PORT: a Precision-Order-Recall MT Evaluation Metric for Tuning

Boxing Chen, Roland Kuhn and **Samuel Larkin**
National Research Council Canada
283 Alexandre-Taché Boulevard, Gatineau (Québec), Canada J8X 3X7
{Boxing.Chen, Roland.Kuhn, Samuel.Larkin}@nrc.ca

Abstract

Many machine translation (MT) evaluation metrics have been shown to correlate better with human judgment than BLEU. In principle, tuning on these metrics should yield better systems than tuning on BLEU. However, due to issues such as speed, requirements for linguistic resources, and optimization difficulty, they have not been widely adopted for tuning. This paper presents PORT[1], a new MT evaluation metric which combines precision, recall and an ordering metric and which is primarily designed for tuning MT systems. PORT does not require external resources and is quick to compute. It has a better correlation with human judgment than BLEU. We compare PORT-tuned MT systems to BLEU-tuned baselines in five experimental conditions involving four language pairs. PORT tuning achieves consistently better performance than BLEU tuning, according to four automated metrics (including BLEU) and to human evaluation: in comparisons of outputs from 300 source sentences, human judges preferred the PORT-tuned output 45.3% of the time (*vs.* 32.7% BLEU tuning preferences and 22.0% ties).

1 Introduction

Automatic evaluation metrics for machine translation (MT) quality are a key part of building statistical MT (SMT) systems. They play two roles: to allow rapid (though sometimes inaccurate) comparisons between different systems or between different versions of the same system, and to perform tuning of parameter values during system training. The latter has become important since the invention of minimum error rate training (MERT) (Och, 2003) and related tuning methods. These methods perform repeated decoding runs with different system parameter values, which are tuned to optimize the value of the evaluation metric over a development set with reference translations.

MT evaluation metrics fall into three groups:

- BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), WER, PER, TER (Snover *et al.*, 2006), and LRscore (Birch and Osborne, 2011) do not use external linguistic information; they are fast to compute (except TER).
- METEOR (Banerjee and Lavie, 2005), METEOR-NEXT (Denkowski and Lavie 2010), TER-Plus (Snover *et al.*, 2009), MaxSim (Chan and Ng, 2008), TESLA (Liu *et al.*, 2010), AMBER (Chen and Kuhn, 2011) and MTeRater (Parton *et al.*, 2011) exploit some limited linguistic resources, such as synonym dictionaries, part-of-speech tagging, paraphrasing tables or word root lists.
- More sophisticated metrics such as RTE (Pado *et al.*, 2009), DCU-LFG (He *et al.*, 2010) and MEANT (Lo and Wu, 2011) use higher level syntactic or semantic analysis to score translations.

Among these metrics, BLEU is the most widely used for both evaluation and tuning. Many of the metrics correlate better with human judgments of translation quality than BLEU, as shown in recent WMT *Evaluation Task* reports (Callison-Burch *et*

[1] PORT: **P**recision-**O**rder-**R**ecall **T**unable metric.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 930–939,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

al., 2010; Callison-Burch *et al.*, 2011). However, BLEU remains the *de facto* standard tuning metric, for two reasons. First, there is no evidence that any other tuning metric yields better MT systems. Cer *et al.* (2010) showed that BLEU tuning is more robust than tuning with other metrics (METEOR, TER, *etc.*), as gauged by both automatic and human evaluation. Second, though a tuning metric should correlate strongly with human judgment, MERT (and similar algorithms) invoke the chosen metric so often that it must be computed quickly.

Liu *et al.* (2011) claimed that TESLA tuning performed better than BLEU tuning according to human judgment. However, in the WMT 2011 "tunable metrics" shared pilot task, this did not hold (Callison-Burch *et al.*, 2011). In (Birch and Osborne, 2011), humans preferred the output from LRscore-tuned systems 52.5% of the time, versus BLEU-tuned system outputs 43.9% of the time.

In this work, our goal is to devise a metric that, like BLEU, is computationally cheap and language-independent, but that yields better MT systems than BLEU when used for tuning. We tried out different combinations of statistics before settling on the final definition of our metric. The final version, PORT, combines precision, recall, strict brevity penalty (Chiang *et al.*, 2008) and strict redundancy penalty (Chen and Kuhn, 2011) in a quadratic mean expression. This expression is then further combined with a new measure of word ordering, *v*, designed to reflect long-distance as well as short-distance word reordering (BLEU only reflects short-distance reordering). In a later section, 3.3, we describe experiments that vary parts of the definition of PORT.

