

# A Linguistically Motivated MT Evaluation System Based on SVM Regression

Muyun Yang, Shuqi Sun, Jufeng Li, Sheng Li, Zhao Tiejun  
School of Computer Science and Technology  
Harbin Institute of Technology  
Harbin, China

{ymy, sqsun, jfli, tjzhao}@mtlab.hit.edu.cn  
lisheng@hit.edu.cn

## Abstract

This paper describes the automatic MT evaluation system by Machine Intelligence and Translation Lab. of Harbin Institute of Technology, for NIST MetricsMATR 2008 evaluation. The system is based on SVM regression and employed many linguistic features. Machine-learning based automatic MT evaluation has recently gained more attention in NLP research. And feature mining is essential for ML. In this work, a novel feature, letter-based BLEU, is introduced, and it turns out to be of good capability to correlate with human assessments and contributes to the system's final performance.

## 1 Motivation

Automatic MT evaluation has recently gained more attention in NLP research. Especially, the exploitation of linguistic information becomes an attracting direction. Since BLEU (Papineni et al., 2001) showed its excellent performance and was examined in all aspects, it was found that only lexical information is likely not sufficient, and that richer linguistic knowledge can capture the aspects of the quality of a translation efficiently. For example, some approaches make use of paraphrasing (Zhou et al., 2006; Kauchak and Barzilay, 2006; Owczarzak et al., 2006), while some approaches take advantages of syntax information (Liu and Gildea 2005; Amigó et al., 2006; Mehay and Brew, 2007; Owczarzak et al., 2007).

Combining metrics together is also a prevalent way to avoid biased evaluation caused by single metric. Giménez and Márquez (2007) showed that compared with metrics limited in lexical dimension, metrics integrating deep linguistic information will be more reliable. In the combina-

tion task, machine learning method is usually applied. Kulesza and Shieber (2004) proposed a SVM classifier based on confidence score, which takes the distance between feature vector and the decision surface as the measure of the MT system's output. Albrecht and Hwa (2007) extended the work of Kulesza and Shieber (2004) and adopted regression SVM, showing that the higher classification accuracy does not guarantee higher correlation with human assessments, and regarding the MT evaluation as a regression problem is more appropriate.

As for ranking, several works (Ye et al. 2007; Duh 2008) argue that it is easier and more natural to rank machine translation systems instead of giving a numeric score to them. Although there were studies showing that ranking can achieve a comparable performance against regression approaches, there are still some difficulties to overcome. For instance, rankings are only available among sentences translated from the same source sentence, and it is meaningless to compare sentences that from different source sentences. This increases the difficulty of calculating sentence level correlation with human assessments.

In consideration of above, we adopted a machine learning approach using Support Vector Machine together with some carefully selected features from several linguistic levels. In section 2, we will introduce this approach in detail. Section 3 will show some experiment results to show its performance. A there will be a discussion in section 4.

## 2 Methodology

In this section, we will first introduce SVM regression based automatic MT evaluation, and then describe the feature set used in our evaluation system.

## 2.1 SVM Regression

Suppose we are given training data  $\{(x_1, y_1), \dots, (x_l, y_l)\} \subset X \times \mathbb{R}$ , where  $X$  denotes the space of the input patterns (e.g.  $X = \mathbb{R}^d$ ). In  $\varepsilon$ -SV regression (Vapnik, 1995), the goal is finding a function  $f(x)$  that has at most  $\varepsilon$  deviation from the actually obtained targets  $y_i$  for all the training data, and at the same time is as flat as possible (Smola and Schölkopf, 2001). In the case of linear regression,  $f(x)$  is formulated as:

$$f(x) = \langle w, x \rangle + b \text{ with } x \in X, b \in \mathbb{R}$$

We can write this problem as a convex optimization problem with a ‘soft-margin’ loss function (Bennett and Mangasarian, 1992) as the formulation stated in Vapnik (1995):

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \\ \text{s.t.} \quad & y_i - \langle w, x \rangle - b \leq \varepsilon + \xi_i \\ & \langle w, x \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \end{aligned}$$

where  $\xi$ ,  $\xi^*$  describe the extent of training error and  $C > 0$  is the trade-off between training error and margin.

