Application of Prize based on Sentence Length in Chunk-based Automatic Evaluation of Machine Translation

Hiroshi Echizen'ya
Hokkai-Gakuen University
S26-Jo, W11-Chome, Chuo-ku,
Sapporo 064-0926 Japan
echi@lst.hokkai-s-u.ac.jp

Kenji Araki
Hokkaido University
N 14-Jo, W 9-Chome, Kita-ku,
Sapporo 060-0814 Japan
araki@ist.hokudai.ac.jp

Eduard Hovy
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
hovy@cmu.edu

Abstract

As described in this paper, we propose a new automatic evaluation metric for machine translation. Our metric is based on chunking between the reference and candidate translation. Moreover, we apply a prize based on sentence-length to the metric, dissimilar from penalties in BLEU or NIST. We designate this metric as **A**utomatic Evaluation of Machine Translation in which the **P**rize is **A**pplied to a **C**hunk-based metric (APAC). Through meta-evaluation experiments and comparison with several metrics, we confirmed that our metric shows stable correlation with human judgment.

1 Introduction

In the field of machine translation, various automatic evaluation metrics have been proposed. Among them, chunk-based metrics such as METEOR(A. Lavie and A. Agarwal, 2007), ROUGE-L(Lin and Och, 2004), and IMPACT(H. Echizen-ya and K. Araki, 2007) are effective. In general, BLEU(K. Papineni et al., 2002), NIST(NIST, 2002), and RIBES(H. Isozaki et al., 2010) use a penalty for calculation of scores because the high score is often given extremely when the candidate translation is short. Therefore, the penalty is effective to obtain high correlation with human judgment. On the other hand, almost all chunk-based metrics use the F-measure based on a precision by candidate translation and a recall by reference. Moreover, they assign a penalty for the difference of chunk order between the candidate translation and the reference, not the penalty for the difference of sentence length. Nevertheless, it is also important for chunk-based metrics to examine the sentence length. In chunk-based metrics, each word's weight depends on the sentence length. For example, the weight of each word is 0.2 (=1/5) when the number of words in a sentence is 5; it is 0.1 (=1/10) when the number of words in a sentence is 10. Therefore, the weight of the non-matched word in the short sentence is large.

To resolve this problem, it is effective for short sentences to give a prize based on the sentence length in the chunk-based metrics. Therefore, we propose a new metric using a prize based on the sentence length. We designate this metric as **A**utomatic Evaluation of Machine Translation in which the **P**rize is **A**pplied to a **C**hunk-based metric (APAC). In our metric, the weight of a non-matched word becomes small for the short sentence by awarding of the prize. It is almost identical to that for a long sentence by awarding of the prize. Therefore, our metric does not depend heavily on sentence length because the weight of non-matched words is constantly small. We confirmed the effectiveness of APAC using meta-evaluation experiments.

2 Score calculation in APAC

The APAC score is calculated in two phases. In the first phase, the chunk sequence is determined between a candidate translation and the reference. The chunk sequence

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 381–386,
Baltimore, Maryland USA, June 26–27, 2014.

is determined using the **L**ongest **C**ommon **S**ubsequence (LCS). Generally, several chunk sequences are obtained using LCS. In that case, APAC determines only one chunk sequence using the number of words in each chunk and the position of each chunk.

For example, in between the candidate translation "In this case, the system power supply is accessory battery 86." and "In this case, the system power supply is the accessory power supply battery 86.", the chunk sequence is "in this case, the system power supply is", "accessory" and "battery 86.", and the chunk sequence is ony one in these sentences. Only one chunk sequence is determined using the number of words in each chunk and the position of each chunk when several chunk sequences are obtained.

The second phase is calculation of the score based on the determined chunk sequence. The Ch_score in Eq. (3) is calculated using the determined chunk sequence. In Eq. (3), ch denotes each chunk and ch_num represents the number of chunks. Moreover, $length(ch)$ is the word number of each chunk. β is the weight parameter for the length of each chunk. For example, in between the candidate translation "In this case, the system power supply is accessory battery 86." and "In this case, the system power supply is the accessory power supply battery 86.", ch_num is 3 ("in this case, the system power supply is", "accessory" and "battery 86."). Therefore, Ch_score is 91 ($=9^{2.0}+1^{2.0}+3^{2.0}$) when β is 2.0.

