Fusion of Word and Letter Based Metrics for Automatic MT Evaluation

Muyun Yang, Junguo Zhu, Sheng Li, Tiejun Zhao
School of Computer Science and Technology, Harbin Institute of Technology, China
{ymy, jgzhu}@mtlab.hit.edu.cn; {lisheng,tjzhao}@hit.edu.cn

Abstract

With the progress in machine translation, it becomes more subtle to develop the evaluation metric capturing the systems' differences in comparison to the human translations. In contrast to the current efforts in leveraging more linguistic information to depict translation quality, this paper takes the thread of combining language independent features for a robust solution to MT evaluation metric. To compete with finer granularity of modeling brought by linguistic features, the proposed method augments the word level metrics by a letter based calculation. An empirical study is then conducted over WMT data to train the metrics by ranking SVM. The results reveal that the integration of current language independent metrics can generate well enough performance for a variety of languages. Time-split data validation is promising as a better training setting, though the greedy strategy also works well.

1 Introduction

The automatic evaluation of machine translation (MT) has become a hot issue for the MT circle. Compared with the slow and costly human evaluation, the automatic evaluation is more desirable for its low-cost and reusability. Moreover, with reliable automatic evaluation scores, researchers can tune parameters of the translation model instantly and consistently for better quality. For example, the statistical MT models are optimized by certain metric scores [Och, 2003; Chiang et al., 2008]. Rule-based MT can choose the best translation rules via the BLEU score [Kenji et al., 2003].

Practically, the work on automatic translation evaluation is targeted at a metric of high correlation to the human judgments. Two pioneering metrics BLEU [Papineni et al., 2002] and NIST [Doddington, 2002] are now the most widely adopted metrics in the open MT evaluation campaigns. Popular as they are, they were still observed an error case in NIST MT evaluation in 2005 even in the system level [Le and Przybocki, 2005]. Calllison-Burch et al. [2006] details the deficits of the BLEU, claiming that the simple surface similarity calculation between the machine and human translations suffers from morphological issues and misses certain important factors in human judgments.

Instead of reckoning on surface level matching, other approaches concentrate on integrating rich linguistic features to model what human may perceive for translation quality [Liu and Gildea, 2005; Owczarzak et al., 2007; Popović and Ney, 2009]. A recent work has further revealed that deeper linguistic techniques contributes to overcome the important surface differences between acceptable MT outputs and human references [Amigò et al., 2009]. Despite the positive effects in better correlation with manual judgments, an obvious drawback of these approaches is that the linguistic features are language dependent, which prevents the wide application of such built metrics.

As a result, the development of automatic MT evaluation metrics falls into a dilemma of choosing the string level matching for wide application or selecting the linguistic-rich strategy for better performance. Faced with an evaluation task involving multiple languages as in WMT campaigns [Callison-Burch et al., 2008; 2009; 2010], a language independent metric of high performance is more desirable.

To address this issue, this paper takes the thread of combining the string metrics for better evaluation performance at the sentence level. To compete with the linguistic approaches, the letter is suggested as a new unit of calculation for the string metrics, approximating the rich linguistic features from the perspective of finer granularity of modeling. Further, an empirical study on the feature selection is carried under the ranking SVM framework. Experiments on WMT data prove that the metric fusion can produce a robust and high-performance metric.

The rest of this paper is arranged as following: Section 2 introduces the related work of MT evaluations: string approaches, linguistic approaches and previous fusions by machine learning. Section 3 proposes the letter based calculation of the present string metrics. Section 4 describes the ranking SVM framework and a time-split data scheme for metric fusion. Section 5 demonstrates the good performance of the so-achieved metric in WMT08/09/10 data by an empirical study. Finally, Section 6 concludes this paper.

2 Related Work

Generally speaking, the present automatic MT evaluation metrics are based on the assumptions that good translations

are more close to the reference translations provided by experts. With the development of MT, it becomes more subtle to capture such similarity with the improvements in MT system.

2.1 MT Evaluation: String vs Linguistic

With regard to the key features involved, there are roughly two strategies in developing MT evaluation metrics: string surface similarity based approach (short for string approach hereafter) and the linguistic feature based approach (short for linguistic approach hereafter).

