Automatic Meta-evaluation of Low-Resource Machine Translation Evaluation Metrics

Junting Yu[1], Wuying Liu[1], Hongye He[2], Lin Wang[3*]
1 Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangdong, CHINA
2 Training Department, Shaanxi provincial Party School of CPC, Shaanxi, CHINA
3 Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai, CHINA
junting_yu@163.com; wyliu@gdufs.edu.cn; hugh5945@163.com; lwang@xdsisu.edu.cn

Abstract—**Meta-evaluation is a method to assess machine translation (MT) evaluation metrics according to certain theories and standards. This paper addresses an automatic meta-evaluation method of machine translation evaluation based on ORANGE - Limited ORANGE, which is applied in low-resource machine translation evaluation. It is adopted when the resources are limited. And take the three n-gram-based metrics - BLEUS, ROUGE-L and ROUGE-S for experiment, which is called horizontal comparison. Also, vertical comparison is used to compare the different forms of the same evaluation metric. Compared with the traditional human method, this method can evaluate metrics automatically without extra human involvement except for a set of references. It only needs the average rank of the references, and will not be influenced by the subjective factors. And it costs less and expends less time than the traditional one. It is good for the machine translation system parameter optimization and shortens the system development period. In this paper, we use this automatic meta-evaluation method to evaluate BLEUS, ROUGE-L, ROUGE-S and their different forms based on Cilin on the Russian-Chinese dataset. The result shows the same as that of the traditional human meta-evaluation. In this way, the consistency and effectiveness of Limited ORANGE are verified.**

Keywords-Automatic Meta-evaluation; Limited ORANGE; BLEUS; ROUGE-L; ROUGE-S

I. INTRODUCTION

The evaluation of machine translation (MT) plays an important role in the field of machine translation. For system developers, it can assess system performance to learn the system problems and improve them in time. And for users, it can provide the basis for translation quality assessment. As is provided in the language specification released by the State Language Work Committee[1], there are mainly two kinds of machine translation evaluation: human evaluation and automatic evaluation. The human evaluation mainly determine the output quality through experts' subjective judgments of certain features such as adequacy and fluency. Automatic evaluation uses a computed sentence similarity to compare MT system outputs according to a set of reference translations. It is objective and not affected by external factors. However, the human method is very subjective, and will cost a lot of resources. Also, it is easy to be affected by external factors, and need a long period to finish the evaluation. All of these disadvantages cause the human evaluation unable to adapt to the fast MT system modification progress and parameter adjustment. This inevitably extends the system development period. So that it is difficult to provide efficient evaluation results for both developers and users. Therefore, researchers are more inclined to use the automatic method to evaluate the system performance quantitatively. Consequently, automatic evaluation has become one of the research hot spots in the machine translation.

There are so many automatic evaluation methods, and it is worth to assess their performances. For example, BLEU[2] can only be used at the corpus level, but smoothing BLEU (BLEUS) can be applied to sentence-level evaluation; BLEU does not consider the matching between the discontinuous subsequences, while ROUGE[3], based on continuous n-gram as BLEU, considers the subsequence with maximum length for candidate and reference matching; and so on. The quality of these automatic evaluation methods needs to be measured by a unified standard, that is, meta-evaluation of machine translation evaluation.

Meta-evaluation of machine translation evaluation, that is, using a certain method to detect which machine translation evaluation metric performance is better, mainly divided into two categories: human meta-evaluation and automatic meta-evaluation. The human meta-evaluation is to obtain the correlation coefficient by the adequacy and fluency scores of the automatic evaluation and the human evaluation score. It is expensive and the subjective factors always cause results inconsistent. What's more, it is difficult to achieve consistent for adequacy and fluency scores. The larger the correlation coefficient, the better the evaluation method performance. While the automatic meta-evaluation is to use the computer to assess the evaluation methods automatically. This method is objective, convenient and easy to implement, and will not be affected by external factors to cause the deviation of the evaluation results. Human meta-evaluation has many shortcomings due to manual intervention, and it is difficult to adapt to the information processing needs of massive data. As a result, it is very important to explore a general

[*] Corresponding Author

and objective automatic meta-evaluation method for the assessment of machine translation evaluation methods.

