Multi-level Evaluation for Machine Translation

Boxing Chen, Hongyu Guo and Roland Kuhn
National Research Council Canada
first.last@nrc-cnrc.gc.ca

Abstract

Translations generated by current statistical systems often have a large variance, in terms of their quality against human references. To cope with such variation, we propose to evaluate translations using a multi-level framework. The method varies the evaluation criteria based on the clusters to which a translation belongs. Our experiments on the WMT metric task data show that the multi-level framework consistently improves the performance of two benchmarking metrics, resulting in better correlation with human judgment.

1 Introduction

The aims of automatic Machine Translation (MT) evaluation metrics, which measure the quality of translations against human references, are twofold. Firstly, they enable rapid comparisons between different statistical machine translation (SMT) systems. Secondly, they are necessary to the tuning of parameter values during system trainings.

To attain these goals, many machine translation metrics have been introduced in recent years. For example, metrics such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and TER (Snover et al., 2006) rely on word n-gram surface matching. Also, metrics that make use of linguistic resources such as synonym dictionaries, part-of-speech tagging, or paraphrasing tables, have been proposed, including Meteor (Banerjee and Lavie, 2005) and its extensions, TER-Plus (Snover et al., 2009), and TESLA (Liu et al., 2011). In addition, attempts to deploy syntactic features or semantic information for evaluation have also been made, giving rise to the STM and DSTM (Liu and Gildea, 2005), DEPREF (Wu et al., 2013) and MEANT family (Lo and Wu, 2011) metrics.

All these evaluation metrics deploy a single evaluation criterion or use the same source of information to evaluate translations. Nevertheless, translations generated by current statistical systems often have widely varying scores, in terms of their quality against human references. As a result, current metrics often perform better for a portion of translations but worse against the others. Consider, for example, two widely used metrics, namely the sentence-level Meteor and BLUE. Figure 1 depicts the distributions of the two metrics' evaluation scores, computed on system outputs for two WMT test sets, i.e., the *newstest2013.fr-en* and *newstest2012.en-cs*. As shown in Figures 1, the variances of the created evaluation scores are large across evaluation metrics as well as test sets.

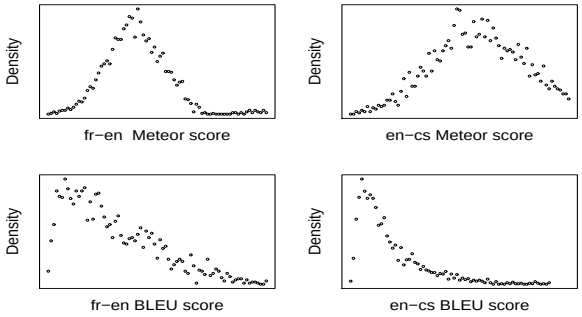

Figure 1: Distributions of translation quality. X-axis is in the range of [0,1].

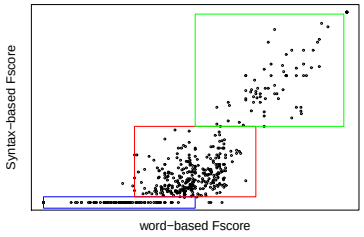

Figure 2: Clusters of translations based on quality. Both X-axis and Y-axis are in the range of [0,1].

Such widely varying evaluation quality, however, may be clustered into multiple sub-regions, as illustrated in Figure 2. Here, we sample 300 sentences from the system output of the *newstest2013.fr-en* test set; we depict the F-measure based on dependency triplet (dependency type, governor word, and dependent word) on the Y-axis against the word-based F-measure on the X-axis. We observe a straight line at the bottom left corner (blue box) of the graph represent-

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 361–365,
Lisboa, Portugal, 17-18 September 2015.

ing sentences which all have dependency triplet F-score of zero; if we want to distinguish between them in terms of their quality score, we must rely on word matching rather than on syntax. The situation in the upper right corner (green box) of the graph is quite different. Here, the word-based F-measure and dependency-based F-measure have a roughly linear correlation, suggesting that a combination of word-based and syntactic information might be a better measure of quality than either alone. These observations imply that a metric may benefit from applying different sources of information at different quality levels.

