Regression and Ranking based Optimisation for Sentence Level Machine Translation Evaluation

Xingyi Song and **Trevor Cohn**
The Department of Computer Science
University of Sheffield
Sheffield, S1 4DP. UK
{xsong2,t.cohn}@shef.ac.uk

Abstract

Automatic evaluation metrics are fundamentally important for Machine Translation, allowing comparison of systems performance and efficient training. Current evaluation metrics fall into two classes: heuristic approaches, like BLEU, and those using supervised learning trained on human judgement data. While many trained metrics provide a better match against human judgements, this comes at the cost of including lots of features, leading to unwieldy, non-portable and slow metrics. In this paper, we introduce a new trained metric, ROSE, which only uses simple features that are easy portable and quick to compute. In addition, ROSE is sentence-based, as opposed to document-based, allowing it to be used in a wider range of settings. Results show that ROSE performs well on many tasks, such as ranking system and syntactic constituents, with results competitive to BLEU. Moreover, this still holds when ROSE is trained on human judgements of translations into a different language compared with that use in testing.

1 Introduction

Human judgements of translation quality are very expensive. For this reason automatic MT evaluation metrics are used to as an approximation by comparing predicted translations to human authored references. An early MT evaluation metric, BLEU (Papineni et al., 2002), is still the most commonly used metric in automatic machine translation evaluation. However, several drawbacks have been stated by many researchers (Chiang et al., 2008a; Callison-Burch et al., 2006; Banerjee and Lavie, 2005), most notably that it omits recall (substituting this with a penalty for overly short output) and not being easily applied at the sentence level. Later heuristic metrics such as METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006) account for both precision and recall, but their relative weights are difficult to determine manually.

In contrast to heuristic metrics, trained metrics use supervised learning to model directly human judgements. This allows the combination of different features and can better fit specific tasks, such as evaluation focusing more on fluency/adequacy/relative ranks or post editing effort. Previous work includes approaches using classification (Corston-Oliver et al., 2001), regression (Albercht and Hwa, 2008; Specia and Gimenez, 2010), and ranking (Duh, 2008). Most of which achieved good results and better correlations with human judgments than heuristic baseline methods.

Overall automatic metrics must find a balance between several key issues: a) applicability to different sized texts (documents vs sentences), b) easy of portability to different languages, c) runtime requirements and d) correlation with human judgement data. Previous work has typically ignored at least one of these issues, e.g., BLEU which applies only to documents (A), trained metrics (Albercht and Hwa, 2008; Specia and Gimenez, 2010) which tend to ignore B and C.

This paper presents ROSE, a trained metric which is loosely based on BLEU, but seeks to further simplify its components such that it can be used for sentence level evaluation. This contrasts with BLEU which is defined over large documents, and must

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 123–129,
Edinburgh, Scotland, UK, July 30–31, 2011.

be coarsely approximated to allow sentence level application. The increased flexibility of ROSE allows the metric to be used in a wider range of situations, including during decoding. ROSE is a linear model with a small number of simple features, and is trained using regression or ranking against human judgement data. A benefit of using only simple features is that ROSE can be trivially ported between target languages, and that it can be run very quickly. Features include precision and recall over different sized n-grams, and the difference in word counts between the candidate and the reference sentences, which is further divided into content word, function word and punctuation. An extended versions also includes features over Part of Speech (POS) sequences.

The paper is structured as follows: Related work on metrics for statistical machine translation is described in Section 2. Four variations of ROSE and their features will be introduced in Section 3. In section 4 we presents the result, showing how ROSE correlates well with human judgments on both system and sentence levels. Conclusions are given at the end of the paper.

2 Related Work

The defacto standard metric in machine translation is BLEU (Papineni et al., 2002). This measures n-gram precision (n normally equal to 1,2,3,4) between a document of candidate sentences and a set of human authored reference documents. The idea is that high quality translations share many n-grams with the references. In order to reduce repeatedly generating the same word, BLEU clips the counts of each candidate N-gram to the maximum counts of that n-gram that in references, and with a brevity penalty to down-scale the score for output shorter than the reference. In BLEU, each n-gram precision is given equal weight in geometric mean, while NIST (Doddington and George, 2002) extended BLEU by assigning more informative n-grams higher weight.