As the limited corpus resources, we propose a automatic meta-evaluation method named Limited ORANGE, which is based on ORANGE (Oracle Ranking for Gisting Evaluation) proposed by Lin[4] in 2004. With this method, we evaluate the performance of the current three automatic n-gram-based evaluation methods, BLEUS[4], ROUGE-L[5] and ROUGE-S[5]. Also, we evaluate some other forms of these three metrics, which consider semantic analysis based on Cilin[6]. In order to maintain consistency with the human meta-evaluation, we propose to evaluate the automatic evaluation metric performance by scoring and sorting the candidate and the reference through the adequacy score, fluency score and translation similarity. Except for the artificial reference translations, the whole evaluation process does not require additional manual intervention, and can be well applied to sentence-level evaluation.

II. LIMITED ORANGE AUTOMATIC META-EVALUATION METHOD

A. Brief introduction of used machine translation evaluation metrics

a) Smoothing BLEU (BLEUS)

Since its introduction by Papineni et al. in 2002[2], BLEU has been widely used in various evaluation activities and has many variants[7]. BLEU is calculated through matching the continuous n-grams between the candidate and a fixed set of references, and then get the precision of n-gram, where n is set 1 to 4 usually. Then take the geometric mean of the modified n-gram precisions and then multiply the result by a brevity penalty factor (BP) to punish candidate that is shorter than reference. So BLEU is defined as:

$$\mathrm{BLEU} = \mathrm{e}^{\min(1-r/c,0)} \times \exp(\textstyle\sum_{n=1}^{N} w_n \log p_n) \tag{1}$$

Where N is the maximum number of n-gram, p_n is the n-gram precision, with the w_n=1/N to be corresponding weight. In the brevity penalty case, r is the length of reference, c is that of candidate.

Although it is widely used as golden standard, BLEU lacks stable, reliable, and valuable sentence-level scores, which is critical to distinguishing system performance. Therefore, it is important for BLEU to avoid zero-precision by smoothing techniques.

In 2004, Lin has proposed smoothing BLEU (BLEUS) for the first time[4]. In this paper, we adopt +1 smoothing in each n-gram precision, as defined in equation (2), to solve the zero- precision.

$$p_n = \frac{\mathrm{Count}_{\mathrm{clip(n-gram)}} + 1}{\mathrm{Count}_{\mathrm{(n-gram)}} + 1} \tag{2}$$

$\mathrm{Count}_{\mathrm{(n\text{-}gram)}}$ and $\mathrm{Count}_{\mathrm{clip(n\text{-}gram)}}$ are respectively the minimum n-grams co-occurring in reference and candidate.

b) ROUGE-L

When matching the continuous n-grams in candidate and reference, BLEU does not describe well the relationship between long-distance discontinuous words. In this respect, in 2004, Lin proposed a method named Recall-Oriented Understudy for Gisting Evaluation (ROUGE) based on discontinuous n-grams to describe the relationship between long-distance units, such as ROUGE-L and ROUGE-S.

ROUGE-L is defined to measure sentence-to-sentence similarity based on the longest common subsequence (LCS) statistics between a candidate and a set of references. This metric considers both the precision and the recall, and uses the LCS-based F-measure to calculate the similarity between the reference X, with length of m, and the candidate Y with length of n.

