BLEUS-syn: Cilin-Based Smoothed BLEU

Junting Yu[1], Wuying Liu[2(✉)], Hongye He[1], and Mianzhu Yi[1]

[1] Luoyang University of Foreign Languages, Luoyang, China
junting_yu@163.com, hugh5945@163.com,
13373781261@163.com
[2] Laboratory of Language Engineering and Computing,
Guangdong University of Foreign Studies, Guangzhou, China
wyliu@gdufs.edu.cn

Abstract. Machine Translation (MT) evaluation is very important for a MT system. In this paper, we investigate an improved Cilin-based smoothed BLEU (BLEUS-syn). As the possible cases that the short translation or English abbreviations in candidate may cause unigram have no matches, this evaluation metric smoothed the traditional BLEUS n-gram. It applied synonym substitution in unigram matching, and calculated the other 2–4-gram. It performed experiments in Russian and Chinese bilingual sentence data set and evaluated the output translations of online translation systems such as Google, Baidu, Bing and Youdao. The experimental results show that the effectiveness of our BLEUS-syn and traditional BLEUS are consistent. The performance of Baidu is the best, that of Youdao is the second, and that of Bing is the worst. Using BLEUS-syn can greatly enhance the performance of traditional BLEUS evaluation. It makes the Baidu BLEUS value improve 6.81%, Youdao improve 6.98%, Google 7.82%, and Bing 7.55%.

Keywords: Cilin · Evaluation · BLEU · BLEUS

1 Introduction

With the popularity of Internet and the arrival of the era of big data, Internet languages become more and the contacts of various countries become more frequent. As the main platform of information, the language translation becomes the key factor. With the multi-language information processing, traditional human translation cannot meet the daily needs, and the machine translation, which can translate multi-language automatically, become a hot topic. With the continuous development of information technology, translation quality, various machine translation systems and models appear. The translation quality and performance index become critical for the machine translation system. The evaluation of MT systems becomes important for the research of machine translation.

Machine translation evaluation usually refers to the quantitative evaluation of a given translation system. It can evaluate the system performance and enable the developers to learn the problems and improve it in time. The text evaluation of MT system mainly has two kinds of human evaluation and automatic evaluation, which is provided in the language specification released by the State Language Work Committee [1]. The human evaluation is mainly scoring the adequacy and fluency of system output by language

M. Yang and S. Liu (Eds.): CWMT 2016, CCIS 668, pp. 102–112, 2016.
DOI: 10.1007/978-981-10-3635-4_9

experts subjectively according to references. But the human translation has many shortcomings, such as strong subjectivity, expensive, easy to be affected by external factors, the long evaluation period and so on. These will cause the human translation unable to adapt to the progress of MT system modification and parameter optimization, extend the system development period, and difficult to provide developers and users with efficient evaluation. As a result, researchers prefer the automatic evaluation.

Automatic evaluation methods can be divided into three categories: the linguistic point of detection, string similarity and machine learning. The linguistic point of detection method proposed by Professor Y Shiwen [2], is not widely used because it doesn't consider the whole of translation and tests the corresponding part of the translation according to the prior definition of a good linguistic test points, which cost higher. The method based on the string similarity becomes the most widely used evaluation method among the single metric evaluation. The best one is BLEU, which is proposed by Papineni [3] in 2002. BLEU matches the n-gram between candidate and reference, and the more n-gram match, the higher score is. Then researchers have made a lot of improvements against the problems such as not applied to sentence level and lack of recall. The most famous and widely used are the smoothed BLEU (BLEUS) [4], ROUGE-N [5] and METEOR [6]. The machine learning method develops fast as the emergence of deep neural networks and the multi-features evaluation.

In view of the maturity of the application, the operating speed and the degree of application, in this paper, we propose a new metric BLEUS-syn based on Cilin [7] and smoothed BLEU [8]. We adopt the synonym match except the exact word match in BLEU smoothing technology to improve the evaluation metric performance.