However, BLEU and NIST have several drawbacks, the first being that BLEU uses a geometric mean over all n-grams which makes BLEU almost unusable for sentence level evaluations [1]. Secondly, BLEU and NIST both use the brevity penalty to replace recall, but Banerjee and Lavie (2005) in experiments show that the brevity penalty is a poor substitute for recall.

Banerjee and Lavie (2005) proposed a METEOR metric, which that uses recall instead of the BP. Callison-Burch et al. (2007; Callison-Burch et al. (2008) show that METEOR does not perform well in out of English task. This may because that Stemmer or WordNet may not available in some languages, which unable to model synonyms in these cases. In addition, the performance also varies when adjusting weights in precision and recall.

Supervised learning approaches have been proposed by many researchers (Corston-Oliver et al., 2001; Duh, 2008; Albercht and Hwa, 2008; Specia and Gimenez, 2010). Corston-Oliver et al. (2001) use a classification method to measure machine translation system quality at the sentence level as being human-like translation (good) or machine translated (bad). Features extracted from references and machine translation include heavy linguistic features (requires parser).

Quirk (2004) proposed a linear regression model which is trained to match translation quality. Albercht and Hwa (2008) introduced pseudo-references when data driven regression does not have enough training data. Most recently, Specia and Gimenez (2010) combined confidence estimation (without reference, just using the source) and reference-based metrics together in a regression framework to measure sentence-level machine translation quality.

Duh (2008) compared the ranking with the regression, with the results that with same feature set, ranking and regression have similar performance, while ranking can tolerate more training data noise.

3 Model

ROSE is a trained automatic MT evaluation metric that works on sentence level. It is defined as a linear model, and its weights will be trained by Support Vector Machine. It is formulated as

$$S = \overrightarrow{w} \cdot f(\overrightarrow{c}, \overrightarrow{r}) \quad (1)$$

where $\overrightarrow{w}$ is the feature weights vector, $f(\overrightarrow{c}, \overrightarrow{r})$ is the feature function which takes candidate transla-

[1] Note that various approximations exits (Lin and Och, 2004; Chiang et al., 2008b)

tion ($\overrightarrow{c}$) and reference ($\overrightarrow{c}$), and returns the feature vector. S is the response variable, measuring the 'goodness' of the candidate translation. A higher score means a better translation, although the magnitude is not always meaningful.

We present two different method for training: a linear regression approach ROSE-reg, trained to match human evaluation score, and a ranking approach ROSE-rank to match the relative ordering of pairs of translations assigned by human judge. Unlike ROSE-reg, ROSE-rank only gives relative score between sentences, such as A is better than B. The features that used in ROSE will be listed in section 3.1, and the regression and ranking training are described in section 3.2

3.1 ROSE Features

Features used in ROSE listed in Table 1 include string n-gram matching, Word count and Part of Speech (POS). String N-gram matching features, are used for measure how closely of the candidate sentence resembles the reference. Both precision and recall are considered. Word count features measure length differences between the candidate and reference, which is further divided into function words, punctuation and content words. POS features are defined over POS n-gram matches between the candidate and reference.

3.1.1 String Matching Features

The string matching features include n-gram precision, n-gram recall and F1-measure. N-gram precision measures matches between sequence of words in the candidate sentence compared to the references,

$$P_n = \frac{\sum_{\text{n-gram} \in \overrightarrow{c}} Count(\text{n-gram}) [\![\text{n-gram} \in \overrightarrow{r}]\!]}{\sum_{\text{n-gram} \in \overrightarrow{c}} Count(\text{ngram})} \quad (2)$$

where $Count$ are the occurrence counts of n-grams in the candidate sentence, the numerator measures the number of predicted n-grams that also occur in the reference.

Recall is also used in ROSE, so clipping was deemed unnecessary in precision calculation, where the repeating words will increasing precision but at the expense of recall. F-measure is also included, which is the harmonic mean of precision and recall.