Results given below show that PORT correlates better with human judgments of translation quality than BLEU does, and sometimes outperforms METEOR in this respect, based on data from WMT (2008-2010). However, since PORT is designed for tuning, the most important results are those showing that PORT tuning yields systems with better translations than those produced by BLEU tuning – both as determined by automatic metrics (including BLEU), and according to human judgment, as applied to five data conditions involving four language pairs.

2 BLEU and PORT

First, define n-gram precision $p(n)$ and recall $r(n)$:

$$p(n) = \frac{\#\text{n-grams}(T \cap R)}{\#\text{n-grams}(T)} \quad (1)$$

$$r(n) = \frac{\#\text{n-grams}(T \cap R)}{\#\text{n-grams}(R)} \quad (2)$$

where T = translation, R = reference. Both BLEU and PORT are defined on the document-level, *i.e.* T and R are whole texts. If there are multiple references, we use closest reference length for each translation hypothesis to compute the numbers of the reference n-grams.

2.1 BLEU

BLEU is composed of precision $P_g(N)$ and brevity penalty BP:

$$BLEU = P_g(N) \times BP \quad (3)$$

where $P_g(N)$ is the geometric average of n-gram precisions

$$P_g(N) = \left(\prod_{n=1}^{N} p(n) \right)^{\frac{1}{N}} \quad (4)$$

The BLEU brevity penalty punishes the score if the translation length $len(T)$ is shorter than the reference length $len(R)$; it is:

$$BP = \min\left(1.0, e^{1-len(R)/len(T)}\right) \quad (5)$$

2.2 PORT

PORT has five components: precision, recall, strict brevity penalty (Chiang *et al.*, 2008), strict redundancy penalty (Chen and Kuhn, 2011) and an ordering measure *v*. The design of PORT is based on exhaustive experiments on a development data set. We do not have room here to give a rationale for all the choices we made when we designed PORT. However, a later section (3.3) reconsiders some of these design decisions.

2.2.1 Precision and Recall

The average precision and average recall used in PORT (unlike those used in BLEU) are the arithmetic average of n-gram precisions $P_a(N)$ and recalls $R_a(N)$:

$$P_a(N) = \frac{1}{N} \sum_{n=1}^{N} p(n) \quad (6)$$

$$R_a(N) = \frac{1}{N} \sum_{n=1}^{N} r(n) \quad (7)$$

We use two penalties to avoid too long or too short MT outputs. The first, the strict brevity penalty (SBP), is proposed in (Chiang *et al*., 2008). Let t_i be the translation of input sentence *i*, and let r_i be its reference. Set

$$SBP = \exp\left(1 - \frac{\sum_i |r_i|}{\sum_i \min\{|t_i|, |r_i|\}}\right) \quad (8)$$

The second is the strict redundancy penalty (SRP), proposed in (Chen and Kuhn, 2011):

$$SRP = \exp\left(1 - \frac{\sum_i \max\{|t_i|, |r_i|\}}{\sum_i |r_i|}\right) \quad (9)$$

To combine precision and recall, we tried four averaging methods: arithmetic (A), geometric (G), harmonic (H), and quadratic (Q) mean. If all of the values to be averaged are positive, the order is $min \le H \le G \le A \le Q \le max$, with equality holding if and only if all the values being averaged are equal. We chose the quadratic mean to combine precision and recall, as follows:

$$Qmean(N) = \sqrt{\frac{(P_a(N) \times SBP)^2 + (R_a(N) \times SRP)^2}{2}} \quad (10)$$

2.2.2 Ordering Measure

Word ordering measures for MT compare two permutations of the original source-language word sequence: the permutation represented by the sequence of corresponding words in the MT output, and the permutation in the reference. Several ordering measures have been integrated into MT evaluation metrics recently. Birch and Osborne (2011) use either Hamming Distance or Kendall's τ Distance (Kendall, 1938) in their metric LRscore, thus obtaining two versions of LRscore. Similarly, Isozaki *et al*. (2011) adopt either Kendall's τ Distance or Spearman's ρ (Spearman, 1904) distance in their metrics.

Our measure, *v*, is different from all of these. We use word alignment to compute the two permutations (LRscore also uses word alignment). The word alignment between the source input and reference is computed using GIZA++ (Och and Ney, 2003) beforehand with the default settings, then is refined with the heuristic *grow-diag-final-and*; the word alignment between the source input and the translation is generated by the decoder with the help of word alignment inside each phrase pair.