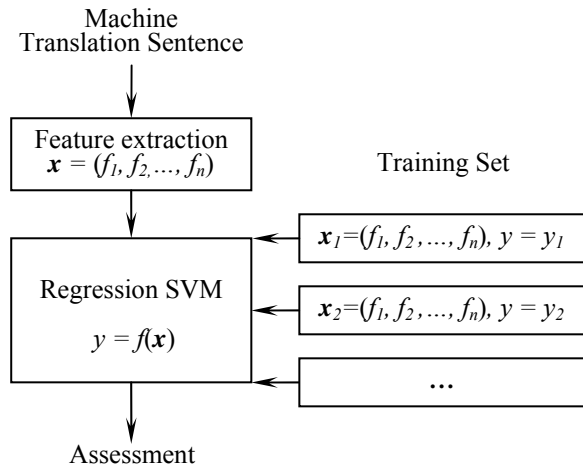


Figure 1: SVM regression based framework of automatic MT evaluation metric

In MT evaluation, the mission is to train a function using feature vectors extracted from translations and human assessments. When given a new translation, one should first construct a feature vector from it, and then use the function obtained in the training phase to calculate a score which representing the quality of the translation approximately. Figure 1 gives a general framework of this task.

Feature selection is essential to all machine learning methods. A good feature set will model the target object well, and will be less prone to over-fitting. Though SVM has the characteristic of less sensitivity to the detection of the feature space good generalization capability, feature selection and feature mining are still playing an important role. The reason lies in three aspects:

1. Trade off between representativeness and computational cost. High-level linguistic features are computationally expensive. Computational cost of SVM only depends on the number of training samples, which is relatively low in practice, while the extraction of high-level linguistic features often involves in searching process, which will probably be the major part of the total computational cost.
2. Avoid from conflict and cross-dependency between features. Sometimes one wants to involve in more features to capture ‘all’ aspects of human evaluation, but we deem that features may conflict with each other and cross-dependent features increase the difficulty in obtaining function which has good capability to generalize. Moreover, in the other hand, Sun et al. (2008) introduced an approach based on SVM regression using only six features, and achieved a comparable result with an approach using 53 features (Albrecht and Hwa 2007).
3. Find key features that can represent more aspects of human evaluation process. We indeed do not have much idea of which factors do human actually emphasize exactly, and features adopted by current approaches are more or less only an approximation to that. However, this is an attracting field which worth deeper investigation.

In section 2.2, we will continue to introduce feature set used in our system.

## 2.2 Feature Set

Our MT evaluation system is designed for participating NIST MetricsMATR2008 contest at sentence level. MT systems to be evaluated are Arabic-English MT systems. The feature set in our system contains altogether 12 features that can fall into three linguistic levels:

**Lexical level:** BLEU, BLEU2, letter-based BLEU, ROUGE-9R, METEOR, translation pre-

recision/recall of content words after morphological reduction

**Phrase level:** Translation precision/recall of noun phrases

**Sentence level:** Entropy of a sentence, Parser score of a sentence, P-norm of Byte-length ratio between a sentence and reference sentence

We can see that features in lexical level constitute the major part of the feature set, indicating that high-level features generated by noisy NLP utilities still need polishing. In the following of this section, we will describe each feature in detail.

#### **Lexical level:**

•BLEU and METEOR (Banerjee and Lavie 2005) are adopted in the original form described in the papers respectively.

•BLEU2 is the individual bigram BLEU score. We found it correlates well with human assessments on develop dataset.

•Letter-based BLEU is calculated as follows:

1. Compute the average word length of reference translation sentences. For example, the average length is  $n$  letters.
2. Split every word in both machine translation and human reference sentences into letters.
3. Calculate cumulative BLEU- $n$  on the split sentences obtained in step 2.

This criterion is novel and has a good correlation with human assessments. It pays more attention to translation's adequacy, since it has a relatively narrow view (one word in average). Besides, it is less sensitive to translation errors of short words, which are mostly function words.

