

Metric and Reference Factors in Minimum Error Rate Training

Yifan He and Andy Way

*CNGL, School of Computing,
Dublin City University,
Dublin, Ireland*

February 22, 2010

Abstract.

In Minimum Error Rate Training (MERT), BLEU is often used as the error function, despite the fact that it has been shown to have a lower correlation with human judgment than other metrics such as METEOR and TER. In this paper, we present empirical results in which parameters tuned on BLEU may lead to sub-optimal BLEU scores under certain data conditions. Such scores can be improved significantly by tuning on an entirely different metric altogether, e.g. METEOR, by .0082 BLEU or 3.38% relative improvement on the WMT08 English–French data.

We analyze the influence of the number of references and choice of metrics on the result of MERT and experiment on different data sets. We show the problems of tuning on a metric that is not designed for the single reference scenario and point out some possible solutions.

Keywords: Minimum Error Rate Training, Machine Translation Evaluation, Log-linear Phrase-Based Statistical Machine Translation, BLEU, METEOR, TER, Chunk Penalty

1. Introduction

Minimum Error Rate Training (MERT: (Och, 2003)) is effective in boosting translation performance in log-linear models of phrase-based statistical machine translation (PB-SMT: (Och and Ney, 2002)) according to both automatic and human judgments. MERT tunes parameters in such models by minimizing translation errors on the error surface of the N-best list of translation candidates according to automatic evaluation metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006). Most of the time the chosen metric is BLEU, but there have also been some efforts to tune against other criteria, such as $\frac{\text{BLEU} - \text{TER}}{2}$ (Dyer et al., 2009) or the IQ_{MT} metric (Lambert et al., 2006).

In this paper, we investigate how the choice of metric and number of references influence the result of MERT. We find that in the single reference scenario (WMT08 English–French), tuning on BLEU leads to significantly inferior BLEU scores than tuning on METEOR. We replicate such results for the reverse language direction by modifying METEOR to use a static chunk penalty for all language pairs, and obtain similar results on another two Chinese–English multiple reference data sets when we use only one of the references.

© 2010 Kluwer Academic Publishers. Printed in the Netherlands.

The rest of the paper is organized as follows. Section 2 reviews MERT and Section 3 describes the different evaluation metrics we use. Section 4 describes and analyzes our experiments in the single reference scenario, while Section 5 deals with the multiple reference scenario. We conclude in Section 6, together with some avenues for further research.

2. Minimum Error Rate Training

In Statistical Machine Translation (SMT), log-linear models (Och and Ney, 2002) are used to incorporate various features h_i into SMT, as in (1):

$$p(\mathbf{e}|\mathbf{f}) = \frac{1}{Z} \exp \sum_{i=1}^n \lambda_i h_i(\mathbf{e}, \mathbf{f}) \quad (1)$$

MERT is a method of finding the best global values for parameters λ_i in such models. MERT tunes the weights λ_i of the features h_i in (1) to minimize the error function on the error surface of the N-best list of a development (or ‘dev’) set, as in (2):

$$\operatorname{argmin}_{\lambda} \operatorname{Err}(e^*(\lambda); \mathbf{ref}) \quad (2)$$

In practice, the function Err is actually approximated by a specific automatic evaluation metric m , in which case MERT is actually optimising on (3):

$$\operatorname{argmin}_{\lambda} \operatorname{err}_m(e^*(\lambda); \mathbf{ref}) \quad (3)$$

where err_m is a specific automatic evaluation metric used to estimate the errors in an output translation.

To date, research in this area has focused on (i) better search strategy and (ii) incorporating more sophisticated data representations into MERT. In the first direction, there are techniques that avoid local optima by using regularization (Cer et al., 2008) and random restarts (Moore and Quirk, 2008). With respect to (ii), one example is to use word lattices instead of an N-best list to approximate the search space (Macherey et al., 2008).

3. Automatic MT Evaluation Metrics

Automatic evaluation metrics enable researchers to quickly validate and optimise translation techniques. Simple n -gram-based metrics such as BLEU are fundamental to the development and tuning of MT systems and are now integral parts of most MERT implementations.