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{3}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{4}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{5}$$

Where LCS(X,Y) is the length of the LCS of X and Y, and when $\partial F_{lcs}/\partial R_{lcs}=\partial F_{lcs}/\partial P_{lcs}$, $\beta= P_{lcs}/R_{lcs}$, which is named relative weight. The equation (5) is ROUGE-L. And the ROUGE-L is 1 when X=Y, while the value is 0 when LCS(X,Y)=0, that is, the candidate Y and the reference X are completely different.

c) ROUGE-S

As ROUGE-L only records the longest common subsequence, and does not record the co-occurrences shorter than the longest one, this method cannot describe translation similarity well. ROUGE-S calculates the similarity between candidate and reference based on the discontinuous skip-bigram, which is any pair of words in their sentence order, allowing for arbitrary gaps. The skip-bigram is similar to the 2-gram in BLEU, but contains discontinuous subsequences. Thus the ROUGE-S can get the translation similarity well. Each sentence with the length of len has C(len, 2) skip-bigrams, where C is combined function. For Chinese, len stands for the number of words after segmentation. Given reference translation X of length m and candidate translation Y of length n, we compute skip-bigram-based F-measure, which is ROUGE-S, as shown in equation (8):

$$R_{skip2} = \frac{SKIP2(X,Y)}{C(m,2)} \tag{6}$$

$$P_{skip2} = \frac{SKIP2(X,Y)}{C(n,2)} \tag{7}$$

$$F_{skip2} = \frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2} + \beta^2 P_{skip2}} \tag{8}$$

SKIP2(X,Y) is the number of skip-bigram matches between X and Y. Relative weight $\beta=P_{skip2}/R_{skip2}$, when $\partial F_{skip2}/\partial R_{skip2}=\partial F_{skip2}/\partial P_{skip2}$.

d) Semantic Space Transformation-based metrics

Machine translation is actually a different encoding of the "same semantics". When automatically evaluating the translation quality of the machine translation system, if the two words in the translations are not exactly the same, they are judged as having zero similarity, which reduces the similarity of the translations. Then this will affect the evaluation performance. So Junting Yu[8] proposed an improved Cilin-based smoothing BLEU (BLEUS-syn) metric with Semantic Space Transformation (SST) algorithm.

The SST algorithm is mainly based on the unigram of the reference. When the candidate and the reference are matched to form a mapping, the exact morphological matching is first performed, and then the synonymy matching based on Cilin is performed. The two stages are carried out in order without overlapping.

BLEUS-syn mainly introduces the synonymy matching into BLEUS. When matching the unigram of candidate and reference, it uses exact matching first, and then synonymy matching based on Cilin, in which stage candidate unigram is replaced by reference unigram. In the 2~4-grams matching, the n-gram series are based on replaced unigram. So the evaluation performance maybe improved obviously.

Also, applying the SST algorithm into the ROUGE-L and ROUGE-S will obtain the ROUGE-L-syn and ROUGE-S-syn metrics respectively.

B. Limited ORANGE

A good evaluation metric should give a higher score to a good translation than a bad one. So a good translation should be ranked higher than a bad one according to their scores. Taking two assumptions that references are good translations and the more a candidate is similar to its references the better; references are usually better than candidates. Therefore, references should be ranked higher than candidates on average if a good automatic evaluation metric is used.

As the short of some data for specific tasks or data that can be applied to restricted fields, the training of statistical machine translation models still faces serious "data sparse" problems[9]. In view of the subjectivity, long period and weak consistency of the human meta-evaluation method, this paper proposes the Limited ORANGE, based on ORANGE, as a automatic method to assess evaluation metrics performances in the case of low-resource language. This method makes automatic meta-evaluation possible because of its objective, concise and convenient implementation, which reduces the contribution of human resources.

The Limited ORANGE is proposed in the absence of corpus resources. When the corpus resources are relatively limited and the scale of the training set is not large enough, the output n-best candidate translation list is poorly readable. So the Limited ORANGE utilizes limited corpus resources and assess the metric performance automatically with the help of online translation systems. In this paper, we intend to use the output Chinese translations of the widely used Russian-Chinese online translation systems on the network as a list of candidate translations, which is set as n-best translations, and Chinese translations in Russian-Chinese aligned corpus as reference translations. Then, we sort the list of candidates and reference, calculate the rank of the reference in the n-best list, and then get the ratio of the reference rank to the length of the n-best list as the Limited ORANGE score.

