human judgments, finding out that automatic metrics capture relatively well the human evaluation criteria in several cases.

This paper is structured as follows: in Section 2, we review relevant literature in the field of MT evaluation. Section 3 provides a brief summary of the metrics under study in this work. Section 4 explains the methods used to evaluate MT. In Section 5, we describe the experimental setup. We show and discuss the results of our evaluation in Section 6. Finally, we conclude in Section 7 by highlighting the main lessons learned from this work.

2 Related work

As stated in the previous section, the automatic evaluation of MT quality is a key element for the effective development of MT. Therefore, it has been studied from long ago (Pierce and Carroll, 1966; White et al., 1994). From here, a large amount of metrics have been proposed. Among them, the most widely used, especially in the academia, is the aforementioned BLEU. Nonetheless it is also widely accepted that BLEU suffers from several limitations when correlating with human judgments (Turian et al., 2003; Tatsumi, 2009) and can be fooled with bad translations (Smith et al., 2016). Other metrics are also common in the literature. This is the case of TER, METEOR (Lavie and Denkowski, 2009), word error rate (WER) (Klakow and Peters, 2002; Morris et al., 2004) or NIST (Doddington, 2002). Despite these efforts, the automatic assessment problem remains open, being organized several evaluation campaigns (Mauro et al., 2017) and shared tasks (Bojar et al., 2017a).

Due to the fragility and ambiguity of the existing metrics, several works attempted to perform a fine-grained evaluation of different MT systems or technologies. With the recent irruption of the neural machine translation (NMT) paradigm, a natural question arises: is NMT better than classical phrase-based statistical machine translation (PB-SMT) systems?

Several works aimed to answer this question. Toral and Sánchez-Cartagena (2017) performed an extensive comparison of NMT and PB-SMT systems, measuring several facets of the translation, such as similarity, fluency or reordering. Error analyses of NMT and PB-SMT have also been reported, either automatic (Bentivogli et al., 2018) or manual (Klubička et al., 2017). The conclusions were alike: NMT handled better verbs and nouns reordering, while the translation of proper nouns was worse.

However, it is still uncertain whether the NMT paradigm works better in situations with scarce data, as pointed out by Koehn and Knowles (2017). A solution to this issue is to add monolingual data. The usage of synthetic data in NMT has reported excellent results in terms of BLEU (Chinea-Rios et al., 2017; Sennrich et al., 2016a); but a study on the importance of adding synthetic data in NMT with respect to the human perception of translation is still missing.

3 Automatic evaluation of machine translation

In the context of this paper, the goal of automatic metrics is to assign scores to MT outputs in a way that they correlate with a human evaluation of the translation quality. In this section we briefly describe the eight metrics compared in this work. These are the most common metrics used for evaluating MT.

3.1 BLEU

BLEU tries to model the correspondence between the output from a MT system and the one produced by a human. The BLEU score is based on the n-gram precision. It counts the number of n-grams from the hypothesis that appear in the reference, dividing this count by the number of n-grams in the hypothesis. This count is clipped to the maximum number of counts that the n-gram has in any sentence of the reference document. BLEU also features a brevity penalty for short translations.

The final BLEU score is computed as a geometric mean of the n-gram precision, modified by the brevity penalty. The maximum order of the n-grams involved in the computation of BLEU is set to 4, as this provides the highest correlation with human evaluation, according to the original experimentation (Papineni et al., 2002).

3.2 METEOR

BLEU only considers n-gram precision, ignoring the recall component. Moreover, it lacks an explicit word matching. METEOR aims to mitigate these issues. METEOR is an alignment-based metric, which computes all valid alignments between the hypothesis and the references. For computing these alignments, it makes use of a stemmer and a synonym database. Therefore, this is a language-dependent metric.

Once the set of alignments is computed, the METEOR metric is a harmonic mean of the unigram precision and unigram recall, modified by an alignment penalty.

3.3 TER

The TER is defined as the minimum number of word edit operations that must be made in order to transform the hypothesis into the reference. The edit operations considered are insertion, substitution, deletion and swapping groups of words. The number of edit operations is normalized by the number of words in the reference sentence. The minimum number of edit operations is obtained by dynamic programming. Note that, unlike BLEU and METEOR, this is an error-based metric. Hence, the lower, the better.