The string approach generally derives a score according to different perspectives to the string similarity between the candidate and reference translations. BLEU, for example, works by comparing machine translations with the expert translations in terms of N-grams shared between them. NIST and ROUGE [Lin and Och, 2004] improve this line by assigning information weight to N-grams or leveraging different types of N-grams, respectively. TER[Snover et al., 2006], WER [Nießen et al., 2000], PER [Tillmann et al., 1997] and GTM [Melamed et al., 2003] interpret the word matches from the point of the edit distance, the word error rate or the word alignment issue. Amazingly simple as these metrics, they have been widely adopted in practices due to good reliability and language independence. However, along with their good correlations with human judgments at the system level, they are often criticized for the defected performance in the sentence level [Blatz et al., 2003].

Linguistic approaches are dedicated to correlate better with human judgments at the sentence-level by various linguistic features. Liu and Gildea [2005] built STM metric by incorporating the syntax structure information. Owczarzak et al. [2007] and Kahn et al. [2009] included dependency structure information in the metric. Padó et al. [2009] further adopted the textual entailment features to model the semantic equivalence in the translation evaluation. Compared with metrics limited in lexical dimension, Giménez and Márquez [2007a] claimed that, metrics integrating deep linguistic information would be more reliable.

For a given MT evaluation, it is reasonable to employ language specific linguistic features since only the target language is involved in. Though it is likely that the automatic linguistic processing results are full of errors, it has been consistently proved such "un-reliable" linguistic knowledge could generalize well across different years' evaluation data [Albrecht and Hwa, 2008]. Therefore, the only defect in linguistic approaches lies with the less adaptability of linguistic features as well as the less availability linguistic tools for languages. In fact, even a POS-tagger for all the languages involved in current MT research would be practically prohibited. Therefore, in the public MT evaluation campaigns, linguistic approaches are hardly applied to substitute those string approaches.

2.2 Metric Combination by Machine Learning

A noticeable fact in the linguistic approaches is the employment of the machine learning to enable the superior sentence-level correlations with human evaluation. In fact, the automatic MT evaluation has almost become another experimental sandbox for various machine learning techniques.

[Corston-Oliver et al., 2001] treated this issue as a classification task and adopted a decision tree to classify between the human translation and the machine translated sentence. [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 [2008] proved that regression SVM based metrics generalize better across different years of data. Nonetheless, [Yang et al., 2007] and [Duh, 2008] argued that ranking was a more reliable approach for human to evaluate translations, and proved the SVM ranking worked well than SVM-regression.

An impressive observation among these studies is the best evaluation results are always achieved by combining linguistic features with other string metrics. In [Duh 2008], for example, a simple integration of the 30 (including 25 linguistic) features could generate 3.1—21.9% improvement on Pearson coefficient with respect to the single best string metric.

However, the linguistic approach does not carry everything before the combination of string metrics. In [Albrecht and Hwa, 2008], for example, the linguistic approach did not achieve a significant improvement over the combination of string metrics. Again in [Padó et al., 2009], the linguistic approach underperformed the string metric combination in the system level evaluation in Urdu (though pretty better in other cases). Table 1 details the above two cases for the specific correlation coefficients.

	String Only	Both (String+linguistic)
[Albrecht and Hwa, 2008]	0.499	0.512
[Padó, 2009] [Urdu : Sentence/ System]	0.501/0.927	0.556/0.810

Table 1: Cases where the string metric combination challenges the linguistic approach.

In this sense, we can see that the machine learning models are powerful in combination of different features, derived from whether string surface or linguistic depth, for a better evaluation metric. If properly designed, the string metrics are promising to generate competitive performance with those linguistic approaches, while preserving the advantage of language independence and wide applicability.

3 Letter Based String Metrics

One can naturally attribute the success of the linguistic approaches to the feature rich strategy in that more features usually bring forth better results. To compete with the rich information from linguistic features, string metrics simply needs a new perspective for informative features.

