LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors

Aaron L. F. HAN *Derek F. WONG* *Lidia S. CHAO*
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory
Department of Computer and Information Science
University of Macau
Macau S.A.R., China
hanlifengaaron@gmail.com, derekfw@umac.mo, lidiasc@umac.mo

ABSTRACT

In the conventional evaluation metrics of machine translation, considering less information about the translations usually makes the result not reasonable and low correlation with human judgments. On the other hand, using many external linguistic resources and tools (e.g. Part-of-speech tagging, morpheme, stemming, and synonyms) makes the metrics complicated, time-consuming and not universal due to that different languages have the different linguistic features. This paper proposes a novel evaluation metric employing rich and augmented factors without relying on any additional resource or tool. Experiments show that this novel metric yields the state-of-the-art correlation with human judgments compared with classic metrics BLEU, TER, Meteor-1.3 and two latest metrics (AMBER and MP4IBM1), which proves it a robust one by employing a feature-rich and model-independent approach.

KEYWORDS : Machine translation, Evaluation metric, Context-dependent *n*-gram alignment, Modified length penalty, Precision, Recall.

Proceedings of COLING 2012: Posters, pages 441–450,
COLING 2012, Mumbai, December 2012.

1 Introduction

Since IBM proposed and realized the system of BLEU (Papineni et al., 2002) as the automatic metric for Machine Translation (MT) evaluation, many other methods have been proposed to revise or improve it. BLEU considered the *n*-gram precision and the penalty for translation which is shorter than that of references. NIST (Doddington, 2002) added the information weight into evaluation factors. Meteor (Banerjee and Lavie, 2005) proposed an alternative way of calculating matched chunks to describe the *n*-gram matching degree between machine translations and reference translations. Wong and Kit (2008) introduced position difference in the evaluation metric. Other evaluation metrics, such as TER (Snover et al., 2006), the modified Meteor-1.3 (Denkowski and Lavie, 2011), and MP4IBM1 (Popovic et al., 2011) are also used in the literature. AMBER (Chen and Kuhn, 2011) declares a modified version of BLEU and attaches more kinds of penalty coefficients, combining the *n*-gram precision and recall with the arithmetic average of F-measure. In order to distinguish the reliability of different MT evaluation metrics, people used to apply the Spearman correlation coefficient for evaluation tasks in the workshop of statistical machine translation (WMT) for Association of Computational Linguistics (ACL) (Callison-Burch et al., 2011; Callison-Burch et al., 2010; Callison-Burch et al., 2009, 2008).

2 Related work

Some MT evaluation metrics are designed with the part-of-speech (POS) consideration using the linguistic tools, parser or POS tagger, during the words matching period between system-output and reference sentences. Machacek and Bojar (2011) proposed SemPOS metric, which is based on the Czech-targeted work by Kos and Bojar (2009). SemPOS conducts a deep-syntactic analysis of the target language with a modified version of similarity measure from the general overlapping method (Gimenez and Marquez, 2007). However, SemPOS only focuses on the English and Czech words and achieves no contribution for other language pairs.

To reduce the human tasks during the evaluation, the methodologies that do not need reference translations are growing up. MP4IBM1 (Popovic et al., 2011) used IBM1 model to calculate scores based on morphemes, POS (4-grams) and lexicon probabilities. MP4IBM1 is not a simple model although it is reference independent. For instance, it needs large parallel bilingual corpus, POS taggers (requesting the details about verb tenses, cases, number, gender, etc.) and other tools for splitting words into morphemes. It performed well on the corpus with English as source language following the metric TESLA (Dahlmeier et al., 2011) but got very poor correlation when English is the target language. For example, it gained the system-level correlation score 0.12 and 0.08 respectively on the Spanish-to-English and French-to-English MT evaluation tasks (Callison-Burch et al., 2011) and these two scores mean nearly no correlation with human judgments.

Reordering errors play an important role in the translation for distant language pairs (Isozaki et al., 2010). But BLEU and many other metrics are both insensitive to reordering phenomena and relatively time-consuming to compute (Talbot et al., 2011). Snover et al. (2006) introduced Translation Edit Rate (TER) and the possible edits include the insertion, deletion, and substitution of words as well as sequences allowing phrase movements without large penalties. Isozaki et al. (2010) paid attention to word order on the evaluation between Japanese and English. Wong and Kit (2008) designed position difference factor during the alignment of words between

reference translations and candidate outputs, but it only selects the candidate word that has the nearest position in principle.