3.4 WER

Metric based on the Levenshtein distance, working at word level. WER is based on the calculation of the number of words that differ between a piece of machine translated text and a reference translation. WER is similar to TER but ignoring the swapping operation. It was originally used for measuring the performance of speech recognition systems, but was also used in the evaluation of machine translation. As TER, the lower the WER, the better.

3.5 PER

Position independent word Error Rate (PER) (Tillmann et al., 1997) is similar to TER and WER but comparing the words in the two sentences without considering the word order. The PER score is always lower than or equal to WER. On the other hand, a shortcoming of the PER is that the word order may be important in some cases. Therefore the best solution is to calculate both word error rates.

3.6 NIST

NIST was designed to improve BLEU by rewarding the translation of infrequently used words. This was intended to prevent the inflation of MT evaluation scores by focusing on common words and high confidence translations. As a result, the NIST metric assigns larger weights to infrequent words. Similarly to BLEU, the final NIST score is computed according to the arithmetic mean of the weighted n-gram matches between the MT outputs and the reference translations. A brevity penalty is also included. The reliability and quality of the NIST metric has been shown to be superior to BLEU in several cases.

3.7 BEER

BEtter Evaluation as Ranking (BEER) (Stanojević and Sima'an, 2014a,b, 2017) is a trained evaluation metric with a linear model that combines subword feature indicators (character n-grams) and global word order features (skip bi-grams) to get a language agnostic and fast to compute evaluation metric. This metric obtained very high correlation values with human evaluations in the last evaluation campaigns (e.g. Bojar et al. (2017a)).

3.8 CHRF

Character n-gram F-score (CHRF) (Popović, 2015) computes the F_β-score on the character n-gram precision and recall. According to Popović (2015), using an F_3-score correlated best with human judgment. Its popularity is increasing, as it has shown to be a reliable metric for NMT systems.

4 Methodology

In this section we describe the human evaluation protocol applied in our work. We also describe how we computed the correlation across metrics.

4.1 Direct Assessment

Following the metrics shared task from WMT'17 (Bojar et al., 2017a), we used the monolingual DA model for evaluating the translation adequacy (Graham et al., 2017).

To obtain a correct measure of the translation quality is difficult to achieve, and the DA setup simplifies this task: unlike classical translation assessment protocols (typically bilingual), this is a simpler framework. In DA, the translation adequacy is structured as a monolingual assessment of semantic similarity, in which the reference translation and the MT hypothesis are displayed to the human evaluator. Assessors rate a translation by scoring how adequately it expresses the meaning of the reference translation. The evaluation scale ranges from 0 (worst) to 100 (perfect).

In order to avoid the skew from the different evaluators, we standardized all the scores. The standard score z of a raw score x is computed as:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ and σ are the average and standard deviation of the scores population, respectively.

4.2 Computing metric correlations

For computing the correlation between two metrics, we applied the widely used Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(h_i - \bar{h})(m_i - \bar{m})}{\sqrt{\sum_{i=1}^{n}(h_i - \bar{h})^2}\sqrt{\sum_{i=1}^{n}(m_i - \bar{m})^2}} \quad (2)$$

where h_i is the human assessment score of the i-th translation hypothesis and m_i is the corresponding scores to that hypothesis given by an automatic metric. $\bar{h}$ and $\bar{m}$ are the human and automatic mean scores, respectively.

The r coefficient ranges from $+1$ to -1, where $+1$ means total positive correlation and -1 denotes total inverse correlation. A value of 0 means that there is no linear correlation between both variables.