2 BLEUS-syn Metric

2.1 Smoothing BLEU

Papineni [3] has proposed the first evaluation metric BLEU based on n-gram in 2002. Then it is widely used in various evaluations. BLEU is calculated through matching the n-grams between candidate and reference. We take the geometric mean of the test sentences' modified precision scores and then multiply the result by a brevity penalty factor (BP). BLEU is defined as:

$$\mathrm{BLUE} = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

And the N is the maximum base element of n-gram, p_n is the precision of n-gram, w_n is the weight of n-gram. Generally, the N is set 4 and w_n is 1/N. The brevity penalty BP is defined as Formula (2), which is used to compensate the lack of recall.

$$\mathrm{BP} = e^{\min(1-r/c,0)} \tag{2}$$

However, the original BLEU is designed for the corpus-level. When any n-gram precision is zero, the final geometric mean will be zero. So BLEU is short of meaningful sentence-level score, which is important for distinguishing system performance.

In order to compute BLEU at sentence level, we apply smoothing technique to deal with the zero precision.

Lin [4] has proposed smoothing BLEU for the first time. Add one count to the n-gram hit and total n-gram count for n > 1, which is shown as Formula (3). Therefore, for candidates with less than n words, we can still get a positive smoothed BLEU score from shorter n-gram matches. If nothing matches BLEU will be zero.

$$p_n = \frac{Count_{clip(n-gram)^{+1}}}{Count_{(n-gram)^{+1}}} \tag{3}$$

$Count_{clip(n\text{-}gram)}$ is the minimum n-gram number in candidate translation, and $Count_{(n\text{-}gram)}$ is that in reference translation. And the BLEUS is calculated as Formula (4).

$$BLEUS = \min\left(e^{(1-r/c)}, 1\right) \times \exp \sum_{n=1}^{N} w_n \log p_n \tag{4}$$

2.2 Word Similarity Computation Based on Cilin

Diversification of language expression increases the difficulty of information processing. Different systems will produce different translations for the same source language. Semantic analysis and synonym match are important for MT evaluation. Word similarity is the base of research on metric evaluation, and it is important for improving metric performance.

The semantic dictionary, such as WordNet, HowNet and Cilin, leads the word similarity computation to be a hot spot.

2.2.1 Cilin Introduction

Cilin is a semantic dictionary compiled by Mei Jiaju and other scholars and published in 1980s. Then Information Retrieval Laboratory in Harbin Institute of Technology completes HIT IR-Lab Tongyici Cilin (Extended) through deleting not widely used words and dictionary expansion, which contains 77,343 words finally [7].

Cilin contains not only synonym words, but relevant words. Only the leaf nodes of its tree hierarchy are sets of words. Concept is the smallest unit of semantic description, and the tree hierarchy is shown as Fig. 1.

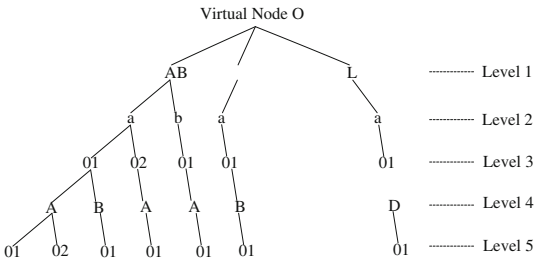


Fig. 1. Cilin tree hierarchy

Five level coding is used in Cilin. The uppercase letters of the alphabet are adopted to represent the major category; medium category with a lowercase letter; minor category is represented by two decimal integers; the fourth level is called word group adopting uppercase letters of alphabet; the fifth grade with two decimal integer is named atomic word group, in each category, there are not many words, a lot of which have only one word that cannot be subdivided any more. With the increasing of the level, the semantic description is more and more detailed.

2.2.2 Cilin Coding Improvement

The paper adopts six-level coding system to facilitate the calculation. Two digits encoding is used for each level, and the English letters are encoded in sequence, such as "A" or "a" substituted by "01", "B" or "b" by "02", and postponed in order. The last two bits we call "mark-bit", which is the sixth level in coding system: "=" is substituted by "01", "#" by "02" and "@" by "03". The new coding system is shown in Table 1. For example, "Da15B02#" becomes "040115020202" in new coding system.