Different words or phrases can express the same meanings, so it is considered commonly in the literature to refer auxiliary synonyms libraries during the evaluation task. Meteor (Banerjee and Lavie, 2005) is based on unigram match on the words and their stems also with additional synonyms database. Meteor-1.3 (Denkowski and Lavie, 2011), an improved version of Meteor, includes ranking and adequacy versions and has overcome some weaknesses of previous version such as noise in the paraphrase matching, lack of punctuation handling and discrimination between word types (Callison-Burch et al., 2011).

3 Proposed metric

According to the analysis above, we see that in the previous MT evaluation metrics, there are mainly two problems: either presenting incomprehensive factors (e.g. BLEU focus on precision) or relying on many external tools and databases. The first aspect makes the metrics result in unreasonable judgments. The second weakness makes the MT evaluation metric complicated, time-consuming and not universal for different languages. To address these weaknesses, a novel metric LEPOR[1] is proposed in this research, which is designed to take thorough variables into account (including modified factors) and does not need any extra dataset or tool. These are aimed at both improving the practical performance of the automatic metric and the easily operating of the program. LEPOR focuses on combining two modified factor (sentence length penalty, *n*-gram position difference penalty) and two classic methodologies (precision and recall). LEPOR score is calculated by:

$$LEPOR = LP \times NPosPenal \times Harmonic(\alpha R, \beta P) \quad (1)$$

The detailed introductions and designs of the features are shown below.

3.1 Design of LEPOR metric

3.1.1 Length penalty:

In the Eq. (1), LP means Length penalty, which is defined to embrace the penalty for both longer and shorter system outputs compared with the reference translations, and it is calculated as:

$$LP = \begin{cases} e^{1-\frac{r}{c}} & if \ \ c < r \\ 1 & if \ \ c = r \\ e^{1-\frac{c}{r}} & if \ \ c > r \end{cases} \quad (2)$$

where c and r mean the sentence length of output candidate translation and reference translation respectively. As seen in Eq. (2), when the output length of sentence is equal to that of the reference one, LP will be one which means no penalty. However, when the output length c is larger or smaller than that of the reference one, LP will be little than one which means a penalty on the evaluation value of LEPOR. And according to the characteristics of exponential function mathematically, the larger of numerical difference between c and r, the smaller the value of LP will be.

[1] LEPOR: **Le**ngth Penalty, **P**recision, *n*-gram **Po**sition difference Penalty and **R**ecall**.**

3.1.2 N-gram position difference penalty:

In the Eq. (1), the $NPosPenal$ is defined as:

$$NPosPenal = e^{-NPD} \tag{3}$$

where NPD means *n*-gram position difference penalty. The $NPosPenal$ value is designed to compare the words order in the sentences between reference translation and output translation. The $NPosPenal$ value is normalized. Thus we can take all MT systems into account whose effective NPD value varies between 0 and 1, and when N equals 0, the $NPosPenal$ will be 1 which represents no penalty and is quite reasonable. When the NPD increases from 0 to 1, the $NPosPenal$ value decreases from 1 to $1/e$ based on the mathematical analysis. Consequently, the final LEPOR value will be smaller. According to this thought, the NPD is defined as:

$$NPD = \frac{1}{Length_{output}} \sum_{i=1}^{Length_{output}} |PD_i| \tag{4}$$

where $Length_{output}$ represents the length of system output sentence and PD_i means the *n*-gram position D-value (difference value) of aligned words between output and reference sentences. Every word from both output translation and reference should be aligned only once (one-to-one alignment). Case (upper or lower) is irrelevant. When there is no match, the value of PD_i will be zero as default for this output translation word.

Output Sentence: $W = \{w_1 w_2 w_3 \dots w_{m_1} | \, m_1 \in (1, \infty)\}$
Reference Sentence: $W^r = \{w_1^r w_2^r w_3^r \dots w_{m_2}^r | \, m_2 \in (1, \infty)\}$
$\forall x \in (1, \infty)$, The Alignment of word w_x:

---

if $\forall y \in (1, \infty): w_x \neq w_y^r$ // ∀ *means for each,* ∃ *means there is/are*
  $(w_x \rightarrow \emptyset)$; // → *shows the alignment*
elseif $\exists! \, y \in (1, \infty): w_x = w_y^r$ // ∃! *means there exists exactly one*
  $(w_x \rightarrow w_y^r)$;
elseif $\exists y_1, y_2 \in (1, \infty): (w_x = w_{y_1}^r) \wedge (w_x = w_{y_2}^r)$ // ∧ *is logical conjunction, and*
  foreach $k \in (-n, -1) \cup (1, n)$
    foreach $j \in (-n, -1) \cup (1, n)$
      if $\exists k_1, k_2, j_1, j_2: (w_{x+k_1} = w_{y_1+j_1}^r) \wedge (w_{x+k_2} = w_{y_2+j_2}^r)$
        if $Distance(w_x, w_{y_1}^r) \leq Distance(w_x, w_{y_2}^r)$
          $(w_x \rightarrow w_{y_1}^r)$;
        else
          $(w_x \rightarrow w_{y_2}^r)$;
      elseif $\exists k_1, j_1: (w_{x+k_1} = w_{y_1+j_1}^r) \wedge (\forall k_2, j_2: (w_{x+k_2} \neq w_{y_2+j_2}^r))$
        $(w_x \rightarrow w_{y_1}^r)$;
      else // *i.e.* $\forall k_1, k_2, j_1, j_2: (w_{x+k_1} \neq w_{y_1+j_1}^r) \wedge (w_{x+k_2} \neq w_{y_2+j_2}^r)$
        if $Distance(w_x, w_{y_1}^r) \leq Distance(w_x, w_{y_2}^r)$
          $(w_x \rightarrow w_{y_1}^r)$;
        else
          $(w_x \rightarrow w_{y_2}^r)$;
else // *when more than two candidates, the selection steps are similar as above*

FIGURE 1 – Context-dependent *n*-gram word alignment algorithm

To calculate the NPD value, there are two steps: aligning and calculating. To begin with, the Context-dependent *n*-gram Word Alignment task: we take the context-dependent factor into consideration and assign higher priority on it, which means we take into account the surrounding context (neighbouring words) of the potential word to select a better matching pairs between the output and the reference. If there are both nearby matching or there is no matched context around the potential words pairs, then we consider the nearest matching to align as a backup choice. The alignment direction is from output sentence to the reference translations. Assuming that w_x represents the current word in output sentence and w_{x+k} (or w_{x+k_i}) means the kth word to the previous ($k < 0$) or following ($k > 0$). While w_y^r (or $w_{y_i}^r$) means the words matching w_x in the references, and w_{y+j}^r (or $w_{y_i+j_i}^r$) has the similar meaning as w_{x+k} but in reference sentence. $Distance$ is the position difference value between the matching words in outputs and references. The operation process and pseudo code of the context-dependent *n*-gram word alignment algorithm are shown in Figure 1 (with "→" as the alignment). Taking 2-gram (n = 2) as an example, let's see explanation in Figure 2. We label each word with its absolute position, then according to the context-dependent *n*-gram method, the first word "A" in the output sentence has no nearby matching with the beginning word "A" in reference, so it is aligned to the fifth word "a" due to their matched neighbor words "stone" and "on" within one (≤ 2) and two (≤ 2) steps respectively away from current position. Then the fourth word "a" in the output will align the first word "A" of the reference due to the one-to-one alignment. The alignments of other words in the output are obvious.

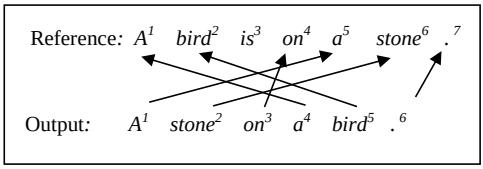
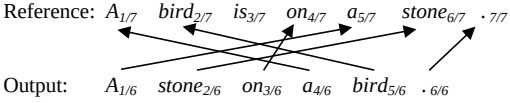


FIGURE 2 – Example of context-dependent *n*-gram word alignment

In the second step (calculating step), we label each word with its position number divided by the corresponding sentence length for normalization, and then using the Eq. (4) to finish the calculation. We also use the example in Figure 2 for the NPD introduction:

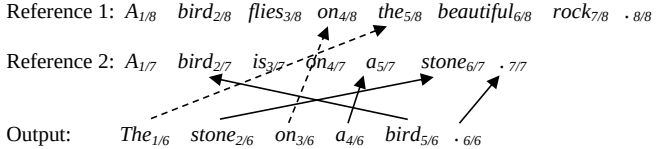


$$NPD = \frac{1}{6} \times [\left|\frac{1}{6} - \frac{5}{7}\right| + \left|\frac{2}{6} - \frac{6}{7}\right| + \left|\frac{3}{6} - \frac{4}{7}\right| + \left|\frac{4}{6} - \frac{1}{7}\right| + \left|\frac{5}{6} - \frac{2}{7}\right|] = \frac{2}{7}$$

In the example, when we label the word position of output sentence we divide the numerical position (from 1 to 6) of the current word by the reference sentence length 6. Similar way is applied in labeling the reference sentence. After we get the NPD value, using the Eq. (3), the values of $NPosPenal$ can be calculated.