	Corpus		$\lvert S\rvert$	$\lvert W\rvert$	$\lvert V\rvert$	$\overline{\lvert W\rvert}$
Training	Common Crawl	En	1.5M	30M	456k	20.0
		Es		31M	522k	20.0
	XRCE – Syn	En	180k	2.2M	54k	9.4
		Es		1.7M	58k	12.2
	IT – Syn	En	150k	2.5M	76k	16.7
		Es		3.0M	78k	20.0
	E-Com – Syn	En	300k	3.2M	100k	10.6
		Es		4.1M	100k	13.6
Test	XRCE	En	1.1k	8.4k	1.6k	7.6
		Es		10.1k	1.7k	9.2
	IT	En	857	15.6k	2.1k	18.2
		Es		17.4k	2.4k	20.3
	E-Com	En	886	7.3k	874	8.2
		Es		8.6k	973	9.7

Table 1: Corpora main figures, in terms of number of sentences ($|S|$), number of words ($|W|$), vocabulary size ($|V|$) and average sentence length ($\overline{|W|}$). Syn indicates synthetic data used for fine-tuning the NMT system. M and k denote millions and thousands, respectively.

5 Experimental setup

Our experimental framework related a domain adaptation task, in the English to Spanish language direction. In our setup, we trained a PB-SMT and a NMT system on the same data, from a general corpus extracted from websites (Common Crawl). We applied these systems to three different domains: printer manuals (XRCE) (Barrachina et al., 2009), information technology[1] (IT) and Electronic Commerce (E-Com). We adapted the NMT system to these domains via synthetic data, as proposed by Chinea-Rios et al. (2017). This method consists in, for each domain, selecting related samples from a large monolingual pool, back-translating them and fine-tuning the general NMT system with these data. Table 1 show the main figures of these datasets. It is worth noting the differences existing between the domains, in terms of sentence length: The Common Crawl and IT domains featured long sentences (with around 20 words per sentence); while the XRCE and E-Com domains had much shorter sentences. This shows that the first two domains contained sentences with much more context than the two latter.

5.1 Machine translation systems

We built an attentional recurrent encoder–decoder NMT system, using the NMT-Keras[2] toolkit. The encoder and decoder were made of long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997). Following Britz et al. (2017), the LSTM, word embedding and attention model dimensions were 512 each. We applied joint byte-pair encoding (Sennrich et al., 2016b), with $32,000$ merge operations. We used Adam (Kingma and Ba, 2014) with a learning rate of 0.0002. For obtaining the translations, we used a beam search with a beam size of 6. The fine-tuning of the systems via synthetic data (denoted by NMT+Syn) was made using vanilla SGD with a learning rate of 0.05.

Our PB-SMT system was built using the standard configuration of Moses (Koehn et al., 2007). The language model was a 5-gram with modified Kneser-Ney smoothing (Kneser and Ney, 1995). The phrase table was generated employing symmetrised word alignments obtained with GIZA++ (Och and Ney, 2003). The weights of the log-linear model were tuned using MERT (Minimum Error Rate Training) (Och, 2003).

The metrics were computed using the scripts provided at the WMT metrics shared task (Bojar et al., 2017b). For all metrics, we used a single reference.

Therefore, we compared three different MT systems in three different tasks. Table 2 shows the BLEU, TER and METEOR scores of our data, as well as the scores given by the human evaluators.

This table reflects the large differences that automatic metrics may produce: for the E-Com task, the NMT system is 0.7 BLEU points worse than Moses, but its TER is more than 30 points worse than Moses. Other inconsistencies in the metrics

[1] http://metashare.metanet4u.eu/qtleapcorpus

[2] https://github.com/lvapeab/nmt-keras

Domain	System	BLEU	TER	METEOR	HUMAN
IT	Moses	33.2	45.8	60.6	58.4
	NMT	34.1	52.8	53.3	64.7
	NMT+Syn	32.2	47.3	58.3	66.3
XRCE	Moses	23.6	61.8	47.5	51.2
	NMT	22.3	78.3	44.7	47.9
	NMT+Syn	23.1	62.0	43.5	47.4
E-Com	Moses	26.2	51.8	46.8	59.7
	NMT	25.5	84.7	45.5	40.7
	NMT+Syn	30.3	52.3	48.9	43.3

Table 2: Human and automatic metrics, for all systems and domains. BLEU, METEOR and HUMAN scores range from 0 to 100, being the higher values, the better. On the other hand, the lower the TER values, the better.

can be found in this table.