Table 1. New coding system of Cilin

Level	1	2	3	4	5	6
Symbol example	D	a	15	B	02	#
Symbol property	Major category	Medium category	Minor category	Word group	Atomic word group	Mark-bit
Coding	04	01	15	02	02	02

2.3 BLEUS Based on Cilin

As Cilin focusing on adequacy, the paper proposed a new smoothed BLEU metric, BLEUS-syn, based on Cilin. This metric mainly introduces the synonym match into BLEUS. It smoothed the precision with n = 1 as Formula (3) to maintain the consistency of the n-gram precisions with different n. Also this smoothing technique can avoid the zero matching in candidate translations because of the short translation and abbreviated form. The pseudo code of the BLEUS-syn algorithm is shown as Fig. 2, which contains two main functions: ***isSynonym*** and ***bleuscalculate***.

When two words *content1* and *content2* arrive, the ***isSynonym*** function will be triggered: (1) It returns index = 1 if *content1* and *content2* are not in Cilin but have the same form; (2) It extracts the twelve-bit-code sets *code1* and *code2* of *content1* and *content2* from "Cilin.xls". It returns index = 1 if the two code sets have the same code.

When the ***bleuscalculate*** function is triggered, it will smooth unigram firstly in ***Ngramprecision*** function: (1) Put the candidate after word segmentation *candi* into seg2, and after de-duplication it is put into array *arr[]*; (2) When calculating the minimum number of unigram in candidate *count*, it will replace the unigram *content2* in *seg2* with the *content* in *arr[]* if the two variables get 1 on ***isSynonym*** function; And

```
1.    // BLEUS based on Cilin (BLEUS-syn)
2.    String: content1; //word in reference
3.    String: content2; //word in candidate
4.
5.    Function Integer isSynonym(content1, content2)
6.  Integer index=0;
7.  if(content1.equals(content 2)) index=1;
8.  List<String>list1 = getCodesByContent(content 1);
9.  List<String>list2 = getCodesByContent(content 2);
10. for (String code1 : list1)
11.   for (String code2 : list2)
12.       If(code1==code2) index=1;break;
13.   return index;
14.
15.   Function List<String> getCodesByContent(String content)
16.   Map<String, String>result=XLSLoad.getDataFromFile("/Cilin.xls");
17.   for (String key : result.keySet())
18.     String val = result.get(key);
19. String[] valItem = val.split(" ");
20. for (String : valItem)
21.   if (string.equals(content)) codes.add(key);break;
22.   return codes;
23.
24.   Function Double: Ngramprecision(ref, candi, N); // N-gram precision of "count 1" smoothing
25.   Function Float: bleuscalculate(ref, candi)
26.   Integer: lr←ref.length(); // length of reference
27.   Integer: lc←candi.length(); // length of candidate
28.
29.   Double:BLEUS4←min(0, (1-lr/lc)) +1/4* (log(Ngramprecision(ref,candi,1)) +
      log(Ngramprecision(ref,candi,2)) + log(Ngramprecision(ref,candi,3))
      + log(Ngramprecision(ref,candi,4)))
30.   Return exp(BLEUS4);
```

Fig. 2. BLEUS-syn algorithm.

record the *count* at the same time; (3) The same way, when counting *max_ref_count* it will replace the unigram *content1* in *seg1* with the *content* in *arr[]* if the two variables get 1 on ***isSynonym*** function; Then put the *seg1* and *seg2* after replacement into *seg11* and *seg22*; (4) Then we get the minimum of *count* and *max_ref_count*, which is called *count_clip*, and calculate the sum of *count_clip* and unigram precision. For N = 2, 3, 4, we obtain the corresponding smoothed precision with *seg11* and *seg22*, and then the final BLEUS value.

BLEUS-syn algorithm has smoothed the traditional BLEU to make the evaluation at sentence level possible. It also has smoothed unigram precision to deal with the zero matching of unigram as a result of short translations and English abbreviations. And the synonym match based on Cilin decreases the precision reduction caused by the diversity of language expression.

3 Metric Performance Analysis

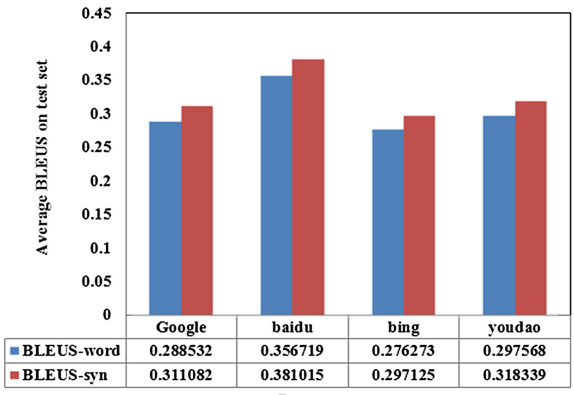
3.1 Corpus and Environment

In the experiment, we use a publicly available benchmark dataset [9], which contains total 8,848 sentence pairs with Russian-Chinese bilingual alignment from 5 websites in news domain. These sentences are different from each other in the form. They are ranked according to Russian sentence length. We proceed the Russian Chinese online translation on Google[1], Baidu[2], Bing[3] and Youdao[4]and get 4 candidate translations from the 4 online translation systems. The Chinese sentences in corpus are considered as reference translations.