When there is multi-references (more than one reference sentence), for instance 2 references, we take the similar approach but with a minor change. The alignment direction is reminded the same (from output to reference), and the candidate alignments that have nearby matching words also

embrace higher priority. If the matching words from Reference-1 and Reference-2 both have the nearby matching with the output word, then we select the candidate alignment that makes the final NPD value smaller. See below (also 2-gram) for explanation:



The beginning output words "the" and "stone" are aligned simply for the single matching. The output word "on" has nearby matching with the word "on" both in Reference-1 and Reference-2, due to the words "the" (second to previous) and "a" (first in the following) respectively. Then we should select its alignment to the word "on" in Reference-1, not Reference-2 for the further reason $\left|\frac{3}{6}-\frac{4}{8}\right| < \left|\frac{3}{6}-\frac{4}{7}\right|$ and this selection will obtain a smaller NPD value. The remaining two words "a" and "bird" in output sentence are aligned using the same principle.

3.1.3 Precision and recall:

Precision is designed to reflect the accurate rate of outputs while recall means the loyalty to the references. In the Eq. (1), $Harmonic(\alpha R, \beta P)$ means the Harmonic mean of αR and βP and is calculated as:

$$Harmonic(\alpha R, \beta P) = (\alpha + \beta)/(\frac{\alpha}{R} + \frac{\beta}{P}) \quad (5)$$

where α and β are two parameters we designed to adjust the weight of R (recall) and P (precision). The two metrics are calculated by:

$$P = \frac{common\_num}{system\_length} \quad (6)$$

$$R = \frac{common\_num}{reference\_length} \quad (7)$$

where $common\_num$ represents the number of aligned (matching) words and marks appearing both in translations and references, $system\_length$ and $reference\_length$ specify the sentence length of system output and reference respectively (Melamed et al., 2003). After we finish the above steps, taking all the variables into Eq. (1), we can calculate the final LEPOR score, and higher LEPOR value means the output sentence is closer to the references.

3.2 Two variants of system-level LEPOR

We have introduced the computation of LEPOR on single output sentence, and we should consider a proper way to calculate the LEPOR value when the cases turn into document (system) level. We perform the system-level LEPOR with two different variants LEPOR-A and LEPOR-B as follow.

$$\overline{LEPOR_A} = \frac{1}{SentNum}\sum_{i=1}^{SentNum} LEPOR_i \quad (8)$$

$$\overline{LEPOR_B} = \overline{LP} \times \overline{PosPenalty} \times \overline{Harmonic(\alpha R, \beta P)} \quad (9)$$

$$\overline{LP} = \frac{1}{SentNum}\sum_{i=1}^{SentNum} LP_i \quad (10)$$

$$\overline{PosPenalty} = \frac{1}{SentNum}\sum_{i=1}^{SentNum} PosPenalty_i \quad (11)$$

$$\overline{Harmonic(\alpha R, \beta P)} = \frac{1}{SentNum}\sum_{i=1}^{SentNum} Harmonic(\alpha R, \beta P)_i \quad (12)$$

where $\overline{LEPOR_A}$ and $\overline{LEPOR_B}$ in Eq. (8) and (9) both represent the system-level score of LEPOR, $SentNum$ specifies the sentence number of the test document, and $LEPOR_i$ in Eq. (8) means the LEPOR value of the ith sentence. As shown above, $\overline{LEPOR_A}$ is calculated by the arithmetic mean of LEPOR value of each sentence. On the other hand, $\overline{LEPOR_B}$ is designed from another perspective, which reflects the system-level values of three factors in LEPOR. To compute $\overline{LEPOR_B}$ using Eq. (9), we should firstly calculate the three system-level factors $\overline{LP}$, $\overline{PosPenalty}$ and $\overline{Harmonic(\alpha R, \beta P)}$ using Eq. (10) to Eq. (12), which are calculated in a similar way to that of $\overline{LEPOR_A}$ by the arithmetic mean.