5.2 Human evaluation experiments

For each domain and MT system, we randomly sampled several translation hypotheses. The samples were arranged in 8 non-overlapping blocks of 40 sentences each. Each block was evaluated by two users. Therefore, each sentence was assessed twice. Table 3 show figures of the distribution of evaluated sentences according to each system and domain. 16 human evaluators participated in our study, all native speakers of the target langauge (Spanish). None of them was a professional translator. Note that, as we are using the DA framework, the evaluators do no require any knowledge of the source language.

$\|S\|$	Domain	MT system		
		Moses	NMT	NMT+Syn
320	IT	40	24	24
	XRCE	40	32	40
	E-Com	32	48	40

Table 3: Figures of the evaluated samples. We show the total number of sentences ($|S|$) and the distribution of sentences from each domain and MT system.

We developed a web page[3] to follow the DA methodology (see Fig. 1 for an example of the front-end). The users were asked to assess *how accurately does the candidate text convey the original semantics of the reference text?*. The ratings ranged from 0 (worst) to 100 (perfect).

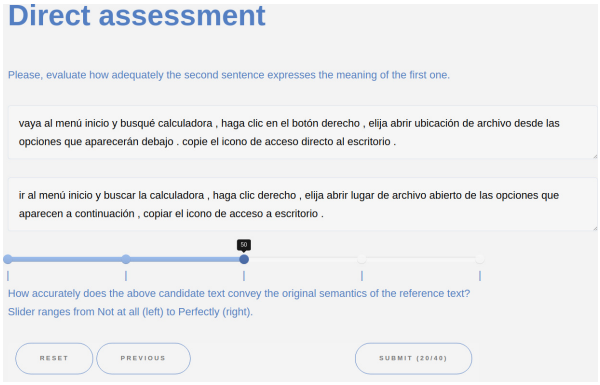


Figure 1: Front-end of the webpage developed for performing the DA protocol. The users were asked to assess how accurately does the candidate text convey the original semantics of the reference text.

6 Results and discussion

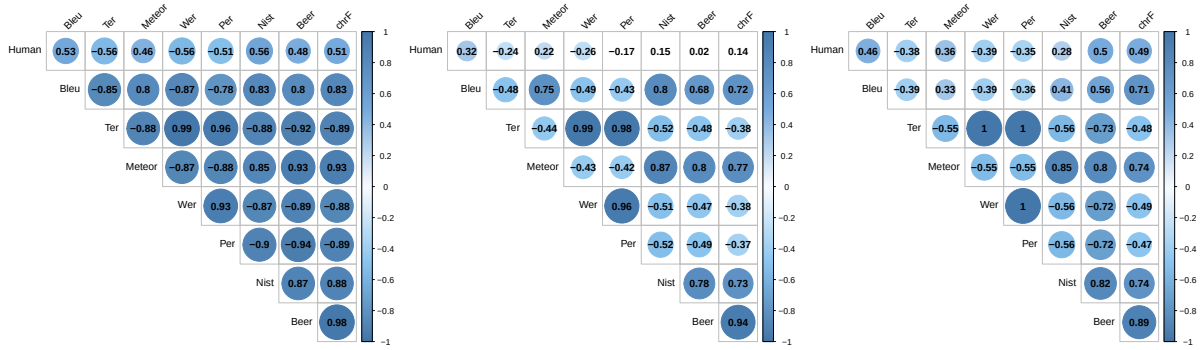
In this section, we present and discuss the results obtained from our experimentation. We analyze all metrics according to the domain and to the translation technology.

6.1 Evaluating the domains

First, we perform the analysis of each domain, regardless the translation technology applied to obtain the translations. Fig. 2 presents the correlation matrix of all metrics, for each domain. Moreover, it also shows whether the correlation of a metric with respect another is statistically significant at $p < 0.05$ (dotted cells) or not (white cell).

It is interesting to observe the large difference in terms of correlation existing between the different domains. The IT domain correlates much better with the human judgments that the other two domains (XRCE and E-Com). The reasons of such differences were found in the different corpora features: IT was a more complex corpus than the other two, featuring longer sentences and more complex syntactic structures. Moreover, in this domain, all automatic metric exhibited a significant correlation with respect to the human judgment. TER, WER and METEOR achieved the highest correlation with the human evaluation, although the differences were statistically non-significant. Therefore, we have not enough evidence to conclude which metric is better for evaluating this domain. The results for the E-Com domain were alike, having all metrics a significant correlation with humans.