Given a source sentence S, select four online translation systems Google, Baidu, Bing, and Youdao, to assist implementing the meta-evaluation research. Take the Chinese outputs of above four online systems as candidates, the corresponding Chinese in Russian-Chinese bilingual news sentence alignment corpus as reference, to construct the experimental corpus. Then each automatic evaluation method is calculated as follows:

(1) Select translation features and calculate translation scores of candidate list and reference;

(2) Sort the reference and the candidate list based on the translation scores, and calculate an average rank of the reference;

(3) Calculate the ratio of the average rank of the reference to the length of the translation list, which is the Limited ORANGE score shown as follows:

$$\mathrm{Score} = \frac{\sum_{i=1}^{S} \mathrm{Rank}(\mathrm{Oracle}_i)}{S(N+1)} \qquad (9)$$

Where, $\mathrm{Rank}(\mathrm{Oracle}_i)$ is the rank of the reference of the source sentence i in the n-best list, S is the number of source sentences in the corpus, N is the length of translation list. The smaller the ratio, the better the automatic evaluation metric performance.

C. Translation sorting method

Through the above analysis, the most important question is how to comprehensively sort the translations? We choose a set of features to represent the translation of the same source sentence. Each feature describes an attribute of the translation. The feature weight indicates the relative importance of the feature to the translation. Any information related to the translation can be encoded as a feature. In order to maintain consistency with traditional human meta-evaluation, we intend to select the 3 features to represent translation information, such as the translation adequacy score - ade, the fluency score - flu and the translation similarity based on the automatic evaluation method - sim. These features must be obtained with the help of linguistic knowledge.

The principle of statistical machine translation is to model, train and then decode through translation models and language models[10], as shown in Figure 1.

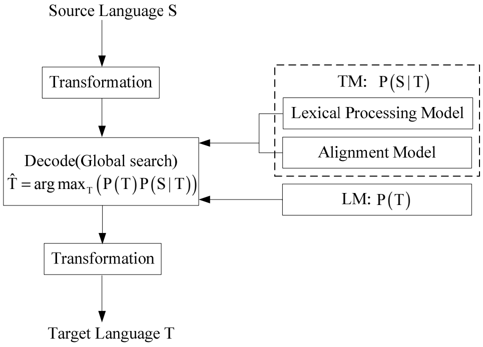


Figure 1. Framework of statistical machine translation system.

And then find the sentence with the highest probability in all possible target T as the translation result of the source S[10]. According to the Bayesian formula, it is to find a closest target sentence $\breve{T}$ by decoding to the source S.

$$\breve{T} = \arg\max_T P(T)P(S \mid T) \qquad (10)$$

Where, P(S|T) is the translation model (TM), including lexical processing and word alignment. It indicates the degree to which the target language T is like the source language S. P(T) is the language model (LM), independent of the source language, indicating the degree to which T is like a sentence in a target language, reflecting the fluency of the translation.

a) Translation adequacy score calculation

The translation adequacy score is expressed by the translation model probability P(S|T), that is, the conditional probability of source language S when given the target language T. The higher the probability value, the higher the translation adequacy score. We train the translation model with the Moses system[11], and use the GIZA++ to achieve word alignment.

Since the experiment corpus comes from four different online Russian-Chinese translation system translations and human reference translations, they are different from the n-best list from the same open source statistical machine translation system. Therefore, it is necessary to find a common "reference point" for the four candidates from different systems and the human reference to obtain their TM probabilities. We select the 1-best translation of the Moses system as the "reference point" to indirectly obtain the adequacy scores of the five translations.

(1) Obtain the 1-best of the Moses and its corresponding TM probability P(S|T).