However, it is well known that BLEU has many limitations (Callison-Burch et al., 2006). METEOR and TER are two metrics that try to improve on BLEU’s matching strategy in different ways: METEOR performs a unigram match and uses a chunk penalty to ensure fluency, while TER basically calculates Edit Distance (Levenshtein, 1966) and uses the *shift* operation to capture reordering.¹

3.1. BLEU

BLEU is the most popular evaluation metric in MT development. Although it suffers from several shortcomings, such as low correlation with human judgment on the sentence level, preference to statistical systems (Callison-Burch et al., 2006) and inconsistency in related evaluation scenarios (Chiang et al., 2008), it is still the automatic evaluation metric used in many translation campaigns and remains the most often used error function in MERT.

BLEU performs n -gram matching between the output and the reference, using n -gram precision with a brevity penalty as the score, as in (4):

$$\text{BLEU}(n) = \prod_{i=1}^n \text{PREC}_i^{\frac{1}{n}} \cdot bp \quad (4)$$

where n is the order of n -gram, PREC_i is the i -gram precision and bp is the brevity penalty, as in (5):

$$bp = \exp(\min(1 - \frac{\text{len}(Ref)}{\text{len}(Out)}, 0)) \quad (5)$$

where $\text{len}(Ref)$ is the length of the reference and $\text{len}(Out)$ is the length of the output. The n -gram matching scheme in BLEU makes it very sensitive to small changes in the output, especially when only one reference is available and linguistic variations are less likely to be captured. It has been shown in evaluation tasks (Callison-Burch et al., 2008) that BLEU has a lower correlation with human judgment than newer metrics that make use of more linguistic resources and better matching strategies, including METEOR and TER.

3.2. METEOR

METEOR tries to solve the problems of BLEU by performing multi-stage unigram matching and adding recall into consideration. With the use of unigram matching, METEOR is less sensitive to variations in word order, and

¹ Other non-string-based MT metrics (e.g. (Owczarzak et al., 2007)) exploit deeper features, but we do not test with such metrics here because their computation is typically heavy which renders them less appropriate for MERT tuning.

with multi-stage matching, METEOR can consider stemming and WordNet semantic information. The METEOR score is calculated as in (6):

$$\text{METEOR} = \frac{PR}{\alpha P + (1 - \alpha)R} \cdot (1 - cp) \text{ in which } cp = \gamma \cdot \left(\frac{\#chunks}{\#matches} \right)^\beta \quad (6)$$

where P is the unigram precision, R is the unigram recall and cp is the chunk penalty, which is used to penalize fluent outputs (cf. Section 4.2).

METEOR uses the different parameters α , β and γ for different target languages. This causes some of the undesired behaviour of MERT that we discuss in Section 4.3.

3.3. TER

TER is an Edit Distance-style evaluation metric. It calculates how many insertions, deletions, modifications and sequence shifts are needed to make the output and reference token sequences identical. The difference between TER and the classical Levenshtein Distance (Levenshtein, 1966) is the sequence shift operation, which allows phrasal shifts in the output to be captured. TER is calculated as in (7):

$$\text{TER} = \frac{\#INS + \#DEL + \#MOD + \#SHIFT}{\text{len}(Ref)} \quad (7)$$

There is no explicit penalty on the length of the sentence in TER and the calculation of TER is based on counting edits/errors. As a result, TER prefers shorter sentences in MERT, as we will show in our experiments.

4. MERT with Single Reference

In some shared tasks and real world applications, only one reference is available in the dev and test sets. In this section we investigate how the choice of metrics and number of references affects MERT in such cases.

In current practice, the metric of choice in tuning is often BLEU. Our experiments show that other metrics should perhaps be chosen if tuning on a single reference data set and the size of the N-best list is limited.

We tune on four single metrics: BLEU, METEOR, METEOR-SCP (Meteor with static chunk penalty, where chunk penalty γ is set to 1.0) and TER, and evaluate the results on our test set with BLEU, METEOR and TER. We also report the length ratio LEN of outputs, where n is the number of references, as in (8):

$$\text{LEN} = \frac{\text{len}(Output)}{\sum_i^n \text{len}(Ref_i)} \cdot n \quad (8)$$

As some metrics are biased to longer/shorter outputs, the length ratio helps us to see whether a change in score is a real improvement, or rather a bias.