PORT uses permutations. These encode one-to-one relations but not one-to-many, many-to-one, many-to-many or null relations, all of which can occur in word alignments. We constrain the forbidden types of relation to become one-to-one, as in (Birch and Osborne, 2011). Thus, in a one-to-many alignment, the single source word is forced to align with the first target word; in a many-to-one alignment, monotone order is assumed for the target words; and source words originally aligned to null are aligned to the target word position just after the previous source word's target position.

After the normalization above, suppose we have two permutations for the same source n-word input. *E.g.*, let P_1 = reference, P_2 = hypothesis:

$$P_1\text{:}\ p_1^1\ p_1^2\ p_1^3\ p_1^4\ \cdots\ p_1^i\ \cdots\ p_1^n$$

$$P_2\text{:}\ p_2^1\ p_2^2\ p_2^3\ p_2^4\ \cdots\ p_2^i\ \cdots\ p_2^n$$

Here, each p_i^j is an integer denoting position in the original source (*e.g.*, p_1^1 = 7 means that the first word in P_1 is the 7th source word).

The ordering metric *v* is computed from two distance measures. The first is absolute permutation distance:

$$DIST_1(P_1, P_2) = \sum_{i=1}^{n} |p_1^i - p_2^i| \quad (11)$$

Let $$v_1 = 1 - \frac{DIST_1(P_1, P_2)}{n(n+1)/2} \quad (12)$$

v_1 ranges from 0 to 1; a larger value means more similarity between the two permutations. This metric is similar to Spearman's ρ (Spearman, 1904). However, we have found that ρ punishes long-distance reorderings too heavily. For instance, v_1 is more tolerant than ρ of the movement of "*recently*" in this example:

Ref: *Recently, I visited Paris*
Hyp: *I visited Paris recently*

Inspired by HMM word alignment (Vogel *et al*., 1996), our second distance measure is based on jump width. This punishes a sequence of words that moves a long distance with its internal order conserved, only once rather than on every word. In the following, only two groups of words have moved, so the jump width punishment is light:

Ref: *In the winter of 2010, I visited Paris*
Hyp: *I visited Paris in the winter of 2010*

So the second distance measure is

$$DIST_2(P_1, P_2) = \sum_{i=1}^{n} |(p_1^i - p_1^{i-1}) - (p_2^i - p_2^{i-1})| \quad (13)$$

where we set $p_1^0 = 0$ and $p_2^0 = 0$. Let

$$v_2 = 1 - \frac{DIST_2(P_1, P_2)}{n^2 - 1} \quad (14)$$

As with v_1, v_2 is also from 0 to 1, and larger values indicate more similar permutations. The ordering measure v_s is the harmonic mean of v_1 and v_2:

$$v_s = 2/(1/v_1 + 1/v_2) \ . \quad (15)$$

v_s in (15) is computed at segment level. For multiple references, we compute v_s for each, and then choose the biggest one as the segment level ordering similarity. We compute document level ordering with a weighted arithmetic mean:

$$v = \frac{\sum_{s=1}^{l} v_s \times len_s(R)}{\sum_{s=1}^{l} len_s(R)} \quad (16)$$

where l is the number of segments of the document, and *len*(*R*) is the length of the reference.

2.2.3 Combined Metric

Finally, *Qmean*(*N*) (Eq. (10) and the word ordering measure *v* are combined in a harmonic mean:

$$PORT = \frac{2}{1/Qmean(N) + 1/v^{\alpha}} \quad (17)$$

Here α is a free parameter that is tuned on held-out data. As it increases, the importance of the ordering measure *v* goes up. For our experiments, we tuned α on Chinese-English data, setting it to 0.25 and keeping this value for the other language pairs. The use of *v* means that unlike BLEU, PORT requires word alignment information.

3 Experiments

3.1 PORT as an Evaluation Metric

We studied PORT as an evaluation metric on WMT data; test sets include WMT 2008, WMT 2009, and WMT 2010 all-to-English, plus 2009, 2010 English-to-all submissions. The languages "all" ("xx" in Table 1) include French, Spanish, German and Czech. Table 1 summarizes the test set statistics. In order to compute the *v* part of PORT, we require source-target word alignments for the references and MT outputs. These aren't included in WMT data, so we compute them with GIZA++.