•ROUGE-9R is calculated according to Lin and Och (2004). It is the case-insensitive translation recall of skip-bigrams within a window of which the size is 9 words.

•Translation precision/recall of content words after lemmatization

English words have plenty of morphological changes. So if a machine translation sentence shares with a human reference sentence some cognates, it contains at least some basic information correct. Moreover, if we look at it in another way, words that do not match in the original text maybe match after lemmatization. Thus, differences between poor translations will be revealed. These two criteria are calculated as follows:

1. POS-tag machine translations and reference translations.

2. In all the sentences, find all content words, and convert them to morphological reduction form.

3. Calculate precision and recall on the word list generated in step 2.

#### **Phrase level:**

•Translation precision/recall of noun phrases

We parse every sentence by Collins parser (Collins, 1999) and extract all noun phrases recursively, and then compute the precision and recall on all the phrases obtained.

In practice, we found that if we use case-insensitive matching phrases, we would receive higher performance.

#### **Sentence level:**

•Entropy

We trained a sub language model on all reference translations using CMU SLM toolkit, and obtained entropy of each machine translation sentence based on this sub LM. The entropy reflects the extent to which a sentence could be generated from the language model, and so represents translation quality. Moreover, a sub language model can represent the certain domain of the whole language.

•Parser score

The parser will generate a score for the parse of a given sentence, i.e. the logarithm of the probability of the sentence. This score may be regarded as a syntactic information based language model score as well as an approximate representation of parse structure. Suppose the score of a sentence is  $s$ ,  $s$  will scale from tens to hundreds negative. We rescaled it with  $-100/s$  so that it could approximately range between 0 and 1.

•P-norm of Byte-length ratio

We speculate that translations to the same source sentence have similar byte-length, and we assume this byte-length has a Gaussian distribution. We estimated the mean  $\mu$  and variation  $\sigma$  of the length of all references to the same source sentence. Let the random variable  $X$  represent the byte-length of a sentence translated from the same sentence, then  $(X-\mu)/\sigma$  will have a 0-1 Gaussian distribution. We will finally employ  $2\{1-P-norm[(X-\mu)/\sigma]\}$  as the feature value.  $P-norm(\cdot)$  function is described by Abramowitz and

Source of Data	MT-06	TRANSTAC
Genre	Newswire	training dialogs
Number of documents/scenario	25	1
Total number of segments	249	16
Source Language	Arabic	Iraqi Arabic
Number of system translations	8	5
Number of reference translations	4	4
Human assessment scores	Score 1-7, adequacy	Score 1-7, adequacy

Table 1: Data set summarization

Stegun (1964; p. 932, equation 26.2.17). Assuming  $\delta$  has a 0-1 Gaussian distribution,  $P\text{-norm}(\delta)$  is the approximation of:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{|\delta|} e^{-z^2/2} dz$$

It is obviously that the higher  $2[1-P\text{-norm}(\delta)]$  is, the closer between  $\delta$  and 0. That is to say that the byte-length of a translation sentence is closer to the average byte-length of reference sentences.

### 3 Experiment

We will show in this section some experimental results during development of our system. After introducing the data set and toolkit used, we will first show the performance of individual criteria, and then show the performance of our system.

#### 3.1 Experimental Settings

NIST provides all participants with a development dataset (LDC2008E43 corpus). The data set information is summarized in Table 1.

We employ SVM-Light (Joachims 1999) to train SVM regression model from human assessment data. Since there is only single score given, we take it as the regression target directly.

In order to compare performance between metrics, we made the following settings:

- For metrics using machine learning method, we split the whole dataset which containing  $249*8+16*5=2072$  sentences into 5 sub sets, and five-fold cross-validation was performed.

- For metrics not using machine learning method, performance will be directly computed on 2072 sentences.

To examine metric's performance, we introduce Pearson's  $r$ , Spearman's  $\rho$  (Siegel, 1956) and Kendall's  $\tau$  (Kendall, 1938).