4.1. EXPERIMENTAL SETTINGS

We run experiments with single reference translations on three sets of data: the English–French and French–English Workshop on Statistical Machine Translation (WMT) 2008 dev set, and Chinese–English NIST-MT (MT06) dev/test sets, and the International Workshop on Spoken Language Translation (IWSLT) 2005 dev set and 2007 test set.

For the WMT experiments, we use the top-1000 sentences in the original development set as the dev set and the remaining 1000 as our test set. We train the translation model and a 4-gram language model on Europarl.² For the NIST/IWSLT experiments, meanwhile, we use the language and translation models trained on data prepared by the organisers, and the first of the reference translations provided. In all experiments, MERT is performed on the 100-best list (a larger N-best list can help but will not change the results significantly) generated by the phrase-based decoder Moses³ with 20 starting points to avoid local optima. MERT is performed with a modified version of ZMERT.⁴ The features we use are the default features of Moses: phrase translation model, language model, distance-based reordering model, word penalty and lexicalized reordering model.

We use NIST BLEU v13 (which uses the closest reference for the brevity penalty),⁵ METEOR 0.7 (without WordNet synonyms)⁶ and TER 0.725⁷ as implementations of the chosen evaluation metrics. We also introduce in Section 4.2 a slightly modified version of METEOR to compensate for the length bias of METEOR.

4.2. ADAPTING METEOR FOR MERT WITH A STATIC CHUNK PENALTY

In Table I we see that tuning on METEOR with the default parameters for English leads to inferior scores than tuning on BLEU when performing MERT on the French–English WMT data set. One reason for this is that tuning on METEOR results in verbose translations, where the output sentences are around 9% longer than the references. However, in the reverse direction, this does not happen and the length of the output is in the normal range.

We assume that this is caused by the different chunk penalties that METEOR assigns to different languages. For French γ (cf. (6) above) is 1.0,

² <http://www.statmt.org/europarl/>

³ <http://www.statmt.org/ Moses/>

⁴ <http://www.cs.jhu.edu/~ozaidan/zmert/>

⁵ <http://www.itl.nist.gov/iad/mig/tests/mt/2008/scoring.html>

⁶ <http://www.cs.cmu.edu/~alavie/METEOR/>

⁷ <http://www.cs.umd.edu/~snoover/tercom/>

Table I. Experimental Results WMT08 French–English. MSCP: METEORSCP. LEN: Length Ratio. Rows are tuning criterion, columns are evaluation scores on dev and test sets.

	Dev set				Test set			
	BLEU	MET	TER	LEN	BLEU	MET	TER	LEN
BLEU	0.3070	0.5449	0.5404	100%	0.3276	0.5552	0.5252	100%
METEOR	0.2938	0.5553	0.5697	108%	0.3142	0.5638	0.5548	109%
TER	0.2735	0.5258	0.5396	92%	0.2946	0.5373	0.5255	93%
MSCP	0.3113	0.5540	0.5382	103%	0.3294	0.5631	0.5255	103%

Table II. Experimental Results WMT08 English–French. LEN: Length Ratio. Rows are tuning criterion, columns are evaluation scores on dev and test sets.

	Dev set				Test set			
	BLEU	MET	TER	LEN	BLEU	MET	TER	LEN
BLEU	0.2297	0.1676	0.6377	100%	0.2429	0.1763	0.6198	99%
MET	0.2363	0.1748	0.6209	97%	0.2511	0.1829	0.6032	96%
TER	0.2285	0.1711	0.5982	89%	0.2392	0.1782	0.5924	89%

but for English γ is 0.28. To fix this, we set a *static* $\gamma = 1.0$ for all target languages. The method is denoted as METEOR with static chunk penalty, METEOR-SCP.

The results in Table I show that tuning on METEOR-SCP leads to .0152 (4.84% relative) better BLEU points than tuning on METEOR, and .0018 points (0.55% relative) higher BLEU compared to tuning on BLEU. It shows that the chunk penalty fixes METEOR’s bias towards longer outputs to some extent, and at the same time preserves METEOR’s better predictive power of translation quality than BLEU.⁸

4.3. RESULTS AND ANALYSIS

The results are given in Tables I (French–English), II (English–French), III and IV (Chinese–English). Scores in bold are significantly (Koehn, 2004) ($p < 0.05$, 1,000 bootstraps) better than the others listed.