$$P = \left\{ \left(\frac{\sum_{i=0}^{RN-1} (\alpha^i \times Ch\_score)}{m^\beta} \right)^{\frac{1}{\beta}} + 0.5 \times Prize\_m \right\} / 2.0 \quad (1)$$

$$R = \left\{ \left(\frac{\sum_{i=0}^{RN-1} (\alpha^i \times Ch\_score)}{n^\beta} \right)^{\frac{1}{\beta}} + 0.5 \times Prize\_n \right\} / 2.0 \quad (2)$$

$$Ch\_score = \sum_{ch \in ch\_num} length(ch)^\beta \quad (3)$$

$$Prize\_m = \frac{1}{log(m)+1} \quad (4)$$

$$Prize\_n = \frac{1}{log(n)+1} \quad (5)$$

$$APAC\ \ score = \frac{(1+\gamma^2)RP}{R+\gamma^2 P} \quad (6)$$

The P and R in Eqs. (1) and (2) respectively denote precision by candidate translation and recall by reference. These are calculated using the Ch_score obtained using Eq. (3). Therein, m and n respectively represent the word numbers of the candidate translation and the reference. Moreover, the chunk sequence determination process is repeated recursively to all common words. The number of determination processes of the chunk sequence is high when the word order of the candidate translation differs from that of the reference. The RN is the number of determination processes of the chunk sequence. Here, α is the parameter for the chunk order. It is less than 1.0. The value of the Ch_score is small when the chunk order between the candidate translation and references differs because the value of $length(ch)$ in each chunk becomes small. For example, in between the candidate translation "In this case, the system power supply is accessory battery 86." and "In this case, the system power supply is the accessory power supply battery 86.", $\left(\frac{\sum_{i=0}^{RN-1} (\alpha^i \times Ch\_score)}{m^\beta} \right)^{\frac{1}{\beta}}$ is 0.773 ($=\sqrt{\frac{91}{169}}=\sqrt{\frac{\sum_{i=0}^{1-1}(0.1^0 \times 91)}{13^{2.0}}}$) and $\left(\frac{\sum_{i=0}^{RN-1} (\alpha^i \times Ch\_score)}{n^\beta} \right)^{\frac{1}{\beta}}$ is 0.596 ($=\sqrt{\frac{91}{256}}=\sqrt{\frac{\sum_{i=0}^{1-1}(0.1^0 \times 91)}{16^{2.0}}}$) when α and β respectively stand for 0.1 and 2.0. The value of RN is 1 because there is no more matching words after the determined chunks ("in this case, the system power supply is", "accessory" and "battery 86.") are removed from the candidate translation "In this case, the system power supply is accessory battery 86." and "In this case, the system power supply is the accessory power supply battery 86.".

Moreover, $Prize\_m$ and $Prize\_n$ in Eqs. (1) and (2) are calculated respectively using Eqs.

(4) and (5). Each is less than 1.0. For example, in the candidate translation "In this case, the system power supply is accessory battery 86." and "In this case, the system power supply is the accessory power supply battery 86.", $Prize\_m$ and $Prize\_n$ respectively stand for 0.473 ($=\frac{1}{1.114+1}=\frac{1}{log(13)+1}$) and 0.454 ($=\frac{1}{1.204+1}=\frac{1}{log(16)+1}$). These values become large in the short sentences. They become small in the long sentences. Therefore, the weight of each non-matched word is small in the short sentences. It is kept small in the long sentences. Finally, the score is calculated using Eq. (6). This equation shows the f-measure based on P and R. In Eq. (6), γ is determined as P/R(C. J. V. Rijsbergen, 1979). The $APAC\ score$ is between 0.0 and 1.0. For example, in the candidate translation "In this case, the system power supply is accessory battery 86." and "In this case, the system power supply is the accessory power supply battery 86.", P and R respectively stand for 0.505 ($=\frac{0.773+0.5\times0.473}{2.0}$) and 0.412 ($=\frac{0.596+0.5\times0.454}{2.0}$). Therefore, $APAC\ score$ is 0.445 ($=\frac{0.521}{1.171}=\frac{(1+1.503)\times0.412\times0.505}{0.412+1.503\times0.505}$) and γ is 1.226 ($=\frac{0.505}{0.412}$)

3 Experiments

3.1 Experimental Procedure

Meta-evaluation experiments are performed using WMT2012(C. Callison-Burch et al., 2012) data and WMT2013(O. Bojar et al., 2013) data, and NTCIR-7(A. Fujii et al., 2008) data and NTCIR-9(A. Goto et al., 2011) data. All sentences by NTCIR data are English patent sentences obtained through Japanese-to-English translation. The number of references is 1. In NTICR-7 data, the average value in the evaluation results of three human judgments is used as the scores of 1–5 from the perspective of adequacy and fluency. In NTCIR-9 data, the evaluation results of one human judgment is used as the scores of 1–5 from the view of adequacy and acceptance. For this meta-evaluation, we used only English and Japanese candidate translations because we can evaluate them in comparison with other languages correctly.