4 Experiments and comparisons

We trained LEPOR on the public ACL WMT 2008[2] data (EN: English, ES: Spanish, DE: German, FR: French and CZ: Czech). The parameters α and β are set to 9 and 1 respectively for all languages pairs except that $\alpha = 1$ and $\beta = 9$ are used for Czech-English translations. For the context-dependent *n*-gram word alignment, we adjust *n* as 2 on all the corpora meaning that we consider both the proceeding and following two words as the context information.

We use the MT evaluation corpora from 2011 ACL WMT[3] for testing. The tested eight corpora are English-to-*other* (Spanish, German, French and Czech) and *other*-to-English news text. Following a common practice (e.g. the TER metric was proposed by the comparison with BLEU and Meteor, the AMBER metric compared with BLEU and Meteor-1.0, the MP4IBM1 compared with BLEU), we compare the scoring results by LEPOR against the three "gold standard" metrics BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and Meteor (version 1.3) (Denkowski and Lavie, 2011). In addition, we select the latest AMBER (modified version of BLEU) (Chen and Kuhn, 2011) and MP4IBM1 (without reference translation) (Popovic et al., 2011) as representatives to examine the quality of LEPOR in this study. The correlation results are shown in Table 1.

Evaluation system	**Correlation Score with Human Judgment**								
	other-to-English				English-to-other				Mean score
	CZ-EN	DE-EN	ES-EN	FR-EN	EN-CZ	EN-DE	EN-ES	EN-FR	
LEPOR-B	0.93	0.62	0.96	0.89	0.71	0.36	0.88	0.84	**0.77**
LEPOR-A	0.95	0.61	0.96	0.88	0.68	0.35	0.89	0.83	**0.77**
AMBER	0.88	0.59	0.86	0.95	0.56	0.53	0.87	0.84	0.76
Meteor-1.3-RANK	0.91	0.71	0.88	0.93	0.65	0.30	0.74	0.85	0.75
BLEU	0.88	0.48	0.90	0.85	0.65	0.44	0.87	0.86	0.74
TER	0.83	0.33	0.89	0.77	0.50	0.12	0.81	0.84	0.64
MP4IBM1	0.91	0.56	0.12	0.08	0.76	0.91	0.71	0.61	0.58

TABLE 1 – Spearman correlation scores of the metrics on eight corpora.

[2] http://www.statmt.org/wmt08/
[3] http://www.statmt.org/wmt11/

The metrics are ranked by their mean (hybrid) performance on the eight corpora from the best to the worst. Table 1 shows that LEPOR-A and LEPOR-B obtained the highest scores among the metrics, and LEPOR-B yields the best results by mean scores. BLEU, AMBER (modified version of BLEU) and Meteor-1.3 perform unsteady with better correlation on some translation languages and worse on others, resulting in medium level generally. TER and MP4IBM1 get the worst scores by the mean correlation. The result proves that LEPOR is a robust metric in all cases by constructing augmented features and also a concise and independent model without using any external tool and database (e.g. AMBER using auxiliary tokenize tool for stem, prefix and suffix matching; Meteor using word stems and synonyms databases etc.). MP4IBM1 does not need the reference translations instead using the POS tagger and word morphemes, but the current correlation is low. Table 1 also releases the information that although the test metrics yield high system-level correlations with human judgments on certain language pairs (e.g. all correlations above 0.83 on Czech-to-English), they are far from satisfactory by synthetically mean scores on total eight corpora (currently spanning from 0.58 to 0.77 only) and there is clearly a potential for further improvement.

Conclusion and perspectives

As we know that better evaluation metrics will be helpful to leading to better machine translations (Liu et al., 2011). This paper proposes a novel automatic evaluation metric LEPOR, which employs rich and augmented evaluation factors such that the result is close to human assessments. From the empirical results, we found that LEPOR can achieve better results compared with the state-of-the-art MT evaluation metrics, including BLEU, TER, Meteor-1.3 and the recently proposed AMBER and MP4IBM1. LEPOR gives good outputs generally on all the testing languages, with the state-of-the-art performance on the Czech-to-English, Spanish-to-English, English-to-Spanish and the mean correlation score without relying on any extra tool and data sources. Actually the correlation coefficient value of LEPOR can be further improved through the adjustment of the parameters α (weighting of recall) and β (weighting of precision), as well as the number of words concurrences used in the context-dependent *n*-gram position difference penalty.