⁸ Recently version 0.8 of METEOR was released with the addition of a length penalty, which can serve a similar purpose to METEOR-SCP in MERT. It remains to be explored how this penalty could be most effectively used in MERT.

We can interpret the results in three respects. Firstly, translation quality as measured by a specific metric is not entirely dependent on the tuned parameters for that metric. For example, in the WMT08 English–French direction, tuning on METEOR can produce significantly better BLEU scores than tuning on BLEU itself (.2363 vs. .2297, i.e. .0066 BLEU points, or 2.87% relative improvement on the dev set; .2511 vs. .2429, i.e. .0082 BLEU points, or 3.38% relative improvement on the test set). In the reverse direction, scores received by METEOR-SCP are at the same level of statistical significance as the best scores on all three metrics. On both MT06 and IWSLT07 test sets, BLEU-tuned results do not receive significantly better BLEU scores than METEOR-SCP-tuned outputs. In some of the results e.g. those in Table II, MERT does not produce the optimal result on the dev set according to the metric it optimizes on. This can be explained by the influence of random factors (choice of dev set, size of N -best list, etc.), but it does show that as MERT does not guarantee a global optimum, it can be trapped when optimising on a less robust metric.

Table III. Experimental Results MT06 Chinese–English Single Reference. MSCP: METEORSCP. LEN: Length Ratio.

	Dev set				Test set			
	BLEU	MET	TER	LEN	BLEU	MET	TER	LEN
BLEU	0.1144	0.3972	0.7681	100%	0.1271	0.4227	0.7101	96%
METEOR	0.1008	0.4120	0.8654	112%	0.1140	0.4343	0.8060	113%
TER	0.0638	0.3301	0.76961	79%	0.0657	0.3450	0.7465	73%
MSCP	0.1088	0.4056	0.8158	107%	0.1272	0.4329	0.7563	106%

Secondly, metrics are often biased to longer or shorter sentences, which prevent some metrics from making correct judgments on translation quality during tuning.

In our experiments, tuning on TER sometimes leads to results that obtain the best TER, but often the worst BLEU and METEOR scores. In the French–English direction, METEOR suffers from a similar problem, as TER favours shorter sentences while METEOR favours longer MT output. The problematic METEOR and TER-measured error surfaces (cf. Section 4.4) might be the root cause of these results.

The length ratio of outputs tuned with different metrics are listed in the tables beside the scores from the specific evaluation metrics. In our WMT08 French–English experiments, TER generates outputs that are 7% shorter than the reference, and METEOR generates 9% longer translations. Note that with

Table IV. Experimental Results IWSLT07 Chinese–English Single Reference. MET: METEOR. MSCP: METEORSCP. LEN: Length Ratio.

	Dev set				Test set			
	BLEU	MET	TER	LEN	BLEU	MET	TER	LEN
BLEU	0.3556	0.5945	0.8345	124%	0.2618	0.4466	0.5659	93%
METEOR	0.3210	0.6013	0.9131	135%	0.2532	0.4545	0.6071	103%
TER	0.3587	0.5906	0.8101	121%	0.2387	0.4325	0.5757	89%
MSCP	0.3471	0.5861	0.8409	125%	0.2584	0.4475	0.5676	93%

respect to Chinese–English, on the MT06 data there is a huge discrepancy with respect to the length of the outputs generated by TER and METEOR, relative to the length of the translations output using BLEU for MERT. We introduced METEOR-SCP in Section 4.2 as an attempt to fix this problem, and the relative length of METEOR-SCP-tuned outputs are considerably shorter than those using METEOR.⁹

Thirdly, our experiments show that in the single reference scenario, BLEU is not the only applicable error function in MERT, even when our aim is simply to improve on the BLEU score itself, regardless of translation quality. In previous shared tasks of WMT, there have been submissions that use other metrics for tuning (e.g. (Dyer et al., 2009)) in order to achieve higher correlation with human judgment. In our experiments, however, tuning on METEOR or METEOR-SCP can be better than tuning on BLEU even if our aim is to obtain a higher BLEU score (cf. Table I for French–English, and Table III for Chinese–English).