We calculated the correlation between the scores by automatic evaluation and the scores by human judgments at the system level and the segment level, respectively. Spearman's rank correlation coefficient is used at the system level. The Kendall tau rank correlation coefficient is used in the segment level.

Moreover, we used BLEU (ver. 13a), NIST (ver. 13a), METEOR (ver. 1.4), and APAC with no prize (APAC_no_p) as the automatic evaluation metrics for comparison with APAC as shown in Eqs. (4) and (5). In APAC_no_p, $\left(\frac{\sum_{i=0}^{RN-1}(\alpha^i\times Ch\_score)}{m^\beta}\right)^{\frac{1}{\beta}}$ as P and $\left(\frac{\sum_{i=0}^{RN-1}(\alpha^i\times Ch\_score)}{m^\beta}\right)^{\frac{1}{\beta}}$ as R are used respectively in Eqs. (1) and (2).

3.2 Experimental Results

Tables 1 and 2 respectively present Spearman's rank correlation coefficients of system-level and Kendall tau rank correlation coefficients of segment-level in WMT2012 data. Tables 3 and 4 respectively show Spearman's rank correlation coefficients of the system-level and Kendall tau rank correlation coefficients of segment-level in WMT2013 data. Moreover, Tables 5 and 6 respectively present Spearman's rank correlation coefficients of system-level and Kendall tau rank correlation coefficients of segment-level in NTCIR-7 data. Tables 7 and 8 respectively show Spearman's rank correlation coefficients of system-level and Kendall tau rank correlation coefficients of the segment level in NTCIR-9 data.

In APAC, 0.1 and 1.2 were used as the values of parameters α and β by the preliminarily experimentally obtained results. In Tables 1–8, "Rank" denotes the ranking based on "Avg." The value of "()" denotes the number of MT systems in Tables 1, 3, 5, and 7. The value of "()" represents the number of sentence pairs in Tables 2, 4, 6, and 8. These values depend on the data.

3.3 Discussion

The results presented in Tables 1–8 indicate that APAC can obtain the most stable correlation coefficients among some metrics. The ranking of APAC is No. 1 through NTCIR data in Tables 5–8. In WMT data of Tables 1–4, the ranking of APAC is the lowest except for Table 3. However, the difference

	cs-en(6)	de-en(16)	es-en(12)	fr-en(15)	Avg.	Rank
APAC	0.886	0.650	0.958	0.811	0.826	5
APAC_no_p	0.886	0.676	0.958	0.807	0.832	3
METEOR	0.943	0.841	0.979	0.818	0.895	1
BLEU	0.886	0.674	0.958	0.796	0.828	4
NIST	0.943	0.700	0.944	0.779	0.841	2

Table 1: Spearman's rank correlation coefficient of system-level in WMT2012 data.

	cs-en(11,155)	de-en(12,042)	es-en(9,880)	fr-en(11,682)	Avg.	Rank
APAC	0.185	0.204	0.209	0.226	0.206	3
APAC_no_p	0.189	0.207	0.208	0.226	0.207	2
METEOR	0.223	0.279	0.248	0.243	0.248	1

Table 2: Kendall tau rank correlation coefficient of the segment level in WMT2012 data.

between the ranking of METEOR, which is the highest, and that of APAC is not larger in WMT data. The correlation coefficients of APAC in NTCIR data of Tables 5–8 are higher than those of METEOR. In Tables 5 and 6, underlining in APAC signifies that the differences between correlation coefficients obtained using APAC and METEOR are statistically significant at the 5% significance level. In Table 7, the correlation coefficients of METEOR, BLEU, and NIST are extremely low. Only one human judgment was used in NTCIR-9 data. As a result, APAC is fundamentally effective for various languages independent of the differences in the grammatical structures between languages: these experimentally obtained results indicate that APAC is the most stable metric.

Moreover, in APAC, the correlation coefficients of the segment level in NTCIR data were increased using the prize of Eqs. (4) and (5). In WMT data, the correlation coefficients are almost identical using the prize. Therefore, use of the prize was fundamentally effective at the segment level. The evaluation quality of segment level is generally very low in the automatic evaluation metrics. Therefore, it is extremely important to improve the correlation coefficient of segment level. Application of the prize is effective to improve the evaluation quality of the segment level.

4 Conclusion

As described in this paper, we proposed a new chunk-based automatic evaluation metric using the prize based on the sentence length. The experimentally obtained results indicate that APAC is the most stable metric.