(2) Compare the similarity - m_i of the 1-best T with the five translations c_i (i=1, 2, 3, 4, 5) to be compared respectively. Chinese, as the target language, is a analytic language. So word order is important for similarity calculation. Then we take Levenshtein Distance to calculate the m_i. When using the Levenshtein distance to match, the exact matching is first performed. If the word forms are different, then the synonymy matching based on Cilin is considered. The two matching steps are performed in a sequence without overlap.

$$m_i = 1 - \frac{d}{\max(l_T, l_{c_i})} \tag{11}$$

Where d stands for the cost of the Levenshtein distance, l_T and l_{ci} represent the length of the 1-best translation T and the translations c_i.

(3) Get the adequacy score-ade_i by multiplying step (1) and step (2), and this is the TM probability of the translation c_i.

$$ade_i = P(S \mid T) \bullet m_i \tag{12}$$

b) Translation fluency score calculation

The translation adequacy score is reflected by the LM probability P(T), which indicates the likelihood that the sequence will be expressed in the target language. Currently the most widely used is n-gram to calculate the probability of the sequence $W=w_1,w_2,\cdots,w_n$.

$$p(W) = p(w_1^n) \approx p(w_1)p(w_1 \mid w_2)\cdots p(w_n \mid w_{n-2}^{n-1}) \tag{13}$$

Due to the limited training corpus, there will be "data sparse". Smoothing techniques are needed to discount the visible events count and give them invisible events. Thus all probabilities are non-zero. We adopt the Katz smoothing[12] to train the language model on the basis of the SRILM toolkit. Then apply the 4-gram model obtained to the Moses system, and get the language model score - $P_{LM}(T)$. We can get the language model probabilities of the translation c_i by the same method as described in TM score.

$$flu_i = P_{LM}(T) \bullet m_i \tag{14}$$

c) Translation similarity calculation

The translation similarity calculation is based on the automatic evaluation metrics. Assume the 1-best of the Moses system as the reference r, and the other 5 translations are the candidate c_i (i=1~5), then the similarity of translation i is shown as follows.

$$sim_i = M_k(r, c_i) \tag{15}$$

Set an appropriate weight for each feature, the final scores of the five translations are:

$$Score_i = \lambda_1 \bullet ade_i + \lambda_2 \bullet flu_i + \lambda_3 \bullet sim_i \tag{16}$$

We set the three features the same weight, that is, $\lambda_1=\lambda_2=\lambda_3=1/3$.

III. Performance Analysis of LIMITED ORANGE

A. Corpus and environment

In the experiment, we apply 52,892 Russian-Chinese bilingual alignment sentence pairs in the domain of news[13]. The corpus is divided into training set and test set by hierarchical sampling. The test set includes 1,057 Russian-Chinese pairs. At the same time, a total of 247,075 entries in Russian-Chinese dictionary were introduced in model training to optimize the bilingual phrase alignment model.

We select four online translation systems, such as Google, Baidu, Bing, and Youdao, which are widely used and have readable translations in Russian-Chinese translation. Take the Chinese outputs in December 2015 as candidate list, and the human Chinese sentences in corpus as references.

We run the experiment on the computer with 8.00GB memory and Intel(R) Core(TM) i7-6700HQ CPU to maintain consistency of results. The experiment is implemented using JAVA on the eclipse platform.

B. Result and discussion

First, on the test set, we take these methods on the four online translation systems, Google, Baidu, Bing, and Youdao, and get the BLEU values. Table I shows the experiment results.

TABLE I. The BLEU values of BLEUS and BLEUS-syn on test set.