4.4. ERROR SURFACES

In order to understand the results in single reference tuning, we investigate the actual error surface of the MERT according to different metrics and the results of MERT tuning. To illustrate how the error surface is measured by different metrics, we present how metric scores alter when the weight for the phrase translation probability $\varphi(f|e)$ changes.

We use the default weights in Moses (0.2 for translation model weights, 0.3 for distortion weights, 0.5 for language model weight and -1 for word

⁹ We do not apply the same methods on TER, as this metric cannot be tuned according to a set of parameters, but its successor TERP (cf. <http://www.umiacs.umd.edu/~snover/terp/>) is designed for use with parameters. We plan to carry out experiments on MERT using TERP in future work.

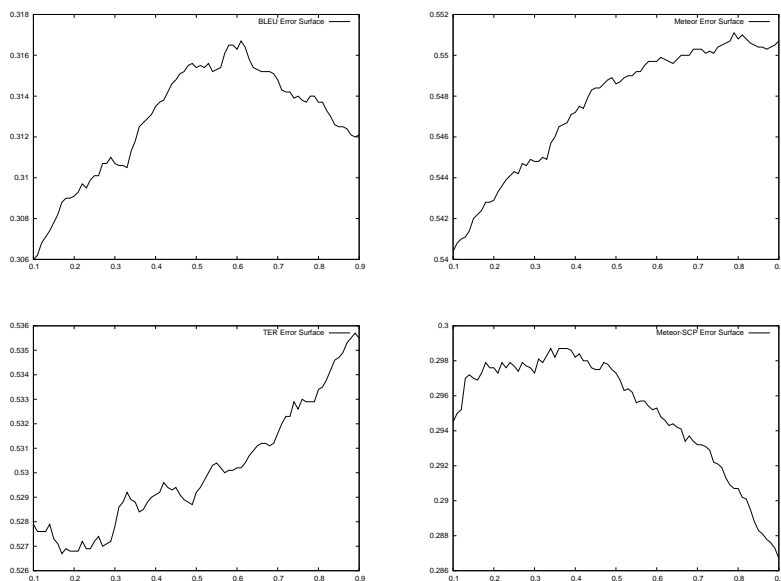


Figure 1. Error surfaces measured by various evaluation metrics. Upper left: BLEU; Upper right: METEOR; Lower left: TER; Lower right: METEOR-SCP.

penalty weight) and sample the weight of $\varphi(f|e)$ between 0.1 and 0.9. The error surfaces are reported in Figure 1 using the WMT08 French–English dev set.

We find the error surfaces measured by METEOR and TER to be problematic, as both curves are monotonically increasing. In the case of METEOR, it would probably lead MERT to converge between 0.8 and 0.9 if only the weight of $\varphi(f|e)$ is tuned. However, as the weight of the language model is set to 0.5, such a large weight for the translation model would not only be counter to the usual practice in MT development, but also would reflect the bias of the error surface. The curve of TER is of a similar shape, but as TER is to be minimized instead of maximized, it would lead MERT to converge at a small weight for $\varphi(f|e)$.

When we use METEOR with a strict chunk penalty of 1.0, the error surface is more similar to the error surface measured by BLEU. Furthermore, we find that in this setting the METEOR-SCP error surface has a stable plateau in the range (0.2, 0.5), which means that even if MERT does not find the global optimum, then local optima in this range will not affect the final result as much as tuning on BLEU. Although we cannot categorically conclude this to be true with just a single example, we feel that the different error surfaces of the various metrics in MERT merit further investigation.