Pearson's  $r$  is calculated as:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}$$

where  $X_i$  and  $Y_i$  are samples from two data sets and  $\bar{X}$ ,  $\bar{Y}$  are means of the data sets.

In principle,  $\rho$  is simply a special case of the Pearson product-moment correlation coefficient in which two sets of data  $X_i$  and  $Y_i$  are converted to rankings  $x_i$  and  $y_i$  before calculating the coefficient. In practice, Spearman's  $\rho$  is calculated as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$d_i = x_i - y_i$  = the difference between the ranks of corresponding values  $X_i$  and  $Y_i$ , and  $n$  = the number of values in each data set (same for both sets).

And if there is too much samples which have the same rank, one should modify the formula to:

$$\rho = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2} \sqrt{n(\sum y_i^2) - (\sum y_i)^2}}$$

Kendall's  $\tau$  is calculated as:

$$\tau = \frac{2P}{\frac{1}{2}n(n-1)} - 1 = \frac{4P}{n(n-1)} - 1$$

where  $n$  is the number of items, and  $P$  is the sum, over all the items, of the number of items ranked after the given item by both rankings

#### 3.2 Individual Performance

We examined each individual criterion's correlation with human assessments. Results are shown in Table 2. Some other criteria that we examined in development but finally not included in the feature set are also shown for comparison.

Translation precision/recall of content words after lemmatization are displayed as LCP, LCR, and Translation precision/recall of noun phrases are displayed as NPP, NPR in Table 2. We also implemented STM with a maximum depth of four (Liu and Gildea 2005). NIST (Doddington, 2002), precision and f-mean of ROUGE-9 (Lin

and Och 2004) are also listed for comparison. More discussions of them will be in section 4.

Criterion	Pearson $r$	Spearman $\rho$	Kendall $\tau$
METEOR	0.705	0.703	0.555
BLEU-L	0.691	0.695	0.553
LCR	0.655	0.652	0.511
ROUGE-9R	0.648	0.679	0.533
BLEU-2	0.625	0.625	0.491
BLEU	0.585	0.589	0.471
LCP	0.562	0.539	0.414
NPR	0.394	0.419	0.321
Entropy	0.389	0.442	0.332
NPP	0.385	0.400	0.307
PSCORE	0.099	0.126	0.093
LOS	-0.261	-0.339	-0.251
NIST	0.673	0.669	0.524
ROUGE-9	0.639	0.660	0.517
ROUGE-9P	0.593	0.613	0.478
STM	0.293	0.334	0.253

Table 2: Correlations with human assessment of individual criterion sorted with Pearson’s  $r$  in descending order. NIST, ROUGE-9, ROUGE-9P and STM are not used in our system; we list them here for comparison.

At the first sight of view, all of the recalls outperform corresponding precisions. This may be caused by the relatively worse performance of NLP toolkits when encountering noisy machine translation outputs. Moreover, METEOR performs the best out of all criteria, and letter-based BLEU (BLEU-L) shows its superiority. As mentioned in section 2.2, letter-based BLEU concentrates on a local area of a sentence (actually one word in average), and tolerates translation errors on short words, much of which are function words that are not so important in adequacy assessments. We also discovered that some poor criterion such as parser score (PSCORE) and byte-length ratio (LOS) have contribution to the final performance, because if we remove them, the final performance would drop.

### 3.3 Criteria Combination

We examined the performance of the SVM regression model by five-cross validation described in section 3.1.

In addition, we noted that the translation precision/recall on some linguistic objects is probably redundant. Although we employed all of them in the system submitted, we are still interested in how the removal of this redundancy would affect system’s performance. Therefore, we will also show the performance of the system

with features ‘NPP’ and ‘LCP’ removed. The results are shown in Table 3.

System	Pearson $r$	Spearman $\rho$	Kendall $\tau$
Submitted	0.783	0.749	0.600
NPP,LCP removed	0.788	0.751	0.602

Table 3: Performance of the system submitted, compared with the system with NPP and LCP removed.