We used Spearman's rank correlation coefficient ρ to measure correlation of the metric with system-level human judgments of translation. The human judgment score is based on the "Rank" only, *i.e.*, how often the translations of the system were rated as better than those from other systems (Callison-Burch *et al.*, 2008). Thus, BLEU, METEOR, and PORT were evaluated on how well their rankings correlated with the human ones. For the segment level, we follow (Callison-Burch *et al.*, 2010) in using Kendall's rank correlation coefficient τ.

As shown in Table 2, we compared PORT with smoothed BLEU (*mteval-v13a*), and METEOR v1.0. Both BLEU and PORT perform matching of *n*-grams up to *n* = 4.

Set	Year	Lang.	#system	#sent-pair
Test1	2008	xx-en	43	7,804
Test2	2009	xx-en	45	15,087
Test3	2009	en-xx	40	14,563
Test4	2010	xx-en	53	15,964
Test5	2010	en-xx	32	18,508

Table 1: Statistics of the WMT dev and test sets.

	Into-En		Out-of-En	
Metric	sys.	seg.	sys.	seg.
BLEU	0.792	0.215	0.777	0.240
METEOR	**0.834**	0.231	**0.835**	0.225
PORT	0.801	**0.236**	0.804	**0.242**

Table 2: Correlations with human judgment on WMT

PORT achieved the best segment level correlation with human judgment on both the "into English" and "out of English" tasks. At the system level, PORT is better than BLEU, but not as good as METEOR. This is because we designed PORT to carry out tuning; we did not optimize its performance as an evaluation metric, but rather, to optimize system tuning performance. There are some other possible reasons why PORT did not outperform METEOR v1.0 at system level. Most WMT submissions involve language pairs with similar word order, so the ordering factor *v* in PORT won't play a big role. Also, *v* depends on source-target word alignments for reference and test sets. These alignments were performed by GIZA++ models trained on the test data only.

3.2 PORT as a Metric for Tuning

3.2.1 Experimental details

The first set of experiments to study PORT as a tuning metric involved Chinese-to-English (zh-en); there were two data conditions. The first is the *small data* condition where *FBIS*[2] is used to train the translation and reordering models. It contains 10.5M target word tokens. We trained two language models (LMs), which were combined loglinearly. The first is a 4-gram LM which is estimated on the target side of the texts used in the *large data* condition (below). The second is a 5-gram LM estimated on English *Gigaword*.

The *large data* condition uses training data from NIST[3] 2009 (Chinese-English track). All allowed bilingual corpora except *UN*, *Hong Kong Laws* and *Hong Kong Hansard* were used to train the translation model and reordering models. There are about 62.6M target word tokens. The same two LMs are used for *large data* as for *small data,* and the same development ("dev") and test sets are also used. The dev set comprised mainly data from the NIST 2005 test set, and also some balanced-genre web-text from NIST. Evaluation was performed on NIST 2006 and 2008. Four references were provided for all dev and test sets.

The third data condition is a French-to-English (fr-en). The parallel training data is from Canadian Hansard data, containing 59.3M word tokens. We used two LMs in loglinear combination: a 4-gram LM trained on the target side of the parallel training data, and the English *Gigaword* 5-gram LM. The dev set has 1992 sentences; the two test sets have 2140 and 2164 sentences respectively. There is one reference for all dev and test sets.

The fourth and fifth conditions involve German--English Europarl data. This parallel corpus contains 48.5M German tokens and 50.8M English tokens. We translate both German-to-English (de-en) and English-to-German (en-de). The two conditions both use an LM trained on the target side of the parallel training data, and de-en also uses the English *Gigaword* 5-gram LM. News test 2008 set is used as dev set; News test 2009, 2010, 2011 are used as test sets. One reference is provided for all dev and test sets.

All experiments were carried out with α in Eq. (17) set to 0.25, and involved only lowercase European-language text. They were performed with MOSES (Koehn *et al*., 2007), whose decoder includes lexicalized reordering, translation models, language models, and word and phrase penalties. Tuning was done with n-best MERT, which is available in MOSES. In all tuning experiments, both BLEU and PORT performed lower case matching of *n*-grams up to $n = 4$. We also conducted experiments with tuning on a version of BLEU that incorporates SBP (Chiang *et al*., 2008) as a baseline. The results of original IBM BLEU and BLEU with SBP were tied; to save space, we only report results for original IBM BLEU here.