[3]The evaluation platform, scores and all data used in this work can be found at: `http://lvapeab.github.io/mt_evaluation.html`.



(a) IT domain. **(b)** XRCE domain. **(c)** E-Com domain.

Figure 2: Metrics correlation across different domains (IT, XRCE, E-Com). Blue circle denote a statistically significant increase in correlation with human assessment for the metric, white cells denote a correlation without statistically significant ($p > 0.05$).

Finally, for the XRCE domain, the correlation of automatic metrics with humans were considerably lower than in the previous task. Moreover, several metrics (NIST, BEER and PER) are unable to properly correlate with humans.

Another interesting result is shown at Fig. 3. We computed a heatmap cluster of the correlation across all metrics. For all domains, the figures are divided in two main clusters. The first one, refers to n-gram-based metrics, such as BLEU, NIST, BEER and METEOR. In the second cluster, we find error-based metrics, TER, WER and TER.

This indicates that the n-gram based metrics and error-based metrics assess different aspects of the translation quality. We therefore recommend to always provide at least one metric from each family, when reporting results of translation quality.

6.2 Evaluating the translation technology

We are interested in study, not only the correlation across domains, but also the behaviors of the different MT systems. We deepen in our analysis, studying each system separately. Fig. 4 shows the correlation results for all metrics according to each domain and MT system.

As in the previous section, we find the most reliable behavior in the IT domain. Most automatic metrics are able to properly correlate with the human criteria. However, the correlations of neural-based system are higher than those obtained by Moses. In this case, the highest correlation values are found in the NMT+Syn system, greater than 0.6 in all cases.

The XRCE domain presents bad results. In this case, the metrics fail to measure the human criteria. Only BLEU for the NMT system is able to properly significantly correlate with the human assessment.

In the E-Com domain, we observe mixed results. The automatic metrics were able to correctly assess the NMT outputs, but failed with Moses. In this latter case, BLEU was the only metric that correlated well with the human evaluation.

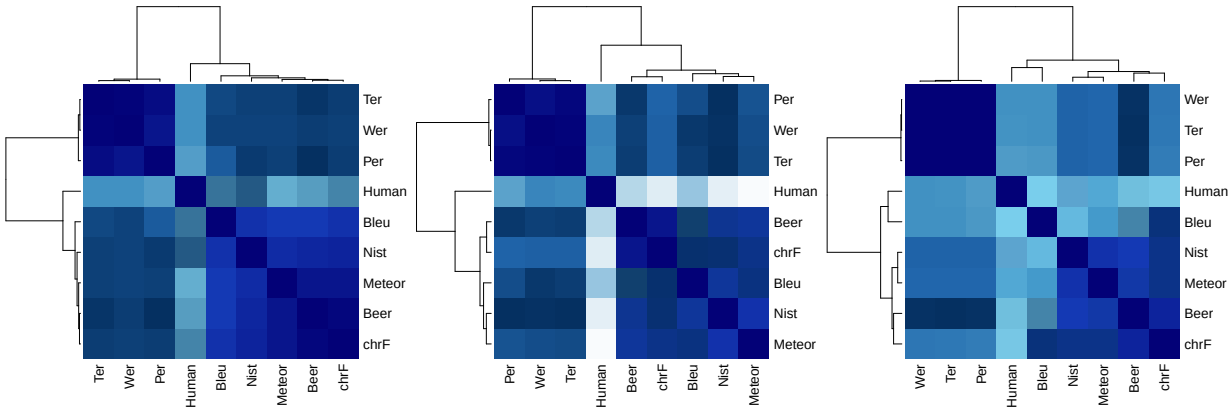
These results suggest that automatic evaluations of NMT systems (either including synthetic data or not) were systematically more reliable than the evaluation of Moses. These differences were especially dramatic as the domain contained more sentences without large contexts nor complex syntactic structures (i.e. XRCE and E-Com). The metrics provided more reliable results for the neural systems; although they can also diverge from the human criteria.