3.1 A Granular Perspective to Letter

Tracing back the study on string metrics, there already accumulates an exhaustive examination for various possible string metrics in terms of matching unit. TER, WER and PER

address the single word match issues. BLEU and NIST weight the adjacent multi word information. ROUGE takes into account the effect of non continuous multi-word.

These metrics are all calculated on word, the most popular lexical unit of most NLP systems. In contrast, in terms of information granularity, the linguistic approaches are actually describing the translation quality in a finer granularity with the increasing of linguistic features. That is to say, with the successive rich linguistic features integrated in the evaluation metrics, linguistic approaches can actually model the string similarity in a much sophisticated granularity space.

Contrary to the substantial features in linguistics, the surface string provides less feature space to explore. However, the fact is that the modeling granularity can be shifted in both the feature space and the calculation unit. In this sense, we are most probably to achieve an improved evaluation results by decreasing the calculation unit of the same model, say, from word to letter.

This hypothesis can be confirmed by the observation of sentence level performance: better performance of bi-gram than other multi-gram BLEU at the sentence level. Taking the data of WMT10 as an example, Table 3 provides the sentence level correlations of BLEU calculated by uni-gram, bi-gram, tri-gram and four-gram (indicated as Bleu_cum1, 2, 3, 4, respectively). The fact of "Bleu_cum2 > Bleu_cum3 > Bleu_cum4" in correlation coefficient indicates that, with small granularity of less word, BLEU method can actually generate better performance at the sentence level.

3.2 Letter Based String Metrics

Compared with word level calculation, the letter is somewhat less exploited in MT evaluation. To achieve a finer granularity of modeling without utilizing linguistic space, this paper suggests the idea of improving the current string metrics by a letter calculation.

In addition to the motivation to have a finer granularity, the letter based calculation has additional advantage of alleviating the influence of morphological inflections to word matching, which is promising in improving the performances of existing string metrics. Take the following two sentences as an example:

S1: Tom is interested in cooking.

S2: Tom has interest in cooker.

The word "*interested*" (in S1) and the word "*interest*" (in S2) would not be matched in original word match as in BLEU (without stemming and synonym information). In the letter based calculation of BLEU, the common string "*i n t e r e s t*" will be captured and awarded. In addition, the long words (which are usually the content word) will benefit from more letter n-grams, and short functional word will be partially under-weighted. In this sense, the method is an alternative to the synonymy dictionary or the stemmer tools. And since no additional linguistic information is required in the process, it may be calculated easier and faster.

The letter based calculation can be devised in certain different ways from the current word based metrics. However, aiming at examining the feasibility of letter based strategy, this paper simply calculates BLEU, NIST, ROUGE, GTM, PER, WER and TER with the letter unit (indicated with L_):

- BLEU: Lbleu-5
- NIST: Lnist-9
- ROUGE: Lrouge-6, Lrouge_L, Lrouge_SU, Lrouge_S, Lrouge_W
- GTM: Lgtm_1, Lgtm_2, Lgtm_3, Lgtm_4
- PER/WER: Lper, Lwer
- TER: Lter

It should be noted that we do not change the model of the above mentioned metrics except counting the letter match. Owing to the space limit, the details of those formulae are omitted here. As for the n-gram sizes of BLEU, NIST and ROUGE, they are roughly fixed by human (slight alternation of n around those figures would not change the performance significantly). The only exception is GTM, whose performance is not stable with letter calculation and the best four results are all kept for later experiments.

4 Metric Fusion by Ranking SVM

4.1 Ranking SVM

To build such a model, various machine learning techniques are qualified. For example, SVM is a popular choice when hypothesizing the translation evaluation modeling as either a classification, a regression, or a ranking task. Here we simply choose Ranking SVM as the framework of later experiments, which has been proved for its priorities over other machine-learning strategies [Duh, 2008].

Ranking SVM is designed to minimize the empirical Kendall's τ. For a given n sets of candidate translations, let r_i^* (i=1,2,…,n) denote the rankings of their qualities in each set of candidate translations. The learner's job is to choose a rank function f that can maximize

$$\tau_S(f) = \frac{1}{n}\sum_{i=1}^{n}\tau(r_i^f, r_i^*)$$

where r_i^f denotes ranking given by function f in each set.

