

A SVM Regression Based Skip-Ngram Approach to MT Evaluation

Bo Wang, Tiejun Zhao, Muyun Yang, Sheng Li
School of Computer Science and Technology, Harbin Institute of Technology
Harbin, China 150001
{bowang, tjzhao, ymy}@mtlab.hit.edu.cn
lisheng@hit.edu.cn

Abstract

This paper describes an automatic MT evaluation metric named SNR by Machine Intelligence and Translation Lab. of Harbin Institute of Technology, for NIST Metrics-MATR 2008 evaluation. The metric extend the idea of skip-bigram with larger span and multiple statistics. SVM regression method is adopted to tune the weights of statistics in the metric. The experimental results show that SNR correlate with human assessments better than mainstream metrics on sentence level.

1 Motivation

Automatic evaluation of machine translation (MT) systems is an important problem for the development of MT technology. Automatic evaluation metrics make it possible to quickly determine the quality of MT system outputs. This is useful for the comparison of various MT systems or the incremental development of single MT system when human judgments are expensive to obtain. Furthermore, automatic MT evaluation metrics are also used for tuning system parameters (Och 2003).

Many automatic evaluation metrics have been proposed to date. Metrics based on various measures of the similarity of strings have been successful and widely accepted such as BLEU. (Papineni et al., 2002) work by comparing MT output with one or more human reference translations and generating a similarity score. TER (Snover et al., 2006) uses edit-distances. GTM (Melamed 2003) calculate the precision and recall of machine translation using longest common subsequence. ROUGE (Lin and Och, 2004) is based on LCS and skip-bigram. METEOR

(Banerjee and Lavie, 2005) uses aligned 1-grams and their linguistic equivalents.

Besides the metrics based on the similarity of strings, richer linguistic features are believed to be more powerful to present the quality of the MT outputs. 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).

Currently, though linguistic features has been proved to be effective to improve the performance of the MT evaluation the metrics based on the string similarity is still the researchers' first choice for system comparison and parameters tuning in most cases. This happens because there are several advantages of string similarity based metrics:

- They are easier to be understood and implemented.
- They are independent with languages.
- They are available when the linguistic parsers are absent.
- They often have less time consumption.

The first three characters lead to better adaptability and the last one is very important for the tuning of parameters.

Our approach is a kind of string similarity based metrics, inspired by the idea of skip-bigram introduced by Rouge. We extend the original method with larger span and multiple statistics. We consider three different co-occurrence measures for N-gram pairs separately including the partial matching, full matching and ordered matching. Two novel metrics are constructed based on the three matching scores. In the first metric, three scores are equally normalized into a general score. In the second metric, SVM regression method is applied to weight the

scores. The experimental results show that both metrics greatly improved the original method and the second metric correlate with human assessments better than all mainstream metrics on sentence level.

2 Methodology

In this section, we will first describe how to calculate three skip-Ngram matching scores, and then introduce the SVM regression based automatic MT evaluation and its implementation in our approach.

2.1 Skip-Ngram Co-Occurrence Statistics

Skip-Ngram is any pair of N-gram in their sentence order, allowing for arbitrary gaps. Skip-Ngram co-occurrence statistics measure the overlap of skip-Ngram between a candidate translation and a set of reference translations. Different with the well-known skip-bigram (Lin, 2004), each N-gram of a skip-Ngram pair is allowed to contain more than one word, but the number of the words in each N-gram are required to be equal. For example:

Reference: Australia reopens embassy in Manila

Candidate: Australia reopens in the Manila embassy

If N=1, reference has following skip-1grams:
 {"Australia, reopens", "Australia, embassy", "Australia, in", "Australia, manila", "reopens, embassy", "reopens, in", "reopens, manila", "embassy, in", "embassy, manila", "in, manila"}

If N=2, reference has following skip-2grams:
 {"Australia reopens, embassy in", "Australia reopens, in manila", "reopens embassy, in manila"}

In our approach we calculate the recall of each reference with the skip-Ngram pairs in it. Each skip-Ngram pair is matched to the candidate in three different ways: it is partially matched if only one of the two members is matched, it is fully matched if both members are matched and it is ordered matched if both mem-

3 Experiment

We will show in this section some experimental results during development of our system. After introducing the data sets and the settings, we will show the performance of our system together with several popular mainstream metrics.

bers are matched in their sentence order. In the above example:

We calculate the score for each matching degree with following formulas:

$$Score_P(c, R) = \text{Max}_{r \in R} \frac{\sum_n \sum_{x, y \in n\text{-grams in } r} (M_P([x, y], c))}{\sum_n \#[x, y]_{x, y \in n\text{-grams in } r}} \quad (1)$$

$$Score_F(c, R) = \text{Max}_{r \in R} \frac{\sum_n \sum_{x, y \in n\text{-grams in } r} (M_F([x, y], c))}{\sum_n \#[x, y]_{x, y \in n\text{-grams in } r}} \quad (2)$$

$$Score_O(c, R) = \text{Max}_{r \in R} \frac{\sum_n \sum_{x, y \in n\text{-grams in } r} (M_O([x, y], c))}{\sum_n \#[x, y]_{x, y \in n\text{-grams in } r}} \quad (3)$$

$$M_P([x, y], c) = \begin{cases} 1 & \text{if } c \text{ contains } x \text{ or } y \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$M_F([x, y], c) = \begin{cases} 1 & \text{if } c \text{ contains } x \text{ and } y \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$M_O([x, y], c) = \begin{cases} 1 & \text{if } c \text{ contains } x \text{ and } y \text{ in order} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where c is a candidate translation and R is a set of corresponding reference translations. $Score_P$, $Score_F$ and $Score_O$ present the matching scores of partial matching, full matching and ordered matching respectively.