In this paper, we propose a multi-level automatic evaluation framework for MT. Our strategy first roughly classifies the translations into different quality levels. Next, it rates the translations by exploiting several different information sources, with the weight on each source depending on its quality level. We apply our method to two metrics: the Meteor and a new metric, DREEM, which is based on distributed representations. Our experiments on the WMT metric task data show that the multi-level framework consistently improves the performance of these two metrics.

2 Multi-level Evaluation

The multi-level evaluation framework works on the sentence level. Specifically, we first assign each test sentence to one of the three categories: low-, medium-, or high-quality translations. Next, we evaluate the translations within each category with a tailored set of weights of the metric on the information sources.

To this end, we deploy a simple strategy for the category clustering. Note that more sophisticate strategies could be deployed; we leave this to our future work. Here, we first use a scoring function to compute a score between the translation and its reference. Next, the category assignment of the translation is then determined by a pre-defined score threshold.

In detail, suppose we have a translation (t) and its reference (r). The multi-level metric scores the translation pair as follows.

$$\text{Score(t,r)} = \begin{cases} M(t,r,w_l) & \text{if } (F(t,r) \leq \theta_1) \\ M(t,r,w_m) & \text{if } (\theta_1 < F(t,r) \leq \theta_2) \\ M(t,r,w_h) & \text{otherwise} \end{cases}$$

where $M(t,r,w)$ is a metric, w is the weight, $F(t,r)$ is the simple classification scoring function. Also, θ is a threshold, and its value is automatically tuned on development data set.

For the classification function, we employ a formula which combines word-based F-measure (denoted as $F_W(t,r)$) and a F-measure (denoted as $F_D(t,r)$) based on dependency triplet (dependency type, governor word, dependent word), as follows:

$$F(t,r) = \lambda \cdot F_W(t,r) + (1-\lambda) \cdot F_D(t,r) \quad (1)$$

where the free parameter λ is tuned on development data.

It is worth noting that, for languages which dependency parser is not available, we only use the word-based F-measure as the classification function. Specifically, we use Equation 1 for Into-English task, and the word-based F-measure for Out-of-English task in this paper.

In a scenario where there are multiple references, we compute the score with each reference, then choose the highest one. In addition, we treat the document-level score as the weighted average of sentence-level scores, with the weights being the reference lengths, as follows.

$$\text{Score}_d = \frac{\sum_{i=1}^{D} \text{len}(r_i)\text{Score}_i}{\sum_{i=1}^{D} \text{len}(r_i)} \quad (2)$$

where Score_i is the score of sentence i, and D is the number of sentences in the document.

3 Evaluation metrics

We apply our multi-level approach to two metrics. The first one is Meteor (Banerjee and Lavie, 2005), which has been widely used for machine translation evaluations. The second one is DREEM, a new metric based on distributed representations generated by deep neural networks.

3.1 Metric Meteor

We use the latest version of Meteor, i.e. Meteor Universal (Denkowski and Lavie, 2014) in this paper. Meteor computes a one-to-one alignment between matching words in a translation and a reference. The space of possible alignments is constructed by exhaustively identifying all possible matches of the following types: exact word matches, word stem matches, synonym word matches, and matches between phrases listed as paraphrases. Alignment is then conducted as a beam search.

From the final alignment, the translation's Meteor score is calculated as follows. First, content

and function words are identified in the hypothesis and reference according to a function word list. Next, the weighted precision and recall using match weights (w_i ...w_n) and content-function word weight (δ) are computed, as follows:

$$P = \frac{\sum_i w_i \cdot (\delta \cdot m_i(t_c) + (1-\delta) \cdot m_i(t_f))}{\delta \cdot |t_c| + (1-\delta) \cdot |t_f|} \quad (3)$$

$$R = \frac{\sum_i w_i \cdot (\delta \cdot m_i(r_c) + (1-\delta) \cdot m_i(r_f))}{\delta \cdot |r_c| + (1-\delta) \cdot |r_f|} \quad (4)$$

These two are then combined into a weighted harmonic mean, where a large α means recall is weighted more heavily.

$$\text{Fmean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} \quad (5)$$

To penalize reorderings, this value is then scaled by a fragmentation penalty based on the number of chunks and number of matched words.

$$\text{Meteor}(t, r) = (1 - \gamma \cdot (\frac{\#\text{chunk}}{\#\text{match}})^{\beta}) \cdot \text{Fmean} \quad (6)$$

In our studies, we fine-tune all the parameters for both multi-level and non-multi-level scoring frameworks.