Some further works are worth doing in the future. First, test on synonym thesaurus: in most cases, translation of word can be re-expressed in different ways, such as multi-words or paraphrases. It will certainly be helpful to the correlation score if a synonym thesaurus is available during the matching of words (Wong et al., 2009). Secondly, evaluate the effectiveness on languages of different topologies: in this paper we use the corpora that cover five languages English, Spanish, Czech, French, and German. That will be good if proposed metric can be tested on more pairs of languages from different families such as Portuguese, Japanese and Chinese etc. Thirdly, employ the Multi-references: another way to replace the synonym is the use of multi-references for evaluation. This can reduce the deviation when calculating the mechanical translation quality. The results will be more reason if we use multi-references. Lastly, in this work, we focus on the lexical information and how can we go beyond this is another direction that worth for further studies.

Acknowledgments

This work is partially supported by the Research Committee of University of Macau, and Science and Technology Development Fund of Macau under the grants UL019B/09-Y3/EEE/LYP01/FST, and 057/2009/A2. The authors are also grateful to the ACL's special interest group in machine translation (SIGMT) association for the offering of the data.

References

Banerjee, S. and Lavie, A. (2005). Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization of the Association for Computational Linguistics*, pages 65-72, Prague, Czech Republic.

Callison-Bruch, C., Koehn, P., Monz, C. and Zaidan, O. F. (2011). Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine translation of the Association for Computational Linguistics(ACL-WMT)*, pages 22-64, Edinburgh, Scotland, UK.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M. and Zaidan, O. F. (2010). Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the 5th Workshop on Statistical Machine Translation, Stroudsburg, Association for Computational Linguistics(ACL-WMT)*, pages 17-53, PA, USA.

Callison-Burch, C., Koehn, P., Monz,C. and Schroeder, J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation, Association for Computational Linguistics (ACL-WMT)*, pages 70-106, Columbus, Ohio, USA.

Callison-Burch, C., Koehn, P., Monz,C. and Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the 4th Workshop on Statistical Machine Translation, the European Chapter of Association for Computational Linguistics (EACL-WMT)*, pages 1-28, Athens, Greece.

Chen, B. and Kuhn, R. (2011). Amber: A modified bleu, enhanced ranking metric. In *Proceedings of the Sixth Workshop on Statistical Machine translation of the Association for Computational Linguistics(ACL-WMT)*, pages 71-77, Edinburgh, Scotland, UK.

Dahlmeier, D., Liu, C. and Ng, H. T. (2011). TESLA at WMT2011: Translation evaluation and tunable metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics (ACL-WMT)*, pages 78-84, Edinburgh, Scotland, UK.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine translation of the Association for Computational Linguistics(ACL-WMT)*, pages 85-91, Edinburgh, Scotland, UK.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research(HLT 2002)*, pages 138-145, San Diego, California, USA.

Gimenez, J. and Marquez, L. (2007). Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics (ACL-WMT)*, pages 256–264, Prague.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H. (2010). Automatic Evaluation of Translation Quality for Distant Language Pairs, In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics(EMNLP)*, pages 944-952, MIT, Massachusetts, USA.

Kos, K. and Bojar, O. (2009). Evaluation of Machine Translation Metrics for Czech as the

Target Language. *Prague Bull. Math. Linguistics*, Vol. 92: 135-148.

Liu, C., Dahlmeier, D. and Ng, H. T. (2011). Better evaluation metrics lead to better machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics(EMNLP)*, pages 375-384, Stroudsburg, PA, USA.

Melamed, I. D., Green, R., Turian, J. P. (2003). Precision and recall of machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, pages 61-63, Edmonton, Canada.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311-318, Philadelphia, PA, USA.

Popovic, M., Vilar, D., Avramidis, E. and Burchardt, A. (2011). Evaluation without references: IBM1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics (ACL-WMT)*, pages 99-103, Edinburgh, Scotland, UK.

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 Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223-231, Boston, USA.

Talbot, D., Kazawa, H., Ichikawa, H., Katz-Brown, J., Seno, M. and Och, F. (2011). A Lightweight Evaluation Framework for Machine Translation Reordering. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics (ACL-WMT)*, pages 12-21, Edinburgh, Scotland, UK.

Wong, B. T-M and Kit, C. (2008). Word choice and word position for automatic MT evaluation. In *Workshop: MetricsMATR of the Association for Machine Translation in the Americas (AMTA)*, short paper, 3 pages, Waikiki, Hawai'I, USA.

Wong, F., Chao, S., Hao, C. C. and Leong, K. S. (2009). A Maximum Entropy (ME) Based Translation Model for Chinese Characters Conversion, *Journal of Advances in Computational Linguistics, Research in Computer Science*, 41: 267-276.