We will improve APAC to obtain higher correlation coefficients in future studies. Particularly, we will strive to improve the correlation coefficients at the segment level. The APAC software will be released by `http://www.lst.hokkai-s-u.ac.jp/~echi/automatic_evaluation_mt.html`.

Acknowledgments

This work was done as research under the AAMT/JAPIO Special Interest Group on Patent Translation. The Japan Patent Information Organization (JAPIO) and the National Institute of Information (NII) provided corpora used in this work. The author gratefully acknowledges support from JAPIO and NII.

References

O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Sortcut and L. Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. Proceedings of the Eighth Workshop on Statistical Machine Translation. pp.1–44.

C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Sortcut and L. Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. Proceedings of the Seventh Workshop on Statistical Machine Translation. pp.10–51.

H. Echizen-ya and K. Araki. 2007. Automatic Evaluation of Machine Translation based on

	cs-en(11)	de-en(17)	es-en(12)	fr-en(13)	ru-en(19)	Avg.	Rank
APAC	0.900	0.904	0.916	0.934	0.709	0.873	3
APAC_no_p	0.909	0.909	0.937	0.934	0.721	0.882	2
METEOR	0.982	0.946	0.923	0.967	0.889	0.941	1
BLEU	0.945	0.897	0.853	0.951	0.614	0.852	4
NIST	0.900	0.828	0.804	0.786	0.465	0.757	5

Table 3: Spearman's rank correlation coefficient of the system level in WMT2013 data.

Metrics	cs-en (85,469)	de-en (128,668)	es-en (67,832)	fr-en (80,741)	ru-en (151,422)	Avg.	Rank
APAC	0.144	0.163	0.169	0.139	0.121	0.147	3
APAC_no_p	0.148	0.167	0.176	0.142	0.123	0.151	2
METEOR	0.222	0.236	0.241	0.194	0.226	0.224	1

Table 4: Kendall tau rank correlation coefficient of the segment level in WMT2013 data.

Recursive Acquisition of an Intuitive Common Parts Continuum. Proceedings of the Eleventh Machine Translation Summit. pp.151–158.

A. Fujii, M. Utiyama, M. Yamamoto and T. Utsuro. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. Proceedings of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access. pp.389–400.

I. Goto, B. Lu, K. P. Chow, E. Sumita and B. K. Tsou. 2011. Overview of the Patent Translation Task at the NTCIR-9 Workshop. Proceedings of the Ninth NTCIR Workshop Meeting. pp.559–578.

H. Isozaki, T. Hirao, K. Duh, K. Sudoh and H. Tsukada. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp.944–952.

A. Lavie and A. Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. Proceedings of the Second Workshop on Statistical Machine Translation.

Chin-Yew Lin and F. J. Och. 2004. Automatic Evaluation of Machine Translation Quality Using the Longest Common Subsequence and Skip-Bigram Statistics. *In Proc. of ACL'04*, 606–613.

NIST. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf*.

K. Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp.311–318.

C. J. Van Rijsbergen. 1979. *Information Retrieval (2nd ed.)*, Butterworths.

	Adequacy(15)	Fluency(15)	Avg.	Rank
APAC	0.872	0.805	0.839	1
APAC_no_p	0.872	0.805	0.839	1
METEOR	0.424	0.380	0.402	5
BLEU	0.582	0.586	0.584	3
NIST	0.578	0.568	0.573	4

Table 5: Spearman's rank correlation coefficient of the system level in NTCIR-7 data.

	Adequacy (1,500)	Fluency (1,500)	Avg.	Rank
APAC	0.494	0.489	0.491	1
APAC_no_p	0.482	0.476	0.479	2
METEOR	0.366	0.383	0.375	3

Table 6: Kendall tau rank correlation coefficient of the segment level in NTCIR-7 data.

	Adequacy (19)	Acceptance (14)	Avg.	Rank
APAC	0.182	0.298	0.240	1
APAC_no_p	0.182	0.298	0.240	1
METEOR	-0.081	0.015	-0.033	4
BLEU	-0.123	0.059	-0.032	3
NIST	-0.344	-0.275	-0.309	5

Table 7: Spearman's rank correlation coefficient of the system level in NTCIR-9 data.

	Adequacy (5,700)	Acceptance (5,700)	Avg.	Rank
APAC	0.250	0.261	0.256	1
APAC_no_p	0.242	0.250	0.246	2
METEOR	0.167	0.217	0.192	3

Table 8: Kendall tau rank correlation coefficient of segment-level in NTCIR-9 data.