We run the experiment on the computer with 8.00 GB memory and Intel(R) Core (TM) i7-6700HQ CPU. Firstly, we implement the traditional BLEUS algorithm to take the BLEUS of the 4 online translation systems. Secondly, we adopt the synonym match to smooth unigram precision; And then we take the other n-gram matches with the segmentation translations after synonym substitution. Finally, we get the BLEU-syn value.

3.2 Result and Discussion

Firstly, we implement the traditional BLEUS algorithm with the test dataset to take the BLEUS through comparing the similarity of reference and 4 candidates. The 4 systems' average BLEUS (BLEUS-word) on the whole test set is shown as Fig. 3.



	Google	baidu	bing	youdao
BLEUS-word	0.288532	0.356719	0.276273	0.297568
BLEUS-syn	0.311082	0.381015	0.297125	0.318339

Fig. 3. BLEUS evaluation results of 4 systems.

[1] http://translate.google.cn/.
[2] http://fanyi.baidu.com/.
[3] https://www.bing.com/translator/.
[4] http://fanyi.youdao.com/.

We take the average of the 200 experimental results as the final result of the group in order to display the results with the chart show. Then we obtain the 45 groups of BLEUS change curve shown as Fig. 4.

Secondly, we implement the BLEUS-syn in the same way with the same corpus to evaluate the translation quality of the 4 systems. We adopt the same experimental process as the BLEUS. Then, we get the results of average BLEUS-syn as shown in Fig. 3 and obtain the 45 groups of BLEUS-syn change curve shown as Fig. 5.

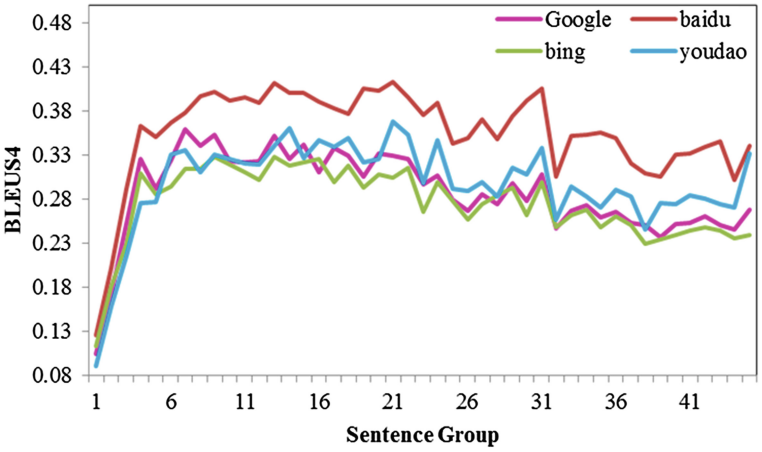


Fig. 4. BLEUS change curves of 4 systems.

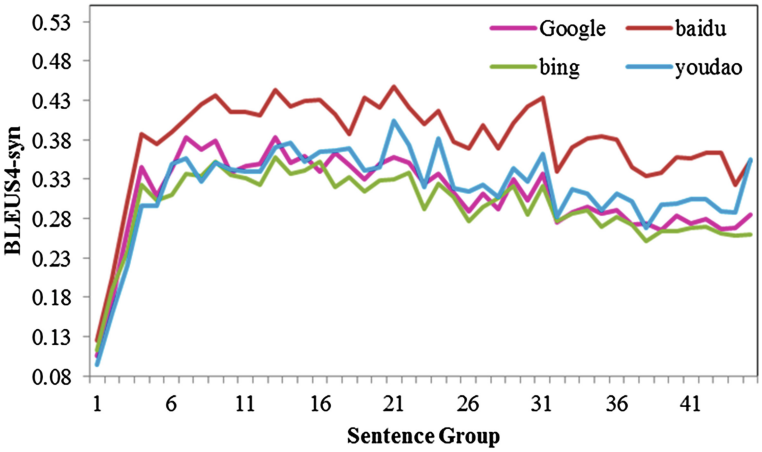


Fig. 5. BLEUS-syn change curves of 4 systems.