Finally, it should be noted that BLEU was metric that correlated best with human criteria in these unstructured domains. However, in domains involving sentences with more complex syntactic structures and longer contexts, BLEU is surpassed by several metrics, like TER.

7 Conclusions

In this work, we studied the behavior of automatic metrics in several translation systems for different domains. Since the metrics provided contradictory results, we conducted a human evaluation, based on the DA protocol. Next, we computed the correlation of the automatic metrics with respect to the human criteria.

Our findings were that automatic metrics were



(a) IT domain. **(b)** XRCE domain. **(c)** E-Com domain.

Figure 3: Cluster metrics correlation across different domains (IT, XRCE, E-Com); blue cell denote a statistically significant($p < 0.05$).

closer to the human as more structured and contextual the task was. When evaluating tasks with short sentences (e.g. samples from a printer manual), the correlation of the automatic metrics with respect to the human greatly fell. We also found that the automatic metrics evaluate surprisingly well NMT systems, while failing in the evaluation of classical phrase-based systems.

Finally, we also found that the metrics were clustered, even in these specific domains, according to their nature, n-gram-based or error-based. Therefore, we recommend to always give error-based and n-gram-based metrics when reporting results on MT quality.

As future work, we intend to develop a metric capable to complement the existing ones, especially when dealing with the aforementioned unstructured corpora.

Acknowledgements: We acknowledge the users who performed the human assessment in an altruistic way. The research leading to these results has received funding from the Generalitat Valenciana under grant PROMETEO/2018/004.

References

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2018). Neural versus phrase-based mt quality: An in-depth analysis on english–german and english–french. *Computer Speech & Language*, 49:52–70.

Bojar, O., Graham, Y., and Kamran, A. (2017a). Results of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513.

Bojar, O., Graham, Y., and Kamran, A. (2017b). Results of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513.

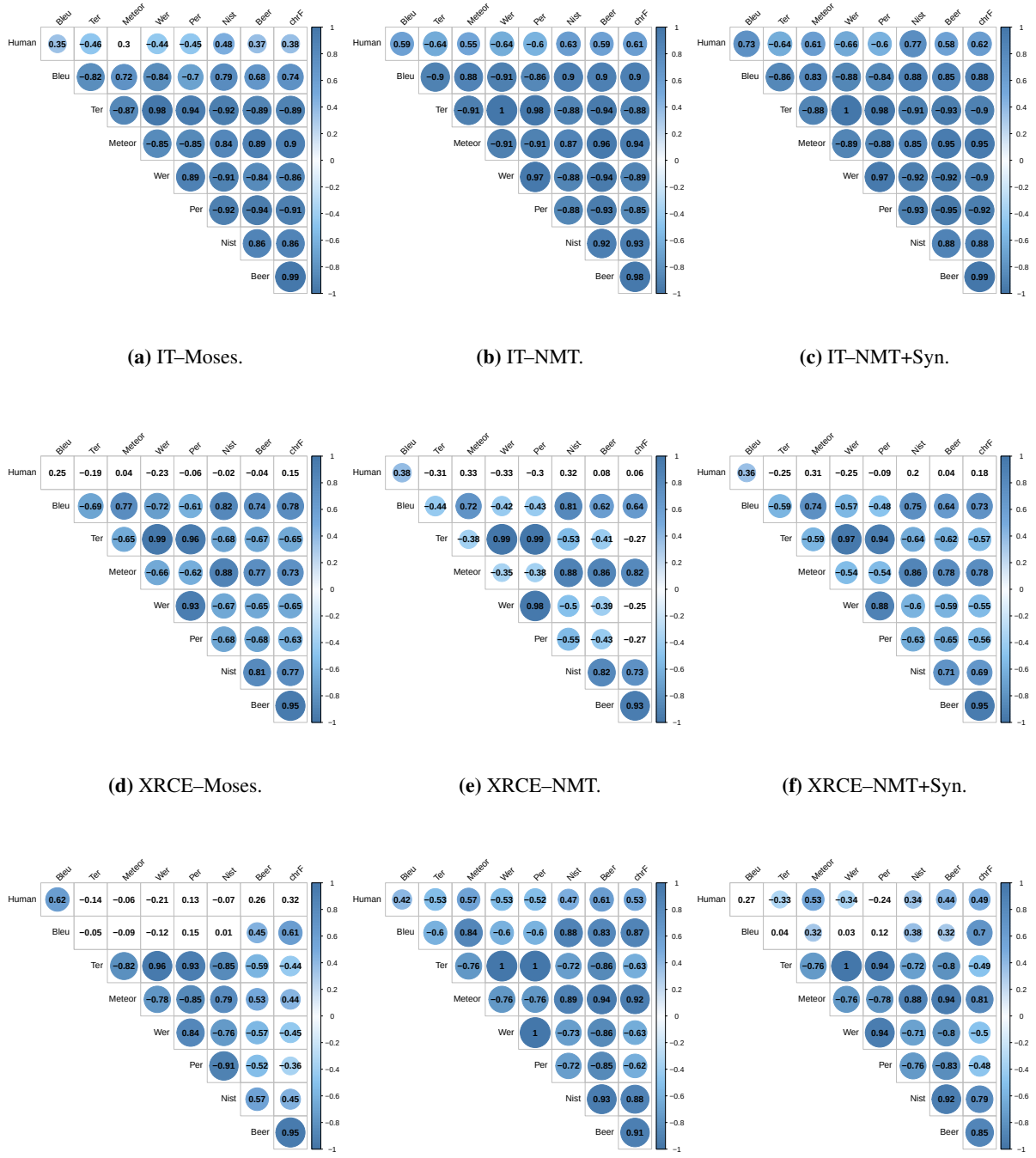
Britz, D., Goldie, A., Luong, T., and Le, Q. (2017). Massive exploration of neural machine translation architectures. *arXiv:1703.03906*.