Further details of ranking SVM based MT evaluation metric can be found in [Yang et al., 2007; Duh, 2008]. In this paper, we use SVM-Light toolkit [Joachims, 1999] to build Ranking SVM models. For all machine translations of the same source sentence, we produce their translation ranking scores by the built model.

Compared with those of millions of documents to rank for a query in information retrieval researches, ranking based MT metrics is not a serious challenge since one can easily find some clues in the hundred to thousand sample sentences to discriminate the MT systems under consideration. Nevertheless, it would be equally challenging if one wants a better metrics for a target data.

4.2 Metrics for Fusion

Features are crucial to machining learning models. To train a multilingual evaluation metric, here only the language independent information is considered. Of course, the string

metrics are good choices.

In addition to the previously mentioned letter based string metrics, the following word based string metrics under various parameters are chosen for fusion in the experiment:

- BLEU: 1-9, individual and cumulative
- NIST: 1-5, individual and cumulative
- ROUGE: 1-4, L, S, SU and W
- GTM: 1-4
- PER, WER and TER
- Meteor[1]

Most of these metrics are calculated with the public tools available (i.e. mteval-v11b, meteor-0.7, rouge-1.5.5, gtm-1.4, and tercom-7.25), with only PER and WER internally built according to Tillmann et al. [1997] and Nießen et al. [2000], respectively.

It remains an open issue to locate key features among these metrics since we are lack of explicit evidences on what human actually uses in perceiving the translation quality. Empirically, Giménez and Màrquez [2007] reported a success of unified combination of the metrics. Here we aim for a further optimized set of these metrics by a greedy search: to incrementally add in one feature with the most performance gain each time. Since such process is prone to local maxima, we resort to a good data allocation to counter balance this issue.

4.3 Data Split by Time

Inherited in the greedy feature selection, data over-fitting is not a trivial issue in addition to local maxima. To deal with this issue, larger data scale or n-fold cross-validation are commonly adopted.

Inspired by the good contribution of development corpus in recent SMT discriminative training, here we devised a strategy to split the training data by time. The motivation to this method is to capture the actual progress of MT system year by year. With the yearly round of technology competition, it is less likely that the previously detected errors still exist in the systems. Meanwhile, for a proved powerful feature for translation quality before, it is very likely to have included this feature in the MT training process, as demonstrated in [Padó et al., 2009]. That is another reason why previous metrics fail to evaluate so properly on current MT results.

In another aspect, the manual evaluation results are also accumulated yearly. For example, WMT campaigns actually provide manual evaluation results from the year 2006 to 2009, which can be then adopted for training the metric for WMT2010.

The basic idea is to split the training data by year: leaving the most updated data for the validation and keeping other data for training. The hypothesis is that a not-too-late data would best testify which metrics are still probably valid for current evaluation. For example, in preparation for WMT10 evaluation metric, an appealing setting would be reserving WMT09 data for validation and using other data for training.

[1] Meteor cannot be calculated in letter owing to its employment of synonym information.

5 Experiments

5.1 Data and Evaluation Method

This paper chooses an open campaign—the WMT shared evaluation task—as the experiment data for its public availability and the large data scale. WMT series involves the manual translation evaluations on English and other four languages: Czech(cz), French (fr), Spanish(es) and German (de). As for the training set, we mixed all the translation pairs without discriminating the language differences. For the testing, we report the results for each individual languages as well as the average results for "into English" and "out of English".

As an extreme to examine the "data split" strategy without relying on large amount of training data, here we restrict to the WMT08 [Calllison-Burch et al., 2008] and WMT09 [Calllison-Burch et al. 2009] data as the training set, and use the WMT10 [Calllison-Burch et al., 2010] as the testing set . This setting is exactly the same as the participants encountered in WMT10 shared evaluation task. The details of the training and testing data are provided in Table 2.