3.2 Representation based metric

Distributed representations for words and sentences have been shown to significantly boost the performance of a NLP system (Turian et al., 2010). A representation-based translation evaluation metric, DREEM, is introduced in (Anonymous, 2015). The metric has shown to be able to achieve state-of-the-art performance, compared to popular metrics such as BLEU and Meteor. Therefore, in this paper, we also adapt this metric for our experiments.

In a nutshell, the DREEM metric evaluates translations by employing three different types of word and sentence representations: one-hot representations, distributed word representations learned from a neural network model, and distributed sentence representations computed with a recursive autoencoder (RAE). Two different RAE-based representations are used in this metric: one is based on a greedy unsupervised RAE, while the other is based on a syntactic parse tree. To combine the advantages of these four different representations, the authors concatenate them to form one vector representation for each sentence.

In detail, suppose that we have the sentence representations for the translations (t) and references (r). The translation quality is measured by DREEM with a similarity score computed with the Cosine function and a length penalty. Let the size of the vector be N. The quality score is calculated as follows.

$$\text{Score}(t, r) = \text{Cos}^{\alpha}(t, r) \times P_{len} \quad (7)$$

$$\text{Cos}(t, r) = \frac{\sum_{i=1}^{i=N} v_i(t) \cdot v_i(r)}{\sqrt{\sum_{i=1}^{i=N} v_i^2(t)} \sqrt{\sum_{i=1}^{i=N} v_i^2(r)}} \quad (8)$$

$$P_{len} = \begin{cases} exp(1 - l_r/l_t) & \text{if } (l_t < l_r) \\ exp(1 - l_t/l_r) & \text{if } (l_t \geq l_r) \end{cases} \quad (9)$$

where α is a free parameter, $v_i(.)$ is the value of the vector element, P_{len} is the length penalty, and l_r, l_t are lengths of the translation and reference, respectively.

To use this metric in the multi-level framework, we keep the parameter α consistent for all levels, but use different weights to combine the representations. That is, we construct the representation vector as follows:

$$V = < w_1 \cdot V_{oh}, w_2 \cdot V_{wd}, w_3 \cdot V_{gRAE}, w_4 \cdot V_{tRAE} > \quad (10)$$

where V_{oh} is the one-hot representation, V_{wd} denotes the word representations, and V_{gRAE} and V_{tRAE} are representations learned with greedy RAE and tree-based RAE, respectively. The weights w_1 ... w_4 are tuned on development data.

4 Experiments

4.1 Settings

We conducted experiments on the WMT metric task data. Development sets include WMT 2012 all-to-English, and English-to-all submissions. Test sets contain WMT 2013, and WMT 2014 all-to-English, plus 2013, 2014 English-to-all submissions. The languages "all" include French, Spanish, German, Czech and Russian. For training the word embedding and recursive auto-encoder model, we used WMT 2014 training data [1]. We used the English, French, German and Czech sentences in "Europarl v7" and "News Commentary" for our experiments. To train the representations for Russian, we used the "Yandex 1M corpus".

4.2 Results

Following WMT 2014's metric task (Machacek and Bojar, 2014), to measure the correlation with

[1] http://www.statmt.org/wmt14/translation-task.html

	Into-English	
metric	seg τ	sys γ
Original BLEU	–	0.821
Sentence BLEU	0.259	0.841
Original Meteor	0.279	0.849
Sentence Meteor	0.279	0.863
$Multi-level_w$ Meteor	0.285	0.871
$Multi-level_{wd}$ Meteor	0.294*	0.885*
DREEM	0.287	0.875
$Multi-level_w$ DREEM	0.293	0.880
$Multi-level_{wd}$ DREEM	0.303*	0.892*

Table 1: Correlations with human judgment on the WMT data for the Into-English task. Results are averaged on all into-English test sets. $Multi-level_w$ stands for only using word-based F-measure as the classification function, while $Multi-level_{wd}$ denotes the use of a combination of word-based F-measure and dependency triplet based F-measure. * indicates the improvement over the non-multi-level metric is statistically significant, with a significance level of 0.05.

human judgment, we employed Kendall's rank correlation coefficient τ for the segment level, and used Pearson's correlation coefficient (γ in the below tables) for the system-level. We tested the significance through bootstrap resampling (confidence level of 95%).