BLEU Value	Google	Baidu	Bing	Youdao
BLEUS	0.197001	0.214467	0.176529	0.190993
BLEUS-syn	0.209472	0.221200	0.194016	0.197361

Table I shows that, on the test set, (1) The evaluation value of BLEUS-syn is higher than that of the BLEUS for the 4 online systems; (2) When using the SST algorithm based on Cilin, the BLEU value of Baidu increases from 0.214467 to 0.221200, with an increase of 3.14%; the BLEU value of Google increases from 0.197001 to 0.209472, with an increase of 6.33%; the Youdao BLEU is increased by 3.33% from 0.190993 to 0.197361; the one of Bing is increased by 9.91% from 0.176529 to 0.194016; (3) After applying the SST algorithm, the performance of each system is improved with different degrees. All of the above analysis, BLEUS-syn performance is better than BLEUS.

In the same way, we get the F value of ROUGE-L and ROUGE-L-syn on the test set, shown as Table II.

TABLE II. The F values of ROUGE-L and ROUGE-L-syn on test set.

F Value	Google	Baidu	Bing	Youdao
ROUGE-L	0.277583	0.277276	0.254985	0.257065
ROUGE-L-syn	0.339582	0.343511	0.312607	0.319314

As shown in Table II, on the test set, the F value of ROUGE-L-syn is higher than that of ROUGE-L. When using the synonymy matching, the performance of systems have different improvements obviously, for example, Google 22.34%, Baidu 23.89%, Bing 22.60%, and Youdao

24.22%. So the ROUGE-L-syn performance is better than ROUGE-L.

Then we experiment with ROUGE-S and ROUGE-S-syn on the test set to obtain the F values of 4 systems and corresponding improvements by using SST. The results are listed in the Table III.

TABLE III. THE F VALUES OF ROUGE-S AND ROUGE-S-SYN ON TEST SET.

F Value	Google	Baidu	Bing	Youdao
ROUGE-S	0.189192	0.206467	0.167508	0.168972
ROUGE-S-syn	0.233596	0.240100	0.210788	0.202358
Improvement after SST	23.47%	16.29%	25.84%	19.76%

As shown in Table III, the performance of ROUGE-S-syn, which uses SST algorithm is superior to that of ROUGE-S.

Then we achieve vertical comparison with the different forms of the same metric, such as BLEUS and BLEUS-syn, ROUGE-L and ROUGE-L-syn, ROUGE-S and ROUGE-S-syn. Table IV shows the Limited ORANGE scores of these metrics.

TABLE IV. THE LIMITED ORANGE SCORES OF THE VERTICAL COMPARISON.

Metric	BLEUS	BLEUS-syn
Limited ORANGE Score	0.003227	0.003112
Metric	ROUGE-L	ROUGE-L-syn
Limited ORANGE Score	0.003112	0.003097
Metric	ROUGE-S	ROUGE-S-syn
Limited ORANGE Score	0.003160	0.003153

The score of BLEUS is 0.003227, and its variant BLEUS-syn using the SST algorithm, has the score of 0.003112, smaller than that of BLEUS. This shows the performance of BLEUS-syn is better than that of BLEUS, which is consistent with the result demonstrated in Table I. And the metrics with SST algorithm, such as ROUGE-L-syn and ROUGE-S-syn, are better in performance than the baseline of ROUGE-L and ROUGE-S respectively. All of these results are the same as the above research shown in Table II and Table III.

The above sets of experimental analysis results show that the Limited ORANGE meta-evaluation method is consistent with the results of the traditional vertical evaluation to assess the different evaluation methods of the same type. This indicates that the Limited ORANGE method is effective for the performance evaluation of vertical comparison.

Secondly, we evaluate the performance of BLEUS, ROUGE-L and ROUGE-S with Limited ORANGE, which is called horizontal comparison. The scores of the three metrics are shown in Table V.

TABLE V. THE LIMITED ORANGE SCORES OF THE HORIZONTAL COMPARISON.