3.2.2 Comparisons with automatic metrics

First, let us see if BLEU-tuning and PORT-tuning yield systems with different translations for the same input. The first row of Table 3 shows the percentage of identical sentence outputs for the two tuning types on test data. The second row shows the similarity of the two outputs at word-level (as measured by 1-TER): *e.g.*, for the two zh-en tasks, the two tuning types give systems whose outputs are about 25-30% different at the word level. By contrast, only about 10% of output words for fr-en differ for BLEU *vs*. PORT tuning.

	zh-en	zh-en	fr-en	de-en	en-de
	small	*large*	Hans	WMT	WMT
Same sent.	17.7%	13.5%	56.6%	23.7%	26.1%
1-TER	74.2	70.9	91.6	87.1	86.6

Table 3: Similarity of BLEU-tuned and PORT-tuned system outputs on test data.

		Evaluation metrics (%)			
Task	Tune	BLEU	MTR	1-TER	PORT
zh-en	BLEU	26.8	55.2	38.0	49.7
small	PORT	**27.2***	**55.7**	38.0	**50.0**
zh-en	BLEU	29.9	58.4	41.2	53.0
large	PORT	**30.3***	**59.0**	**42.0**	**53.2**
fr-en	BLEU	38.8	**69.8**	54.2	57.1
Hans	PORT	38.8	69.6	**54.6**	57.1
de-en	BLEU	20.1	55.6	38.4	39.6
WMT	PORT	**20.3**	**56.0**	38.4	**39.7**
en-de	BLEU	13.6	43.3	30.1	31.7
WMT	PORT	13.6	43.3	**30.7**	31.7

Table 4: Automatic evaluation scores on test data. * indicates the results are significantly better than the baseline (p<0.05).

[2] LDC2003E14

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

Table 4 shows translation quality for BLEU- and PORT-tuned systems, as assessed by automatic metrics. We employed BLEU4, METEOR (v1.0), TER (v0.7.25), and the new metric PORT. In the table, TER scores are presented as 1-TER to ensure that for all metrics, higher scores mean higher quality. All scores are averages over the relevant test sets. There are twenty comparisons in the table. Among these, there is one case (French-English assessed with METEOR) where BLEU outperforms PORT, there are seven ties, and there are twelve cases where PORT is better. Table 3 shows that fr-en outputs are very similar for both tuning types, so the fr-en results are perhaps less informative than the others. Overall, PORT tuning has a striking advantage over BLEU tuning.

Both (Liu *et al.*, 2011) and (Cer *et al.*, 2011) showed that with MERT, if you want the best possible score for a system's translations according to metric M, then you should tune with M. This doesn't appear to be true when PORT and BLEU tuning are compared in Table 4. For the two Chinese-to-English tasks in the table, PORT tuning yields a better BLEU score than BLEU tuning, with significance at $p < 0.05$. We are currently investigating why PORT tuning gives higher BLEU scores than BLEU tuning for Chinese-English and German-English. In internal tests we have found no systematic difference in dev-set BLEUs, so we speculate that PORT's emphasis on reordering yields models that generalize better for these two language pairs.

3.2.3 Human Evaluation

We conducted a human evaluation on outputs from BLEU- and PORT-tuned systems. The examples are randomly picked from all "to-English" conditions shown in Tables 3 & 4 (*i.e.*, all conditions except English-to-German).

We performed pairwise comparison of the translations produced by the system types as in (Callison-Burch *et al.*, 2010; Callison-Burch *et al.*, 2011). First, we eliminated examples where the reference had fewer than 10 words or more than 50 words, or where outputs of the BLEU-tuned and PORT-tuned systems were identical. The evaluators (colleagues not involved with this paper) objected to comparing two bad translations, so we then selected for human evaluation only translations that had high sentence-level (1-TER) scores. To be fair to both metrics, for each condition, we took the union of examples whose BLEU-tuned output was in the top n% of BLEU outputs and those whose PORT-tuned output was in the top n% of PORT outputs (based on (1-TER)). The value of n varied by condition: we chose the top 20% of zh-en *small*, top 20% of en-de, top 50% of fr-en and top 40% of zh-en *large*. We then randomly picked 450 of these examples to form the manual evaluation set. This set was split into 15 subsets, each containing 30 sentences. The first subset was used as a common set; each of the other 14 subsets was put in a separate file, to which the common set is added. Each of the 14 evaluators received one of these files, containing 60 examples (30 unique examples and 30 examples shared with the other evaluators). Within each example, BLEU-tuned and PORT-tuned outputs were presented in random order.