4.5. EXAMINING TRANSLATION OUTPUT

We provide the following example to show how different the tuned outputs appear, in addition to judging the translation quality with automatic evaluation scores. BLEU-, METEOR- and TER-oriented tuning give the following output for the same input sentence:

REF: *it is important that our products are safe , but should not go over the top with extreme actions .* (20 tokens)

BLEU-TUNED: *it is important that our products are safe , but we must keep going beyond certain limits by carrying out actions extremes .* (23 tokens)

METEOR-TUNED: *it is important that our products are safe , but we must be careful not to go beyond certain limits as a combination of extreme actions .* (27 tokens)

TER-TUNED: *it is important that our products are safe , but we must not exceed certain limits extreme through actions .* (20 tokens)

It is clear that these outputs follow the tendencies which we detailed in Section 4.3, namely that tuning on METEOR and TER leads to longer or shorter outputs respectively than tuning on BLEU. In the METEOR-tuned sentence, the erroneous ‘as a combination of’ is used instead of ‘with’, possibly due to METEOR’s bias towards longer output. It is also worth noting that the BLEU-tuned output omits ‘not’, in so doing changing the meaning of the sentence completely, while METEOR- and TER-tuned outputs both keep it.

5. MERT with Multiple References

These experiments are run with the same settings as in the single reference scenario, on MT06 and IWSLT datasets, using 4 references from each dataset. Note again that only one reference is provided for WMT tasks.

The results are shown in Tables V and VI. The bold numbers are significantly better than others ($p < 0.05$, 1,000 bootstraps).

In this scenario, optimising on each metric most often produces the best scores on that metric. We suspect that multiple references improve the estimation power of the evaluation metrics and generate more stable results.

On MT06, we see in Table V that although METEOR-SCP cannot produce higher BLEU scores than tuning on BLEU as in the single reference scenario, it still significantly improves upon the original version of METEOR in both BLEU and TER scores, and the length ratio is again more acceptable. It even receives a better (original) METEOR score (.4648 vs. .464, an improvement of .008 points on the test set). These show again how the default value of

Table V. Experimental Results MT06 Chinese–English Multiple References. MSCP: METEOR-SCP. LEN: Length Ratio.

	Dev set				Test set			
	BLEU	MET	TER	LEN	BLEU	MET	TER	LEN
BLEU	0.2112	0.4544	0.6835	98%	0.2007	0.4547	0.6962	102%
METEOR	0.1917	0.4653	0.7458	110%	0.1784	0.4640	0.8197	121%
TER	0.1739	0.4083	0.6289	77%	0.1792	0.4103	0.6354	77%
MSCP	0.2021	0.4632	0.7159	105%	0.1932	0.4648	0.7611	113%

Table VI. Experimental Results IWSLT07 Chinese–English Multiple References. MET: METEOR. MSCP: METEOR-SCP. LEN: Length Ratio.

	Dev set				Test set			
	BLEU	MET	TER	LEN	BLEU	MET	TER	LEN
BLEU (B)	0.4149	0.6261	0.7168	114%	0.3292	0.4955	0.4871	89%
MET (M)	0.3884	0.6332	0.7570	122%	0.3234	0.5072	0.5136	96%
TER (T)	0.4070	0.6151	0.6886	110%	0.3110	0.4809	0.4876	85%
MSCP	0.3858	0.6260	0.7769	124%	0.3252	0.5020	0.5165	96%

γ in METEOR causes a bias to verbose outputs during tuning. On IWSLT, however, we see in Table VI that using the strict chunk penalty of 1.0 has a smaller effect. The possible reason for this is that changing the Chunk Penalty is less likely to fix the length of the outputs when outputs are short, such as those in the IWSLT test set.

Moreover, multiple references amplify the bias of METEOR and TER towards longer/shorter sentences. On MT06, METEOR with the original chunk penalty leads to outputs that can be 21% longer than the references, while TER leads to 23% shorter outputs. All these length ratios are worse than those tuned with a single reference, and are away from the rational range. In such cases, these biases hinder these two metrics—believed to have better predictive power than BLEU—from materializing their advantage over BLEU in MERT.

Multiple references also give BLEU better predictive power. Papineni et al. (2002) use multiple references in BLEU to capture variations in translation, and there is no other means to allow variation in BLEU.

6. Conclusion and Future Work

In this paper, we explored the effect of using different automatic evaluation metrics and different numbers of references while tuning with MERT.