Metric	BLEUS	ROUGE-L	ROUGE-S
Limited ORANGE Score	0.003227	0.003112	0.003160

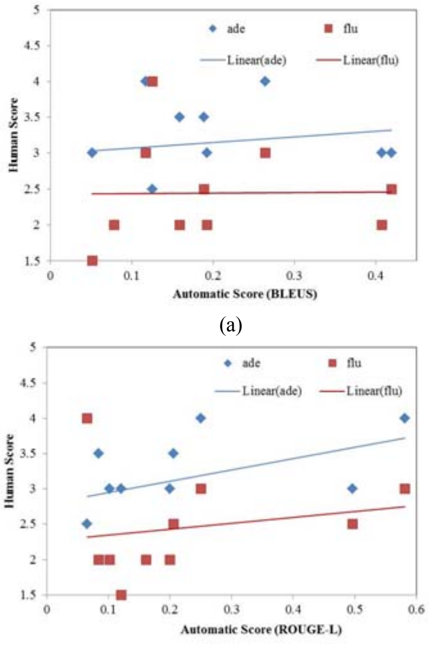
As shown in Table V, ROUGE-L has the lowest score of 0.003112. ROUGE-S is 0.003160, BLEUS is the highest 0.003234. As analyzed above, the lower the Limited ORANGE score, the better the evaluation metric performance. Therefore, the ROUGE-L performance is the best, the ROUGE-S is second, and the performance of BLEUS is the worst. The main reason is that during the evaluation, ROUGE-L compares the translation similarity based on the longest common subsequence, and does not set the fixed length of the n-gram. The matching is flexible and not limited to the length of the n-gram. In this case, it will not affect the number of matching n-grams in the translations. ROUGE-S and BLEUS well reflect the matching of each n-gram. ROUGE-S combines the advantages of ROUGE-L and BLEUS, which includes both the continuous n-gram of BLEU and the discontinuous n-gram of ROUGE-L. ROUGE-S can not only ensure the adequacy of the translation, but also capture information of distant words. In a word, the performance of ROUGE-S is better than that of BLEUS.

To verify the utility of the automatic meta-evaluation method, we adopt the Pearson correlation coefficient r_{xy} to evaluate the three metrics, which is to measure the quality of evaluation metrics by determining the correlation between the scores of the evaluation metrics and scores of fluency and adequacy. Suppose the data point on the test set containing the variable automatic score x and the human score y is $\{(x_i, y_i)\}$, the Pearson correlation coefficient is defined:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \qquad (17)$$

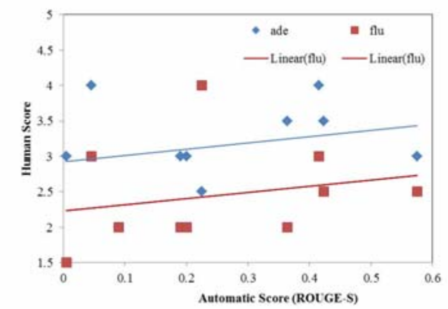
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the averages of sample X and Y respectively. The variable n is the source sentences number of test set. The higher the r_{xy}, the better the evaluation metric performance.

According to the results of vertical evaluation, the performance of Baidu system is the best. So we choose the automatic evaluation scores of Baidu to obtain the Pearson correlation coefficient r_{xy}. Figure 2 shows the correlation of automatic evaluation and human evaluation.



(a)

(b)



(c)

Figure 2. The correlation of automatic evaluation and human evaluation.

The x-axis indicates automatic scoring based on different automatic evaluation methods, and the y-axis is human score. The adequacy and fluency score are expressed by ade and flu respectively, which are added linear trends. (1) Both the correlation of BLEUS adequacy and fluency are the worst of the three methods. (2) The fluency correlation between ROUGE-L and ROUGE-S is basically the same. But the ROUGE-L adequacy correlation is better than that of ROUGE-S. (3) The correlation of ROUGE-S adequacy and fluency is equally good. It can be concluded that the performance of ROUGE-L is better than that of ROUGE-S and BLEUS, and the performance of BLEUS is the worst. So the experimental results of Limited ORANGE are consistent with the results of the human meta-evaluation. Thus, we verify the consistency of the Limited ORANGE method.