Finally, we compare the experimental results of the BLEUS-syn and the traditional BLEUS. Figures 4 and 5 present the average BLEUS and BLEUS-syn results. The horizontal coordinate is the sentence group sequence and the vertical coordinates are the traditional BLEUS4 and BLEUS4-syn scores respectively. The two figures show that the evaluation results of BLEUS and BLEUS-syn are consistent: (1) The two algorithms' average BLEUS4 have the same trend in the whole range; (2) Baidu system performance is the best and its BLEUS4 score is the highest; Youdao performance is slightly worse than Baidu, but better than that of Google; the translation quality of Bing is the worst and its BLEUS4 score is the lowest; (3) The score of BLEUS4 was lower in the first few groups, and then this value increases dramatically with the increase of sentence length. The main reason is that there may be English abbreviations or short translation when the sentence length is short, so that the candidate match with the reference worse, the n-gram matched will be less and the BLEUS4 is reduced. But the value of BLEUS4 tends to be stable with the increase of the sentence length.

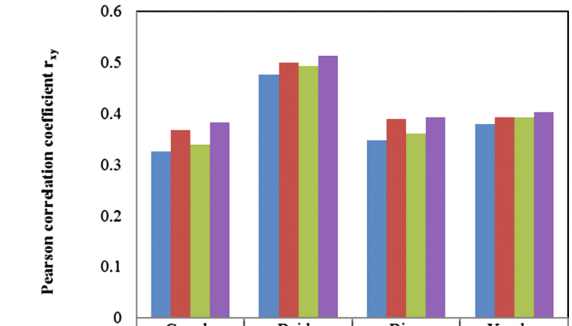
Figure 3 shows that, in the whole test set, (1) The evaluation value of BLEUS-syn is higher than that of the traditional BLEUS for the 4 online systems; (2) The 4 systems on Russian Chinese translation is consistent that Baidu performance is the best, Youdao the second, Google the third and Bing is the worst; (3) When using the synonym match based on Cilin, the BLEUS value of Baidu increases from 0.356719 to 0.381015, with an increase of 6.81%; the BLEUS value of Youdao increases from 0.297568 to 0.318339, with an increase of 6.98%; the Google BLEUS is increased by 7.82% from 0.288532 to 0.311082; the one of Bing is increased by 7.55% from 0.276273 to 0.297125; (4) we use longitudinal comparison to compare the performance of different smoothing algorithms based on the same evaluation metric-BLEU; and we only apply the average BLEU value for evaluation, which is convenient and clear, and conducive to the evaluation metric performance parameters' adjustment and optimization; This will greatly save resources and time, and improve the efficiency; We will apply horizontal comparison for different types of metrics; (5) Google performance improves the most and Baidu the least; this result will play a very good role in the system integration.

We adopt longitudinal comparison to evaluate the performance of BLEUS and BLEUS-syn in the above experiment. The quality of these evaluation metrics is usually measured by determining the correlation of the scores assigned by the evaluation metrics to scores assigned by a human evaluation metric, most commonly fluency and adequacy. In this paper, Pearson correlation coefficient r_{xy} is used to evaluate the two measurements. The higher the coefficient, the better the performance of the evaluation metric.

Suppose the data point on test set, which includes variable automatic scoring x and manual scoring y, is set to $\{(x_i,y_i)\}$. Then the Pearson correlation coefficient r_{xy} is defined as follows:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{5}$$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 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 correlation coefficient r_{xy} does not depend on sample size [10]. Its value ranges from –1 to 1. The positive correlation



	Google	Baidu	Bing	Youdao
ade-BLEUS	0.325697286	0.475909753	0.346638057	0.378885757
flu-BLEUS	0.366523711	0.499488521	0.38878689	0.391633477
ade-BLEUS-syn	0.338901677	0.492122503	0.360357465	0.39269855
flu-BLEUS-syn	0.383148651	0.513486896	0.391987261	0.402092694

Fig. 6. Pearson correlation coefficient r_{xy} between the automatic score and human scores of adequacy and fluency.

coefficient indicates that variable X and variable Y tend to increase or decrease at the same time. On the contrary, the negative correlation coefficient indicates that the variable Y will decrease with the increase of the variable X or increase with the decrease of X.