After receiving the 14 annotated files, we computed Fleiss's Kappa (Fleiss, 1971) on the common set to measure inter-annotator agreement, κ_{all}. Then, we excluded annotators one at a time to compute κ^i (Kappa score without *i*-th annotator, *i.e.*, from the other 13). Finally, we filtered out the files from the 4 annotators whose answers were most different from everybody else's: *i.e.*, annotators with the biggest $\kappa_{all} - \kappa^i$ values.

This left 10 files from 10 evaluators. We threw away the common set in each file, leaving 300 pairwise comparisons. Table 5 shows that the evaluators preferred the output from the PORT-tuned system 136 times, the output from the BLEU-tuned one 98 times, and had no preference the other 66 times. This indicates that there is a human preference for outputs from the PORT-tuned system over those from the BLEU-tuned system at the p<0.01 significance level (in cases where people prefer one of them).

PORT tuning seems to have a bigger advantage over BLEU tuning when the translation task is hard. Of the Table 5 language pairs, the one where PORT tuning helps most has the lowest BLEU in Table 4 (German-English); the one where it helps least in Table 5 has the highest BLEU in Table 4 (French-English). (Table 5 does <u>not</u> prove BLEU is superior to PORT for French-English tuning: statistically, the difference between 14 and 17 here is a tie). Maybe by picking examples for each condition that were the easiest for the system to translate (to make human evaluation easier), we mildly biased the results in Table 5 against PORT tuning. Another possible factor is reordering. PORT differs from BLEU partly in modeling long-distance reordering more accurately; English and French have similar word order, but the other two language pairs don't. The results in section 3.3 (below) for Qmean, a version of PORT without word ordering factor *v*, suggest *v* may be defined suboptimally for French-English.

	PORT win	BLEU win	equal	total
zh-en *small*	**19** **38.8%**	18 36.7%	12 24.5%	49
zh-en *large*	**69** **45.7%**	46 30.5%	36 23.8%	151
fr-en Hans	14 32.6%	**17** **39.5%**	12 27.9%	43
de-en WMT	**34** **59.7%**	17 29.8%	6 10.5%	57
All	**136** **45.3%**	98 32.7%	66 22.0%	300

Table 5: Human preference for outputs from PORT-tuned *vs*. BLEU-tuned system.

3.2.4 Computation time

A good tuning metric should run very fast; this is one of the advantages of BLEU. Table 6 shows the time required to score the 100-best hypotheses for the dev set for each data condition during MERT for BLEU and PORT in similar implementations. The average time of each iteration, including model loading, decoding, scoring and running MERT[4], is in brackets. PORT takes roughly 1.5 – 2.5 as long to compute as BLEU, which is reasonable for a tuning metric.

	zh-en *small*	zh-en *large*	fr-en Hans	de-en WMT	en-de WMT
BLEU	3 (13)	3 (17)	2 (19)	2 (20)	2 (11)
PORT	5 (21)	5 (24)	4 (28)	5 (28)	4 (15)

Table 6: Time to score 100-best hypotheses (average time per iteration) in minutes.

3.2.5 Robustness to word alignment errors

PORT, unlike BLEU, depends on word alignments. How does quality of word alignment between source and reference affect PORT tuning? We created a dev set from Chinese Tree Bank (CTB) hand-aligned data. It contains 588 sentences (13K target words), with one reference. We also ran GIZA++ to obtain its automatic word alignment, computed on CTB and FBIS. The AER of the GIZA++ word alignment on CTB is 0.32.

In Table 7, CTB is the dev set. The table shows tuning with BLEU, PORT with human word alignment (PORT + HWA), and PORT with GIZA++ word alignment (PORT + GWA); the condition is zh-en *small*. Despite the AER of 0.32 for automatic word alignment, PORT tuning works about as well with this alignment as for the gold standard CTB one. (The BLEU baseline in Table 7 differs from the Table 4 BLEU baseline because the dev sets differ).