But at the same time, it can be seen that the human meta-evaluation involves optimization of the two objective functions of the adequacy and fluency correlation coefficients, which are always inconsistent difficultly. The comprehensive evaluation results based on the two objective functions have a lot to do with the subjective factors of people's scores. But the Limited ORANGE has only one objective function to optimize, which makes the meta-evaluation of MT evaluation operability perfectly reflected because of its objectivity. The meta-evaluation results are objective and concise. It can verify the improvement of the evaluation method performance more effectively and quickly. It can also be applied to research the languages with low-resource or in some specific fields or some special tasks. Thanks to its automation, it can free up some human resources.

IV. CONCLUSIONS

This paper mainly proposes a Limited ORANGE automatic meta-evaluation method based on ORANGE to assess the performances of the automatic MT evaluation metrics. Suppose this method works through the average rank of reference translations, without any manual intervention except the prepared references. The higher the ranking, the better the performance of metric. Compared with the traditional human meta-evaluation, the method proposed in the paper can significantly shorten evaluation time and save energy.

We take vertical comparison and horizontal comparison with BLEUS, ROUGE-L, ROUGE-S and their variants using SST algorithm on the test set. In order to be consistent with the traditional human meta-evaluation, three factors of the adequacy score, the fluency score and the translation similarity based on the evaluated metrics are selected to represent the translation in the process of evaluation. Both the vertical comparison and horizontal comparison are verified consistent with traditional human meta-evaluation. So the Limited ORANGE automatic meta-evaluation method is proved to be effective and feasible.

Further research will concern that further optimizing the selection and calculation of translation features. And choose some other languages to experiment to increase the generalization of this method.

ACKNOWLEDGMENT

The research is supported by the Social Science Foundation of Shanghai (No. 2019BYY028), the Key Project of State Language Commission of China (No. ZDI135-26), the Natural Science Foundation of Guangdong Province (No. 2018A030313672) and the Key Project of Guangzhou Key Research Base of Humanities and Social Sciences: Guangzhou Center for Innovative Communication in International Cities (No. 2017-IC-02).

REFERENCES

[1] Assessment Specifications of Machine Translation Systems. GF 2006

[2] Papinen K., Roukos S., Ward T., et al. BLEU: a method for automatic evaluation of machine translation [C]. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics 2003 (ACL 2003), 2002: 311-318.

[3] Lin C. Y.. ROUGE: A Package for Automatic Evaluation of Summaries [C]. In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. 2004.

[4] Lin C. Y., Och F. J.. ORANGE: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation [C]. In Proceedings of the International Committee on Computational Linguistics 2004 (COLING-2004), 2004.

[5] Lin C. Y., Och F. J.. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics [C]. In Proceedings of Association for Computational Linguistics 2004 (ACL 2004), 2004.

[6] Mei J., Zhu Y., Gao Y., et al. Tongyici Cilin (Extended) [M]. HIT IR-Lab. 1996.

[7] Yvette Graham, Barry Haddow, Philipp Koehn. Translationese in Machine Translation Evaluation. arXiv:1906.09833 [cs.CL]. 2019

[8] Junting Yu, Wuying Liu, Hongye He, et al. BLEUS-syn: Cilin-Based Smoothed BLEU [C]. In the 12th China Workshop on Machine Translation (CWMT 2016), 2016: 102-112

[9] ZHANG Bo. The Computational Models of Natural Language Processing [J]. Journal of Chinese Information Processing, 2007, 21(3): 3-7

[10] Feng zhiwei. Formal Models of Natural Language Processing [M]. University of Science and Technology of China Press. 2010.1: 564-565

[11] Koehn P.. Moses-Statistical Machine Translation System-User Manual and Code Guide [A]. 2015

[12] Katz S. M.. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer [J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1987, 35(3): 400-401

[13] Du W., Liu W., Yu J., et al. Russian-Chinese Sentence-level Aligned News Corpus [C]. In Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015), 2015: 213.