Chinea-Rios, M., Peris, Á., and Casacuberta, F. (2017). Adapting neural machine translation with parallel synthetic data. In *Proceedings of the Second Conference on Machine Translation*, pages 138–147.

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

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.



(a) IT–Moses. **(b)** IT–NMT. **(c)** IT–NMT+Syn.

(d) XRCE–Moses. **(e)** XRCE–NMT. **(f)** XRCE–NMT+Syn.

(g) E-Com–Moses. **(h)** E-Com–NMT. **(i)** E-Com–NMT+Syn.

Figure 4: Metric correlations for each system (Moses, NMT, NMT+Syn), for all domains (IT, XRCE, E-Com).

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Klakow, D. and Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28.

Klubička, F., Toral, A., and Sánchez-Cartagena, V. M. (2017). Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Lavie, A. and Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine translation*, 23(2-3):105–115.

Mauro, Cettolo, F., Luisa, B., Jan, N., Sebastian, S., Katsuitho, S., Koichiro, Y., and Christian, F. (2017). Overview of the IWSLT 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

Morris, A. C., Maier, V., and Green, P. (2004). From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2765–2768.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Pierce, J. R. and Carroll, J. B. (1966). Language and machines: Computers in translation and linguistics.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Smith, A., Hardmeier, C., and Tiedemann, J. (2016). Climbing mount BLEU: The strange world of reachable high-BLEU translations. *Baltic Journal of Modern Computing*, 4(2):269.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*, pages 223–231.

Stanojević, M. and Sima'an, K. (2014a). Evaluating word order recursively over permutation-forests. In *Proceedings of the Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 138–147.

Stanojević, M. and Sima'an, K. (2014b). Fitting sentence level translation evaluation with many dense features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206.

Stanojević, M. and Sima'an, K. (2017). Alternative objective functions for training mt evaluation metrics. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 20–25.

Tatsumi, M. (2009). Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *Proceedings of the Machine Translation Summit*, pages 332–339.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated dp based search for statistical translation. In *Fifth European Conference on Speech Communication and Technology*.

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

Turian, J. P., Shea, L., and Melamed, I. D. (2003). Evaluation of machine translation and its evaluation. In *Proceedings of the Machine Translation Summit*, pages 386–393.

White, J., OConnell, T., and OMara, F. (1994). The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas*, pages 193–205.