Tune	BLEU	MTR	1-TER	PORT
BLEU	25.1	53.7	36.4	47.8
PORT + HWA	**25.3**	54.4	**37.0**	**48.2**
PORT + GWA	**25.3**	**54.6**	36.4	48.1

Table 7: PORT tuning - human & GIZA++ alignment

Task	Tune	BLEU	MTR	1-TER	PORT
zh-en	BLEU	26.8	55.2	38.0	49.7
small	PORT	**27.2**	**55.7**	38.0	**50.0**
	Qmean	26.8	55.3	**38.2**	49.8
zh-en	BLEU	29.9	58.4	41.2	53.0
large	PORT	**30.3**	**59.0**	**42.0**	**53.2**
	Qmean	30.2	58.5	41.8	53.1
fr-en	BLEU	38.8	**69.8**	54.2	57.1
Hans	PORT	38.8	69.6	**54.6**	57.1
	Qmean	38.8	**69.8**	**54.6**	57.1
de-en	BLEU	20.1	55.6	**38.4**	39.6
WMT	PORT	**20.3**	56.0	**38.4**	**39.7**
	Qmean	**20.3**	**56.3**	38.1	**39.7**
en-de	BLEU	13.6	43.3	30.1	31.7
WMT	PORT	13.6	43.3	**30.7**	31.7
	Qmean	13.6	**43.4**	30.3	31.7

Table 8: Impact of ordering measure *v* on PORT

3.3 Analysis

Now, we look at the details of PORT to see which of them are the most important. We do not have space here to describe all the details we studied, but we can describe some of them. *E.g.*, does the ordering measure *v* help tuning performance? To answer this, we introduce an intermediate metric. This is Qmean as in Eq. (10): PORT without the ordering measure. Table 8 compares tuning with BLEU, PORT, and Qmean. PORT outperforms Qmean on seven of the eight automatic scores shown for *small* and *large* Chinese-English.

[4] Our experiments are run on a cluster. The average time for an iteration includes queuing, and the speed of each node is slightly different, so bracketed times are only for reference.

However, for the European language pairs, PORT and Qmean seem to be tied. This may be because we optimized α in Eq. (18) for Chinese-English, making the influence of word ordering measure *v* in PORT too strong for the European pairs, which have similar word order.

Measure *v* seems to help Chinese-English tuning. What would results be on that language pair if we were to replace *v* in PORT with another ordering measure? Table 9 gives a partial answer, with Spearman's ρ and Kendall's τ replacing *v* with ρ or τ in PORT for the zh-en *small* condition (CTB with human word alignment is the dev set). The original definition of PORT seems preferable.

Tune	BLEU	METEOR	1-TER
BLEU	25.1	53.7	36.4
PORT(*v*)	**25.3**	**54.4**	**37.0**
PORT(ρ)	25.1	54.2	36.3
PORT(τ)	25.1	54.0	36.0

Table 9: Comparison of the ordering measure: replacing ν with ρ or τ in PORT.

Task	Tune	ordering measures		
		ρ	τ	*v*
NIST06	BLEU	0.979	0.926	0.915
	PORT	0.979	**0.928**	**0.917**
NIST08	BLEU	0.980	0.926	0.916
	PORT	**0.981**	**0.929**	**0.918**
CTB	BLEU	0.973	0.860	0.847
	PORT	**0.975**	**0.866**	**0.853**

Table 10: Ordering scores (ρ, τ and ν) for test sets NIST 2006, 2008 and CTB.

A related question is how much word ordering improvement we obtained from tuning with PORT. We evaluate Chinese-English word ordering with three measures: Spearman's ρ, Kendall's τ distance as applied to two permutations (see section 2.2.2) and our own measure *v*. Table 10 shows the effects of BLEU and PORT tuning on these three measures, for three test sets in the zh-en *large* condition. Reference alignments for CTB were created by humans, while the NIST06 and NIST08 reference alignments were produced with GIZA++. A large value of ρ, τ, or *v* implies outputs have ordering similar to that in the reference. From the table, we see that the PORT-tuned system yielded better word order than the BLEU-tuned system in all nine combinations of test sets and ordering measures. The advantage of PORT tuning is particularly noticeable on the most reliable test set: the hand-aligned CTB data.

What is the impact of the strict redundancy penalty on PORT? Note that in Table 8, even though Qmean has no ordering measure, it outperforms BLEU. Table 11 shows the BLEU brevity penalty (BP) and (number of matching 1- & 4- grams)/(number of total 1- & 4- grams) for the translations. The BLEU-tuned and Qmean-tuned systems generate similar numbers of matching n-grams, but Qmean-tuned systems produce fewer n-grams (thus, shorter translations). *E.g.*, for zh-en *small*, the BLEU-tuned system produced 44,677 1-grams (words), while the Qmean-trained system one produced 43,555 1-grams; both have about 32,000 1-grams matching the references. Thus, the Qmean translations have higher precision. We believe this is because of the strict redundancy penalty in Qmean. As usual, French-English is the outlier: the two outputs here are typically so similar that BLEU and Qmean tuning yield very similar n-gram statistics.