We calculate the general score " $Score(c, R)$ " for c by equally normalizing the three matching scores:

$$Score(c, R) = \text{Max}_{r \in R} \frac{\sum_n \sum_{x, y \in n\text{-grams in } r} M([x, y], c)}{3 * \sum_n \#[x, y]_{x, y \in n\text{-grams in } r}} \quad (7)$$

$$M([x, y], c) = M_P([x, y], c) + M_F([x, y], c) + M_O([x, y], c) \quad (8)$$

3.1 Experimental Settings

NIST provides all participants with a development dataset (LDC2008E43 corpus) which is described in Table 1.

Source of Data	LDC2008E43
Genre	Newswire
Number of documents	25
Total number of segments	249
Source Language	Arabic
Number of system translations	8
Number of reference translations	4
Human assessment scores	Score 1-7

Table 1: Data set description

In this work, we also use the three scores of different matching degrees as the features of the MT evaluation and model the metric using SVM-regression. We employ SVM-Light (Joachims 1999) to train SVM regression model from human assessments data. Since there is only single score given by test data set, we take it as the regression target directly. For in-year tests, we do 5-fold validation on single data set.

As a further switch, we also use the Porter stemmer to get the stem of the words in candidate translation and references.

We compare the performance between metrics using Pearson’s r which is calculated as:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}} \quad (9)$$

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

Furthermore, all sentences are segmented beforehand and the matching between the words is case-insensitive.

3.2 Skip-Ngram Performance

We examine the performance of the skip-Ngram metric with and without the SVM regression methods, which are named as SNR and SN respectively. Both metrics are switched with four parameters settings and tested. Results are shown in Table 2 together with scores of six mainstream metrics. In the table, Rouge-4, Rouge-9 and Rouge* present the Rouge using skip-bigram with a window of 4, 9 and free number of words separately, In row 7-14, the names of the novel metrics are formatted as “SN(R)-N-[Stem]” where “N” indicate the maximum length N of the skip-Ngram and “Stem” indicate whether the words are stemmed or not.

As shown in the results, the stemmed skip-Ngram-regression using 4-gram (SNR-4-Stem) performs the best out of all metrics. When we

focus on the metrics without regression method, all SN series metrics outperform all mainstream methods but METEOR. While SNR series with regression method achieve great improvement and exceed the METEOR. Comparing with the Rouge, the basic metric of this work, all novel metrics performs better.

Criterion	Correlation
METEOR	0.705
ROUGE-4	0.654
ROUGE-9	0.663
ROUGE*	0.655
BLEU	0.609
GTM	0.543
SN-1	0.686
SN-4	0.665
SN-1-Stem	0.689
SN-4- Stem	0.670
SNR-1	0.716
SNR-4	0.741
SNR-1- Stem	0.720
SNR-4- Stem	0.745

Table 2: Pearson’s correlations with human assessment of skip-Ngram and skip-Ngram-regression with four switches. Six mainstream metrics are listed for comparison.

4 Discussion and future work

Though the SN and SNR series shows impressive capability from the overview the best parameters settings are still uncertain. We take two groups of experiments with SN and SNR. For each group of experiments we use four different parameters settings. It is clear that stemming is always helpful in all experiments. In the experiments with SNR longer gram improve the performance while it’s contrary in the experiments with SN.

Nowadays, the most popular solution of the evaluation of machine translation is the introduction of richer linguistic features, which are automatically weighted by the machine learning methods. It’s undoubted that linguistic features is the most credible way to catch the inherent characters of languages but this idea highly rely on the reliability of the linguistic parsers and will be helpless when the parsers are absent. The question is: how much can we learn from the plain words of the languages? This approach proves that there is still something we can find. In future, we will try to find the key features cause the different performance of evaluation metrics on various languages and try to catch these features through the plain words.

Acknowledgments

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

Reference

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.

I. Dan Melamed, Ryan Green, Joseph P. Turian, 2003, Precision and recall of machine translation, In Proceedings of HLT/NAACL 2003.

Franz Josef Och. Minimum Error Rate Training for Statistical Machine Translation. In "ACL 2003: Proc.of the 41st Annual Meeting of the Association for Computational Linguistics", Japan, Sapporo, July 2003.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In Proceedings of NLHNAACL.

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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas, 2006.