ID	Description
1-4	n-gram precision, n=1...4
5-8	n-gram recall, n=1...4
9-12	n-gram f-measure, n=1...4
13	Average n-gram precision
14	Words count
15	Function words count
16	Punctuation count
17	Content words count
18-21	n-gram POS precision, n=1...4
22-25	n-gram POS recall, n=1...4
26-29	n-gram POS f-measure, n=1...4
30-33	n-gram POS string mixed precision, n=1...4

Table 1: ROSE Features. The first column is the feature number. The dashed line separates the core features from the POS extended features.

With there are multiple references, the n-gram precision error uses the same strategy as BLEU: n-grams in candidate can match any of the references. For recall, ROSE will match the n-grams in each reference separately, and then choose the recall for the reference with minimum error.

3.1.2 Word Count Features

The word count features measure the length difference between a candidate sentence and reference sentence. In a sentence, content words are more informative than function words (grammatical words) and punctuation. Therefore, the number of content word candidate is a important indicator in evaluation. In this case, besides measuring the length at whole sentences, we also measure difference in the number of *function words*, *punctuation* and *content words*. We normalise by the length of the reference which allows comparability between short versus long sentences. In multiple reference cases we choose the ratio that is closest to 1.

3.1.3 Part of Speech Features

The string matching features and word count features only measure similarities on the lexical level, but not over sentence structure or synonyms. To add this capability we also include Part of Speech (POS) features which work similar to the String Matching features, but using POS instead of words. The fea-

tures measure precision, recall and F-measure over POS n-grams (n=1...4). In addition, we also include features that mixed string and POS.

The string/POS mixed feature is used for handling synonyms. One problem in string n-gram matching is not being able to deal with the synonyms between the candidate translation and the reference. One approach for doing so is to use an external resource such as WordNet (Banerjee and Lavie, 2005), however this would limit the portability of the metric. Instead we use POS as a proxy. In most of the cases, synonyms share the same POS, so this can be rewarded by forming n-grams over a mixture of tokens and POS. During the matching process, both words and its POS shall be considered, if either matches between reference and candidate, the n-gram matches will be counted.

Considering the example in table 2, candidate 1 has better translation than candidate 2 and 3. If only the string N-gram matching is used, that will give the same score to candidate 1, 2 and 3. The n-gram precision scores obtained by all candidate sentences in this example are: 2-gram = 1, 3-gram = 0. However, we can at least distinguish candidate 1 is better than candidate 3 if string POS mixed precision is used , n-gram precision for candidate 1 will be: 2-gram = 2, 3-gram = 1, which ranks candidate 1 better than candidate 3.

Example
reference: A/DT red/ADJ vehicle/NN
candidate 1: A/DT red/ADJ car/NN
candidate 2: A/DT red/ADJ rose/NN
candidate 3: A/DT red/ADJ red/ADJ

Table 2: Evaluation Example

3.2 Training

The model was trained on human evaluation data in two different ways, regression and ranking.These both used SVM-light (Joachims, 1999). In the ranking model, the training data are candidate translation and their relative rankings were ranked by human judge for a given input sentence. The SVM finds the minimum magnitude weights that are able to correctly rank training data which is framed as a series of constraints reflecting all pairwise comparisons. A soft-margin formulation is used to allow training errors with a penalty (Joachims, 2002). For regression, the training data is human annotation of post-edit effort (this will be further described in section 4.1). The Support vector Regression learns weights with minimum magnitude that limit prediction error to within an accepted range, again with a soft-margin formulation (Smola and Schlkopf, 2004).

A linear kernel function will be used, because non-linear kernels are much slower to use and are not decomposable. Our experiments showed that the linear kernel performed at similar accuracy to other kernel functions (see section 4.2).

4 Experimental Setup

Our experiments test ROSE performance on document level with three different Kernel functions: linear, polynomial and radial basis function. Then we compare four variants of ROSE with BLEU on both sentence and system (document) level.