Task	Tune	1-gram	4-gram	BP
zh-en	BLEU	32055/44677	4603/39716	0.967
small	Qmean	31996/43555	4617/38595	0.962
zh-en	BLEU	34583/45370	5954/40410	0.972
large	Qmean	34369/44229	5987/39271	0.959
fr-en	BLEU	28141/40525	8654/34224	0.983
Hans	Qmean	28167/40798	8695/34495	0.990
de-en	BLEU	42380/75428	5151/66425	1.000
WMT	Qmean	42173/72403	5203/63401	0.968
en-de	BLEU	30326/62367	2261/54812	1.000
WMT	Qmean	30343/62092	2298/54537	0.997

Table 11: #matching-ngram/#total-ngram and BP score

4 Conclusions

In this paper, we have proposed a new tuning metric for SMT systems. PORT incorporates precision, recall, strict brevity penalty and strict redundancy penalty, plus a new word ordering measure *v*. As an evaluation metric, PORT performed better than BLEU at the system level and the segment level, and it was competitive with or slightly superior to METEOR at the segment level. Most important, our results show that PORT-tuned MT systems yield better translations than BLEU-tuned systems on several language pairs, according both to automatic metrics and human evaluations. In future work, we plan to tune the free parameter α for each language pair.

References

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of ACL Workshop on Intrinsic & Extrinsic Evaluation Measures for Machine Translation and/or Summarization.

A. Birch and M. Osborne. 2011. Reordering Metrics for MT. In Proceedings of ACL.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz and J. Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In Proceedings of WMT.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In Proceedings of EACL.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki and O. Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In Proceedings of WMT.

C. Callison-Burch, P. Koehn, C. Monz and O. Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In Proceedings of WMT.

D. Cer, D. Jurafsky and C. Manning. 2010. The Best Lexical Metric for Phrase-Based Statistical MT System Optimization. In Proceedings of NAACL.

Y. S. Chan and H. T. Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In Proceedings of ACL.

B. Chen and R. Kuhn. 2011. AMBER: A Modified BLEU, Enhanced Ranking Metric. In: Proceedings of WMT. Edinburgh, UK. July.

D. Chiang, S. DeNeefe, Y. S. Chan, and H. T. Ng. 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms. In Proceedings of EMNLP, pages 610–619.

M. Denkowski and A. Lavie. 2010. Meteor-next and the meteor paraphrase tables: Improved evaluation support for five target languages. In Proceedings of the Joint Fifth Workshop on SMT and MetricsMATR, pages 314–317.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of HLT.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. In *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382.

Y. He, J. Du, A. Way and J. van Genabith. 2010. The DCU dependency-based metric in WMT-MetricsMATR 2010. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 324–328.

H. Isozaki, T. Hirao, K. Duh, K. Sudoh, H. Tsukada. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. In Proceedings of EMNLP.

M. Kendall. 1938. A New Measure of Rank Correlation. In Biometrika, 30 (1–2), pp. 81–89.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of ACL, pp. 177-180, Prague, Czech Republic.

A. Lavie and M. J. Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. Machine Translation, 23.

C. Liu, D. Dahlmeier, and H. T. Ng. 2010. TESLA: Translation evaluation of sentences with linear-programming-based analysis. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 329–334.

C. Liu, D. Dahlmeier, and H. T. Ng. 2011. Better evaluation metrics lead to better machine translation. In Proceedings of EMNLP.

C. Lo and D. Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In Proceedings of ACL.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In Proceedings of ACL-2003. Sapporo, Japan.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. In Computational Linguistics, 29, pp. 19–51.

S. Pado, M. Galley, D. Jurafsky, and C.D. Manning. 2009. Robust machine translation evaluation with entailment features. In Proceedings of ACL-IJCNLP.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of ACL.

K. Parton, J. Tetreault, N. Madnani and M. Chodorow. 2011. E-rating Machine Translation. In Proceedings of WMT.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of Association for Machine Translation in the Americas.

M. Snover, N. Madnani, B. Dorr, and R. Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece.

C. Spearman. 1904. The proof and measurement of association between two things. In American Journal of Psychology, 15, pp. 72–101.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM based word alignment in statistical translation. In Proceedings of COLING.