	Data ID	# of pairs	# of segments
Training Data	WMT08	34503	14275
	WMT09	35779	13757
Testing Data: WMT10	(en-cz)	11027	3803
	(en-fr)	6682	2666
	(en-es)	12334	3981
	(en-de)	4504	1788
	Total en-others	34547	12238
	(cz-en)	4378	1706
	(fr-en)	6433	2961
	(es-en)	8339	3743
	(de-en)	9342	3200
	Total others-en	28492	11610

Table 2: Statistics of WMT Data.

The WMT translation evaluation is conducted by manual ranking sentences relative to each other. And since no two translations are awarded an equal rank, the automatic evaluation results can be correlated with the human ranking with a variant of Kendall's τ according to [Calllison-Burch et al., 2010]:

$$\tau = \frac{\#\ of\ concordant\ pairs - \#\ of\ discordant\ pairs}{total\ pairs}$$

In the experiments, this Kendall's tau is computed and compared at the segment level only, which is the thorniest part of MT evaluation at present.

5.2 Results

5.2.1 String Metrics by Word and Letter

The first experiment is designed to compare the performance of current string metrics. Results including that of Meteor[Banerjee and Lavie, 2005] are provided in Table 3, where the upper half is the results of word based metrics and the lower half is the letter based calculations.

According to Table3, BLEU, ROUGE, WER and TER

have cases where letter based calculation is better than word, but NIST, GTM and PER cannot benefit from this unit alternation. For the English translations, a simple calculation of BLEU in letter raises this classical metric among the best. Equally powerful ones include the letter based ROUGE in two variants.

Metrics	"Others-to-English"				
	cz-en	fr-en	de-en	es-en	avg
Meteor	0.45	0.38	0.40	0.42	0.41
Bleu_cum_1	0.44	0.32	0.37	0.41	0.38
Bleu_cum_2	0.44	0.36	0.39	0.42	0.40
Bleu_cum_3	0.41	0.33	0.33	0.39	0.37
Bleu_cum_4	0.37	0.31	0.28	0.34	0.32
*Nist_cum_5	0.43	0.34	0.38	0.41	**0.39**
ROUGE_1	0.44	0.35	0.40	0.40	0.40
ROUGE_L	0.45	0.36	0.40	0.41	0.41
*ROUGE_S	0.45	0.36	0.41	0.41	**0.41**
*ROUGE_SU	0.45	0.36	0.41	0.41	**0.41**
ROUGE_W	0.44	0.36	0.40	0.41	0.40
*GTM	0.42	0.35	0.37	0.42	**0.39**
*PER	0.41	0.32	0.32	0.38	**0.36**
WER	0.38	0.34	0.32	0.39	0.36
TER	0.40	0.35	0.33	0.40	0.37
*Lbleu_cum_5	0.48	0.40	0.45	0.44	**0.44**
Lnist_cum_9	0.39	0.36	0.35	0.38	0.37
*Lrouge_6	0.45	0.37	0.43	0.41	**0.41**
*Lrouge_L	0.49	0.38	0.42	0.45	**0.44**
Lrouge_S	0.39	0.26	0.27	0.35	0.32
Lrouge_SU	0.39	0.26	0.27	0.35	0.32
*Lrouge_W	0.49	0.39	0.44	0.44	**0.44**
Lgtm_1	0.38	0.33	0.33	0.35	0.35
Lgtm_2	0.41	0.35	0.38	0.38	0.38
Lgtm_3	0.37	0.31	0.34	0.35	0.34
Lgtm_4	0.35	0.29	0.32	0.33	0.32
Lper	0.36	0.30	0.26	0.31	0.31
*Lwer	0.39	0.35	0.40	0.40	**0.39**
*Lter	0.44	0.37	0.40	0.42	**0.41**

Table 3: Word vs letter based string metrics in "other-en" track of WMT10 data. (*indicating the better metric of the same model)

As for the "en-others" data of WMT10[2], nearly all letter based calculations outperform (or at least equally) with the word based metrics. One possible reason is that those languages suffer more from morphological match owing to the rich inflections. In this sense, letter based calculation alone can improve the string metrics to certain extent.