Take the 4 Russian-Chinese machine translation systems with BLEUS and BLEUS-syn to get the Pearson correlation coefficient r_{xy} between the automatic evaluation score and human scores of adequacy and fluency, which is shown in Fig. 6.

As can be seen in Fig. 6, the Pearson correlation coefficient of adequacy and fluency are all improved as using the Cilin-based SST algorithm for the traditional BLEUS method. For example, the adequacy correlation coefficient of Baidu increases from 0.475910 to 0.492123, Google from 0.325697 to 0.338902, Youdao from 0.378886 to 0.392699, and that of Bing from 0.346638 to 0.360357. And the fluency correlation coefficient is the same. Use of synonym matching does not influence the fluency of candidate translation and simultaneously improve the adequacy of the translation. And the translation is still readable.

The results of the longitudinal comparisons above are consistent with the results of the human evaluation. The above experiment results show that the longitudinal comparison method to evaluate the metrics with different parameter settings based on the same method is effective. This method is conducive to the adjustment and optimization of evaluation metrics, and is more convenient. Also it can significantly save energy and time and improve the timeliness.

In this paper, we use the significance test to verify the effectivity of experimental results. Also, this method can be applied to the offline open source system. It can greatly enhance the MT system performance in the case of the corpus size is not limited. The BLEUS-syn algorithm can greatly improve the performance of traditional BLEUS algorithm. And it plays a very good role in the MT evaluation with Chinese as the target language.

In this paper, we use the significance test to verify the effectivity of experimental results. Also, this method can be applied to the offline open source system. It can greatly enhance the MT system performance in the case of the corpus size is not limited. The BLEUS-syn algorithm can greatly improve the performance of traditional BLEUS algorithm. And it plays a very good role in the MT evaluation with Chinese as the target language.

4 Conclusion

This paper proposes an improved smoothed BLEU evaluation metric (BLEUS-syn). This metric has smoothed n-gram of the traditional BLEUS in the light of zero matching caused by English abbreviations or short translation, and introduced synonym match in unigram matching, and then calculated the other n-gram precisions. The results of the new algorithm and the traditional BLEUS algorithm are consistent from the longitudinal comparison. It will greatly enhance the performance of traditional BLEUS algorithm, especially in machine translation with Chinese as target language.

Further research will concern that the relevant word in Cilin, HowNet and other metrics with synonym match such as ROUGE, METEOR. Also we will evaluate these different types of metrics with horizontal comparison, for example, ORANGE and traditional human evaluation.

Acknowledgments. The research is supported by the Key Project of State Language Commission of China (Resource Construction and Application of Low-Resource Languages for the 21st Century Maritime Silk Road) and the Featured Innovation Project of Guangdong Province (No. 2015KTSCX035).

References

1. Assessment Specifications of Machine Translation Systems. GF 2006
2. Yu, S.: Automatic evaluation of output quality for Machine Translation systems. Mach. Transl. **8**, 117–126 (1993)
3. Papineni, K., Roukos, S., Ward, T., et al.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics 2002, pp. 311–318 (2002)
4. Lin, C.Y., Och, F.J.: ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In: Proceedings of COLING-2004 (2004)
5. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Proceedings of Workshop on Text Summarization Branches Out, Post-conference Workshop of ACL 2004 (2004)
6. Banerjee, S., Lavie, A.: Meteor: an automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (2005)
7. Mei, J., Zhu, Y., Gao, Y., et al.: Tongyici Cilin (Extended). HIT IR-Lab (1996)

8. Chen, B., Cherry, C.: A systematic comparison of smoothing techniques for sentence-level BLEU. In: Proceedings of the Ninth Workshop on Statistical Machine Translation 2014, pp. 362–367 (2014)
9. Du, W., Liu, W., Yu, J., et al.: Russian-Chinese sentence-level aligned news corpus. In: EAMT 2015 (2015)
10. Koehn, P.: Moses-statistical machine translation system- user manual and code guide (2015)