We tuned the weights for the Into-English and Out-of-English tasks separately. According to the tuned thresholds, about 25% of the translations are classified to low-quality translations, around 20% belong to high-quality translations, and the rest fall in the medium-quality category.

Experimental results conducted on the Into-English and Out-of-English tasks are reported in Tables 1 and 2. We also compared to the standard de facto metric BLEU (Papineni et al., 2002).

Results, as shown in Tables 1 and 2, indicate that the representation-based metric DREEM obtained better performance than BLEU and Meteor on both tasks at both segment and system levels. The multi-level versions of these metrics consistently improved the performance over the non-multi-level ones on both segment and system levels.

4.3 Further Analysis

In addition to showing the superior performance of the multi-level framework, our experiments also indicate the following observations.

Firstly, for BLEU and Meteor, document-level score computed by weighted averaging sentence-level scores can get better system-level correlation with human judgment, compared to that of the original document-level score which is computed from aggregate statistics accumulated over the entire document.

	Out-of-English	
metric	seg τ	sys γ
Original BLEU	–	0.843
Sentence BLEU	0.221	0.846
Original Meteor	0.228	0.845
Sentence Meteor	0.228	0.853
$Multi-level_w$ Meteor	0.234	0.861
DREEM	0.236	0.904$^\#$
$Multi-level_w$ DREEM	0.241	0.922$^{\star\#}$

Table 2: Correlations with human judgment on the WMT data for Out-of-English task. Results are averaged over all out-of-English test sets. $^\#$ indicates DREEM is significantly better than its corresponding version of Meteor, with a significance level of 0.05. * indicates the improvement over the non-multi-level metric is statistically significant.

task	low	medium	high
Into-En	0.93	0.81	0.75
Out-of-En	0.99	0.90	0.81

Table 3: The value of parameter α in multi-level Meteor.

Secondly, for Meteor, recall received a larger weight for low-quality translations than for high-quality translations. For instance, as depicted in Table 3, the parameter α in Meteor is higher for low-quality translations.

Finally, the syntax feature received higher weight for high-quality translations than for low-quality translations. In contrast, as shown in Table 4, the surface n-gram feature was assigned larger weight for low-quality translations .

task	low	medium	high
one-hot	0.23	0.11	0.05
word vec	0.42	0.42	0.40
greedy RAE	0.18	0.20	0.20
tree RAE	0.17	0.27	0.35

Table 4: The weights of each representation in the multi-level DREEM tuned for Into-English task. The syntax-based tree RAE representation received higher weight for high-quality translations, while one-hot representation received higher weight for low-quality translations.

5 Conclusions

Translations generated by statistical systems typically have a large variance in terms of their scores against human references. Motivated by such observation, we propose a multi-level framework. It enables a metric to deploy different criteria for various quality levels of translations. Our experiments on the WMT metric task data show that the multi-level strategy consistently improves the performance of two benchmarking metrics on both segment and system levels.

References

Anonymous. 2015. Representation based translation evaluation metrics. In *Proceedings of the Association for Computational Linguistics (ACL)*, Beijing, China, July.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

George Doddington. 2002. Authomatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference*, page 128132, San Diego, CA.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Ann Arbor, MI.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2011. Better evaluation metrics lead to better machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 375–384, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Chi-kiu Lo and Dekai Wu. 2011. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 220–229, Portland, Oregon, USA, June. Association for Computational Linguistics.

Matous Machacek and Ondrej Bojar. 2014. Results of the wmt14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA, June. 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 of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, July. ACL.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Matthew G. Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: Paraphrase, semantic, and alignment enhancements to translation edit rate. In *Machine Translation*, volume 23, pages 117–127.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *ACL*, pages 384–394.

Xiaofeng Wu, Hui Yu, and Qun Liu. 2013. DCU participation in WMT2013 metrics task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 435–439, Sofia, Bulgaria, August. Association for Computational Linguistics.