5.2.2 Cross Validation and Data Split

The second experiment is designed to examine the proper handling of the data allocation for metric fusion. As for the two sets (WMT 08, 09) available, we first merge them as a larger scale of data for greedy search, and then resort the two-fold and five-fold cross validation. Besides, we further include a straightforward division of two data by the year, using WMT08 for training and WMT09 for test, and vice versa. The so-achieved metrics are tested on WMT10 data and the results are in Table 4.

According to Table 4, first of all, we can see a blind combination of all metrics does not guarantee a better result than the best single metric (at least for English translations).

Metrics	"Others-English"				
	cz-en	fr-en	de-en	es-en	avg
Best of Single Metric	0.49	0.40	0.45	0.45	0.44
(08+09)_all	0.46	0.38	0.41	0.44	0.42
(08+09)_greedy	0.52	0.41	0.46	0.47	0.46
(08+09)_5fold	0.51	0.42	0.46	0.47	**0.47**
(08+09)_2fold	0.52	0.41	0.46	0.47	0.46
08train, 09 dev	0.52	0.41	0.46	0.47	**0.47**
09train, 08 dev	0.50	0.41	0.46	0.47	0.46
	"English-Others"				
	en-cz	en-fr	en-de	en-es	avg
Best of Single Metric	-	-	-	-	0.40
(08+09)_all	0.39	0.46	0.36	0.45	0.41
(08+09)_greedy	0.40	0.46	0.38	0.44	0.42
(08+09)_5fold	0.40	0.46	0.38	0.45	0.42
(08+09)_2fold	0.40	0.46	0.38	0.44	0.42
08train, 09 dev	0.40	0.47	0.38	0.45	**0.43**
09train, 08 dev	0.40	0.46	0.37	0.44	0.42

Table 4: Fusion of string metrics under different conditions. (08_ or 09_ indicates to use the data separately and (08+09) indicates to use both of the data; Greedy refers to the "best in" feature augmentation, and 5fold to the five-fold corss validation)

Then, it is clear that the greedy search method could optimize the metric set for a better fusion result. Compared with the simple union of both training sets, the random n-fold cross validation is not of much help. As for the time based data split, it would not hurt the performance, at least. The best result appears[3] when the WMT09 data is left for validation, which partially supports our hypothesis on the function of the most up-dated data.

5.2.3 Compared with Linguistic Approaches

The next experiment compares the proposed methods with those linguistic related metrics submitted to WMT2010, including the Stanford, the DCU-LFG, the TESLA/TESLA-M, the IQmt-DR/IQmt-ULCh, the SEPIA and the SemPOS-BLEU. Details in these metrics can be found in [Calllison-Burch et al., 2010], and results are summarized in Table 5.

According to Table 5, the proposed method successfully outperforms the various linguistic approaches involved in this latest campaign, as well as the best result of WMT10. It is interesting that most linguistic approaches are not deployed for translations in other languages, which remind us the restrictions in developing a general-purpose linguistic metric for multiple languages.

5.3 Discussion

This section examines the optimized feature set, trying to provide an additional insight into the differences of the string metrics.

We first removes the letter calculation results from the combination and re-run all above the experiments. Although

[2] The details of this part are omitted - due to space limitations.

[3] The difference of the best result is significant at 99% level (p=0.01), with only the exception against "(08+09)_5fold" in "other-en" track, with p=0.10.

Metrics	"Others-English"				
	cz-en	fr-en	de-en	es-en	avg
Best of WMT10	0.50	0.41	0.46	0.46	**0.46**
Our approach	0.51	0.42	0.46	0.46	**0.47**
TELSA	0.45	0.40	0.43	0.44	0.43
IQmt-UCLh	0.45	0.38	0.39	0.43	0.41
Stanford	0.45	0.35	0.42	0.42	0.41
TELSA-M	0.40	0.34	0.39	0.39	0.38
SEPIA	0.41	0.33	0.36	0.41	0.38
IQmt-DR	0.37	0.35	0.36	0.40	0.37
SemPOS-BLEU	0.33	0.27	0.32	0.37	0.32
DCU-LFG	0.30	0.21	0.23	0.32	0.27
	"English-Others"				
	en-cz	en-fr	en-de	en-es	avg
Best of WMT10	0.40	0.45	0.37	0.43	**0.41**
Our approach	0.41	0.46	0.37	0.43	**0.43**
TELSA	0.26	0.39	0.29	0.38	0.33
Stanford	0.27	0.40	0.25	0.39	0.36
TELSA-M	0.30	0.36	0.31	0.41	0.34

Table 5: Comparison to linguistic metrics submitted to WMT10 as well as the best result in WMT10.

the best result appears under the same conditions again, a significant drop of 0.03~0.04 is observed for all combinations in both tracks. Therefore, the letter based string metrics contribute remarkably to the string metric combination.