The BLEU version we used here is NIST Open MT Evaluation tool mteval version 13a, smoothing was disabled and except for the sentence level evaluation experiment. The system level evaluation procedure follows WMT08 (Callison-Burch et al., 2008), which ranked each system submitted on WMT08 in three types of tasks:

- **Rank:** Human judges candidate sentence rank in order of quality. On the document level, documents are ranked according to the proportion of candidate sentences in a document that are better than all of the candidates.
- **Constituent:** The constituent task is the same as for ranking but operates over chosen syntactic constituents.
- **Yes/No:** WMT08 Yes/No task is to let human judge decide whether the particular part of a sentence is acceptable or not. Document level Yes/No ranks a document according to their number of YES sentences

Spearman's rho correlation was used to measure the quality of the metrics on system level. Four target languages (English, German, French and Spanish) were used in system level experiments. ROSE-

reg and ROSE-rank were tested in all target language sets, but ROSE-regpos was only tested in the into-English set as it requires a POS tagger. On the sentence level, we compare sentences ranking that ranked by metrics against human ranking. The evaluation quality was examined by Kendall's tau correlation, and tied results from human judges were excluded.

Rank	**es-en**	**fr-en**	**de-en**	**avg**
Linear	0.57	0.97	0.69	0.74
Polynomial	**0.62**	0.97	**0.71**	**0.76**
RBF	0.60	**0.98**	0.62	0.73
Constituent				
Linear	0.79	0.90	0.39	0.69
Polynomial	0.80	0.89	**0.41**	**0.70**
RBF	**0.83**	**0.93**	0.34	**0.70**
Yes/No				
Linear	**0.92**	**0.93**	**0.67**	**0.84**
Polynomial	0.86	0.90	0.66	0.81
RBF	0.87	**0.93**	0.65	0.82

Table 3: ROSE-reg in with SVM kernel functions

Metric	**Kendall's tau**
BLEU-smoothed	**0.219**
ROSE-reg	0.120
ROSE-regpos	0.164
ROSE-rank	0.206
ROSE-rankpos	0.172

Table 4: Sentence Level Evaluation

4.1 Data

Training data used for ROSE is from WMT10 (Callison-Burch et al., 2010) human judged sentences. A regression model was trained by sentences with human annotation for post editing effort. The three levels used in WMT10 are 'OK', 'EDIT' and 'BAD', which we treat as response values of 3, 2 and 1. In total 2885 sentences were used in the regression training. The ranking model was trained by sentences with human annotating sentence ranking, and tied results are allowed in training. In this experiment, 1675 groups of sentences were used for training, and each group contains five sentences, which are manually ranked from 5 (best) to 1 (worst). In order to test the ROSE's ability to adapt the language without training data, ROSE was only trained with English data.

The testing data on sentence level used in this paper is human ranked sentences from WMT09 (Callison-Burch et al., 2009). Tied rankings were removed, leaving 1702 pairs. We only consider translations into English sentences. On system level, the testing data are the submissions for 'test2008' test set in WMT08 (Callison-Burch et al., 2008). ROSE, and BLEU were compared with human ranked submitted system in 'RANK', 'CONSTITUENT' and 'YES/NO' tasks.

English punctuation and 100 common function words list of four languages in this experiment were generated. English POS was tagged by NLTK (Bird and Loper, 2004).

4.2 Results and Discussion

Table 3 shows the results of ROSE-reg with three different SVM Kernel functions. Performance are similar among three different Kernel functions. However, the linear kernel is the fastest and simplest and there is no overall winner. Therefore, linear Kernel function was used in ROSE.

The results of Kendall's tau on sentence level evaluation are shown in Table 4. According to Table 4 ROSE-rank has the highest score in all versions of ROSE. The score is close to the smoothed version of BLEU. Results also showed adding POS feature helped in improving accuracy in the regression model, but not in ranking, The reason for this is not clear, but it may be due to over fitting.

Table 5 and Table 6 are the Spearman's rho in system ranking. Table 5 is the task evaluation for translation into English. ROSE-rank performed the best in the system ranking task. Also, ROSE-regpos is the best in the syntactic constituents task. This may because of ROSE-rank is a ranking based metric and ROSE-regpos incorporates POS that contains more linguistic information. Table 6 shows the results of evaluating translations from English. According to the table, ROSE performs less accurately than for the into-English tasks, but overall the ROSE scores are similar to those of BLEU.