When only one reference was available, we showed that tuning on BLEU yielded inferior BLEU scores than tuning on METEOR or its variants. We also showed that the bias of the original METEOR metric towards verbose outputs on some languages could be fixed by simply keeping the length penalty invariant.

When multiple references were available, tuning on BLEU led to more convincing results. The possible reason for this was that multiple references enabled better prediction from BLEU, and at the same time magnified the length bias of METEOR and TER.

We contend that BLEU is not designed for the single reference scenario, in which case it may be better to rely on more discerning metrics such as METEOR, or a combination of different metrics, when it comes to capturing translation variances. For future research, it would be interesting to find more reliable objective functions for MERT. He and Way (2009) combine different metrics to give a better objective function than any single metric in the single reference scenario. It would also be beneficial to build metrics that can avoid any bias towards certain characteristics of the output sentences other than quality.

Acknowledgements

This research is funded by Science Foundation Ireland¹⁰ grant number 07/CE/I1142.

References

- Banerjee, S. and A. Lavie: 2005, 'METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments'. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, MI, pp. 65–72.
- Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, and J. Schroeder: 2008, 'Further Meta-Evaluation of Machine Translation'. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, OH, pp. 70–106.
- Callison-Burch, C., M. Osborne, and P. Koehn: 2006, 'Re-evaluation the Role of Bleu in Machine Translation Research'. In: *EACL-2006, 11th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*. Trento, Italy, pp. 249–256.
- Cer, D., D. Jurafsky, and C. Manning: 2008, 'Regularization and Search for Minimum Error Rate Training'. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, OH, pp. 26–34.

¹⁰ <http://www.sfi.ie>

- Chiang, D., S. DeNeefe, Y. S. Chan, and H. T. Ng: 2008, 'Decomposability of Translation Metrics for Improved Evaluation and Efficient Algorithms'. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, HI, pp. 610–619.
- Dyer, C., H. Setiawan, Y. Marton, and P. Resnik: 2009, 'The University of Maryland Statistical Machine Translation System for the Fourth Workshop on Machine Translation'. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, pp. 145–149.
- He, Y. and A. Way: 2009, 'Improving the Objective Function in Minimum Error Rate Training'. In: *Proceedings of the Twelfth Machine Translation Summit*. Ottawa, ON, Canada, pp. 238–245.
- Koehn, P.: 2004, 'Statistical Significance Tests for Machine Translation Evaluation'. In: *Proceedings the 2004 Conference of Empirical Methods in Natural Language Processing (EMNLP-2004)*. Barcelona, Spain, pp. 388–395.
- Lambert, P., J. Giménez, M. R. Costa-jussà, E. Amigó, R. E. Banchs, L. Màrquez, and J. A. R. Fonollosa: 2006, 'Machine Translation System Development Based on Human Likeness'. In: *Proceedings of the IEEE/ACL Workshop on Spoken Language Technology*. Palm Beach, Aruba, pp. 246–249.
- Levenshtein, V. I.: 1966, 'Binary Codes Capable of Correcting Deletions, Insertions, and Reversals'. *Soviet Physics Doklady* **10**(8), 707–710.
- Macherey, W., F. Och, I. Thayer, and J. Uszkoreit: 2008, 'Lattice-based Minimum Error Rate Training for Statistical Machine Translation'. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, HI, pp. 725–734.
- Moore, R. C. and C. Quirk: 2008, 'Random Restarts in Minimum Error Rate Training for Statistical Machine Translation'. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK, pp. 585–592.
- Och, F. J.: 2003, 'Minimum error rate training in statistical machine translation'. In: *41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan, pp. 160–167.
- Och, F. J. and H. Ney: 2002, 'Discriminative Training and Maximum Entropy Models for Statistical Machine Translation'. In: *40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, pp. 295–302.
- Owczarzak, K., J. van Genabith, and A. Way: 2007, 'Labelled Dependencies in Machine Translation Evaluation'. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, pp. 104–111.
- Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu: 2002, 'Bleu: a Method for Automatic Evaluation of Machine Translation'. In: *40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, pp. 311–318.
- Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul: 2006, 'A study of translation edit rate with targeted human annotation'. In: *AMTA 2006, Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Visions for the Future of Machine Translation*. Cambridge, MA, pp. 223–231.