As for the specific features selected by the ranking SVM for fusion, we have the followings for the best performance:

- Word level: Bleu_ind_1, Meteor, Rouge_L, TER
- Letter level: Lbleu_cum_5, Lgtm_4, Lrouge_L, Lrouge_S

This set lies in the core of all the metric sets after optimization. A fusion of these sets would bring in the following additional ones under different conditions:

- Word level: Bleu_ind_2, Bleu_cum_1, Bleu_cum_2, Bleu_cum_5, Bleu_cum_6, Rouge_W, GTM1
- Letter level: Lter, Lgtm_1, Lgtm_2, Lgtm_3

It is interesting that most of the selected metric is strong in itself. Candidates from NIST, PER and WER have never been selected. The reason may be that the information carried among them has been well provided by other metrics.

However, by no means the above metrics can be deemed as either fundamental or enough to MT evaluation. As for automatic MT evaluation, a paradox is that a new factor of evaluation can always be included in MT building, and hence losing much of its power in subsequent evaluations. Therefore, with machine learning technique, it is most probable that the evaluation metric inherently requires re-training with the progress of MT system.

6 Conclusion and Future Work

Various string similarity based metrics have been proposed, and then combined with different linguistics features for a better evaluation metric. To facilitate the good applicability to a variety of languages, this paper focuses on the combination of the string metrics exclusively for an equally good—if not better—performance to those successfully proposed linguistic approaches.

To augment the string metrics with new features, this paper suggests the letter based calculation of current string metrics. As a finer granularity of modeling, the letter based calculation itself improves most of the popular string metrics. In addition, letter based matching alleviates the morphological issue in the word based calculation of these metrics.

Taking the ranking SVM framework, this paper conducts an empirical study for the proper combination of the string metrics. A variant of the cross validation, i.e. the data split by time strategy, is described. The hypothesis derives from the observation that along with the progress in MT system, only the metric capable of discriminating most updated translations is promising to predicting further MT results. Experiments on WMT data confirm that the proposed method generates good results with the greedy feature selection, achieving the most consistently evaluations on translations in various languages (as compared with those provided latest implementations of the linguistic approaches).

Looking toward the future work, a proper handling of letter based string similarity is scheduled for a detailed examination. Also, other criteria for feature selection, e.g. PCA or information gain, is worth exploring. Last but not least, the linguistic features would be investigated, if possible, for their potential in further promoting metrics in addition to the letter and word level sting similarity information.

Acknowledgements

This work is supported by the National High Technology Research and Development Program of China (863 Program, No. 2011AA01A207), and the NSF China (No. 61272384 & 61105072).

References

[Albrecht and Hwa, 2008] Joshua S. Albrecht and Rebecca Hwa, Regression for Machine Translation Evaluation at the Sentence Level. *Machine Translation*. 22: 1-27. 2008.

[Amigó et al., 2009] Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Felisa Verdejo. The Contribution of Linguistic Features to Automatic Machine Translation Evaluation. In *proceedings ACL-IJCNLP 2009*, 2009.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. 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. 2005.

[Blatz and Kulesza, 2003] John Blatz and Alex Kulesza. Confidence Estimation for Machine Translation. *Technical Report, Johns Hopkins University, Natural Language Engineering Workshop*, 2003.

[Callison-Burch et al., 2008] Chris Callison-Burch,

Cameron Fordyce, and Philipp Koehn, Chrisof Monz and Josh Schroeder. Further Meta-evaluation of Machine Translation. In *Proceedings of WMT08*. pages: 70-106, 2008

[Callison-Burch et al., 2006] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of EACL 2006*, page: 249-256, 2006

[Callison-Burch et al., 2009] Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of WMT09*, 2009.