Rank	es-en	fr-en	de-en	avg
BLEU	0.66	0.97	0.69	0.77
ROSE-reg	0.57	0.97	0.69	0.74
ROSE-rank	**0.85**	0.96	**0.76**	**0.86**
ROSE-regpos	0.59	**0.98**	0.71	0.76
ROSE-rankpos	0.83	0.96	0.69	0.82
Constituent				
BLEU	0.78	0.92	0.30	0.67
ROSE-reg	**0.79**	0.90	0.39	0.69
ROSE-rank	0.66	0.92	0.33	0.64
ROSE-regpos	**0.79**	0.90	**0.41**	**0.70**
ROSE-rankpos	0.64	**0.93**	0.31	0.63
Yes/No				
BLEU	**0.99**	**0.96**	0.66	**0.87**
ROSE-reg	0.92	0.93	**0.67**	0.84
ROSE-rank	0.78	**0.96**	0.61	0.78
ROSE-regpos	0.97	0.93	0.66	0.85
ROSE-rankpos	0.81	**0.96**	0.57	0.78

Table 5: System Level evaluation that translation into English

5 Conclusion

We presented the ROSE metric to make up for several drawbacks of BLEU and other trained metrics. Features including string matching, words ratio and POS were combined by the supervised learning approach. ROSE's overall performance was close to BLEU on system level and sentence level. However, it is better on tasks ROSE was specifically trained, such as ROSE-rank in the system level ranking task and ROSE-regpos in the syntactic constituents task. Results also showed that when training data is not available in the right language ROSE produces reasonable results.

Smoothed BLEU slightly outperformed ROSE in sentence evaluation. This might be due to the training data not being expert judgments, and consequently very noisy. In further work, we shall modify the training method to better tolerate noise. In addition, we will modify ROSE by substitute less informative features with more informative features in order to improve its performance and reduce over fitting.

Rank	es-en	fr-en	de-en	avg
BLEU	**0.85**	**0.98**	0.88	**0.90**
ROSE-reg	0.75	**0.98**	0.93	0.89
ROSE-rank	0.69	0.93	**0.94**	0.85
Constituent				
BLEU	**0.83**	**0.87**	0.35	**0.68**
ROSE-reg	0.73	**0.87**	**0.36**	0.65
ROSE-rank	0.72	0.78	0.32	0.61
Yes/No				
BLEU	0.75	**0.97**	0.89	0.87
ROSE-reg	0.72	**0.97**	**0.93**	0.87
ROSE-rank	**0.82**	0.96	0.87	**0.88**

Table 6: System Level evaluation that translation from English

References

Josha S. Albercht and Rebecca Hwa. 2008. Regression for machine translation evaluation at the sentence level. *Machine Translation*, 22:1–27.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the ACL-05 Workshop*.

Steven Bird and Edward Loper. 2004. Nltk: The natural language toolkit. In *Proceedings of the ACL demonstration session*, pages 214–217, Barcelona, July.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *In EACL*, pages 249–256.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan.

2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics. Revised August 2010.

David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. 2008a. Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 610–619, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Chiang, Yuval Marton, and Philip Resnik. 2008b. Online large-margin training of syntactic and structural translation features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 224–233, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. In *proceedings of the Association for Computational Linguistics*.

Doddington and George. 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 '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194, Columbus,Ohio,, June.

T. Joachims. 1999. Making large-scale svm learning practical. *Advances in Kernel Methods - Support Vector Learning,*.

T. Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

C Quirk. 2004. Training a sentence-level machine translation confidence measure. In *In: Proceedings of the international conference on language resources and evaluation*, pages 825–828, Lisbon, Portugal.

Alex J. Smola and Bernhard Schlkopf. 2004. A tutorial on support vector regression. *STATISTICS AND COMPUTING*, 14:199–222.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation.

L. Specia and J. Gimenez. 2010. Combining confidence estimation and reference-based metrics for segment-level mt evaluation. In *The Ninth Conference of the Association for Machine Translation in the Americas*, Denver,Colorado.