The overall result is relatively better than individual criterion, showing the combination power of SVM regression. Besides, we can see that after eliminating the redundancy of precision/recall by removing features ‘NPP’ and ‘LCP’, the performance increased slightly, confirming the suspicion that the redundancy would harm the final performance.

## 4 Discussion and future work

The feature set we used was generated in two phases. We first empirically employed some basic linguistic features, and then added other features such as METEOR, ROUGE, BLEU, etc. by examining their correlation with human assessments and the resulting final performance by adding them to the feature set. Therefore, because ROUGE-9R outperforms ROUGE-9, ROUGE-4, and ROUGE-\* (Lin and Och 2004), we adopted ROUGE-9R finally. Likewise, METEOR, BLEU, BLEU-2, BLEU-L were add into the feature set. On the other hand, though NIST correlates well with human assessments, the final performance is not benefitted from adding NIST to the feature set. We also intended to employ more features at phrase level or syntax level, but after we tried several features including STM, we found all of them would make the final performance drop.

Feature selection is not completely successful in this system. The empirically employed features were not examined carefully, resulting redundancy in the feature set. In future study, we will look for a uniform and efficient method of selecting and mining features.

The correlations between many criteria (such as BLEU, NIST and METEOR) and human assessments are much higher than previous studies. It may be caused by the relatively good quality of machine translations. That means n-gram matching approaches will not suffer from data sparseness so much.

The human assessments aim at adequacy only, this may be essential to explain some criteria’s

high performance such as letter-based BLEU. When we involve in fluency assessments, would these criteria work well? We will investigate it in future studies.

We also speculate that one single feature set could not represent the quality aspects of both good translations and bad translations. Is it more appropriate creating different feature sets for translations of different qualities? We hope to address this issue in our future work as well.

## Acknowledgments

This research is supported by Natural Science Foundation of China (Grant No. 60773066) and National 863 Project (Grant No. 2006AA01Z150)

## Reference

- Joshua Albrecht and Rebecca Hwa. 2007. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. In Proceedings of ACL.
- Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. 2006. MT Evaluation: Human-Like vs. Human Acceptable. In Proceedings of COLING-ACL06.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.
- K. P. Bennett and O. L. Mangasarian, 1992. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34.
- George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In Proceedings of the 2nd IHLT.
- Kevin Duh, Ranking vs. Regression in Machine Translation Evaluation. In Proceedings of the Third Workshop on Statistical Machine Translation, June, Columbus Ohio, pages 191-194
- Jesús Giménez and Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In Proceedings of the ACL Workshop on Statistical Machine Translation.
- Jesús Giménez and Lluís Màrquez. 2008. Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations. In Proceedings of IJCNLP, pages 319–326.
- David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In Proceedings of NLHNAACL.
- Maurice Kendall, 1938. A New Measure of Rank Correlation, *Biometrika*, 30, 81-89.
- Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation.
- Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In Proceedings of ACL.
- Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.
- Dennis Mehay and Chris Brew. 2007. BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation. In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI).
- Karolina Owczarzak, Declan Groves, Josef Van Genabith, and Andy Way. 2006. Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, RC22176, IBM. Technical report, IBM T.J. Watson Research Center.
- S. Siegel and N.J. Catellan, 1988. *Non-parametric Statistics for the Behavioral Sciences*, McGraw-Hill, 2nd edition.
- Alex J. Smola and Bernhard Schölkopf, 2001. A tutorial on support vector regression. *Statistics and Computing*, Forthcoming.
- Shuqi Sun, Yin Chen and Jufeng Li, 2008. A Re-examination on Features in Regression Based Approach to Automatic MT Evaluation. In Proceedings of the ACL-08: HLT Student Research Workshop, June, Columbus Ohio, pages 25-30.
- V. Vapnik, 1995. *The Nature of Statistical Learning Theory*. Springer, New York,
- Yang Ye, Ming Zhou and Chin-Yew Lin. 2007. Sentence Level Machine Translation Evaluation as a Ranking. In Proceedings of the ACL Second Workshop on Statistical Machine Translation, Prague, Czech Republic, June.