[Callison-Burch et al., 2010] Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan, Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of WMT10*, 2010. (*http://www.cs.jhu.edu/~ccb/publications/findings-of-the-wmt10-shared-tasks.pdf*)

[Chiang et al., 2008] David Chiang, Yuval Marton and Philip Resnik. Online Large-Margin Training of Syntactic and Structural Translation Features. In *Proceedings of ACL08*, pages: 224-233, 2008

[Corston-Oliver et al, 2001] Simon Corston-Oliver, Michael Gamon and Chris Brockett. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of the ACL01*, pages: 148-155, 2001.

[Doddington et al., 2002] George Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd HLT*, pages: 138-145, 2002.

[Duh, 2008] Kevin Duh. Ranking vs. Regression in Machine Translation Evaluation. In *Proceedings of WMT08* 2008.

[Giménez and Màrquez, 2007] Jesús Giménez and Lluìs Màrquez. A Smorgasbord of Features for Automatic MT Evaluation. In *Proceedings of WMT 07*, pages: 195-198, 2007.

[Joachims, 1999] Thorsten Joachims. Making large-scale SVM learning practical. *Advances in Kernel Methods-Support Vector Learning*. 1999. MIT-Press.

[Kahn et al., 2009] Kahn G. Jeremy, Mathew Snover and Mari Ostendorf, Expected Dependency Pair Match: Predicting Translation Quality with Expected Syntactic Structure. *Machine Translation*. 23: 169-179, 2009.

[Kenji et al., 2003] Imamura Kenji, Eiichiro Sumita, and Yuji Matsumoto. Feedback Cleaning of Machine Translation Rules Using Automatic Evaluation. In *Proceedings of the ACL03*, page: 311-318 , 2003.

[Kulesza and Shieber, 2004] Alex Kulesza and Stuart M. Shieber. A Learning Approach to Improving Sentence-level MT Evaluation. In *Proceedings of the 10th TMI*, 2004.

[Le and Przybocki, 2005] Audrey Le and Mark Przybocki. NIST 2005 Machine Translation Evaluation Official Results. In *Official Release of Automatic Evaluation Scores for all Submission*, 2005.

[Lin and Och, 2004] Chin-Yew Lin and Franz Josef Och. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of ACL04*. pages: 605-612, 2004.

[Liu and Gildea, 2005] Ding and Daniel Gildea. Syntactic Features for Evaluation of Machine Translation. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages: 25-32, 2005.

[Melamed et al., 2003] Dan I., Ryan Green, and Joseph P. Turian. Precision and Recall of Machine Translation. In *Proceedings of HLT-NAACL03*, 2003.

[Nießen et al., 2000] Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of LREC2000*, 2000.

[Och, 2003] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the ACL03*, pages: 160-167, 2003

[Owczarzak et al., 2007] Karolina Owczarzak, Josef van Genabith, and Andy Way. Evaluating Machine Translation with LFG Dependencies. *Machine Translation* 21(2): 95-119, 2007.

[Padó et al., 2009] Sebastian Padó, Daniel Cer, Michel Galley, Dan Jurafsky, and Christopher D. Manning. Measuring Machine Translation Quality as Semantic Equivalence: A Metric Based on Entailment Features. *Machine Translation* 23: 181-193, 2009.

[Papineni et al., 2002] Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL02*, page: 311-318, 2002.

[Popović and Ney, 2009] Maja Popović and Hermann Ney. Syntax-oriented Evaluation Measures for Machine Translation Output. In *Proceedings of WMT09*, pages, 29-32, 2009.

[Snover et al., 2006] Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, pages: 223–231, 2006.

[Tillmann et al., 1997] Christoph Tillmann, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. Accelerated DP-based Search for Statistical Translation. In *Proceedings of European Conference on Speech Communication and Technology*, 1997.

[Yang et al., 2007] Ye Yang, Ming Zhou and Chin-Yew Lin. Sentence Level Machine Translation Evaluation as a Ranking. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation*, 2007.