Automatic Evaluation Metric for Machine Translation that is Independent of Sentence Length

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@media.eng.hokudai.ac.jp

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

Abstract

We propose new automatic evaluation metric to evaluate machine translation. Different from most similar metrics, our proposed metric does not depend heavily on sentence length. In most metrics based on f-measure comparisons of reference and candidate translations, the relative weight of each mismatched word in short sentences is larger than it in long sentences. Therefore, the evaluation score becomes disproportionally low in short sentences even when only one non-matching word exists. In our metric, the weight of each mismatched word is kept small even in short sentences. We designate our metric as **A**utomatic Evaluation Metric that is **I**ndependent of Sentence **Le**ngth (AILE). Experimental results indicate that AILE has the highest correlation with human judgments among some leading metrics.

1 Introduction

Various automatic evaluation metrics for machine translation have been proposed through the metrics task on the Workshop on statistical Machine Translation (WMT). One can identify three kinds of automatic evaluation metrics (C. Liu et al., 2010): the heavyweight linguistic approach, which corresponds to RTE (S. Padó et al., 2009) and ULC (J. Giménez and L. Márquez, 2007); the lightweight linguistic approach, which corresponds to METEOR (A. Lavie and A. Agarwal, 2007) and MaxSim (Y. Seng Chan and H. Tou Ng, 2008) and the non-linguistic approach, which includes BLEU (K. Papineni et al., 2002), TER (M. Snover et al., 2006), RIBES (H. Isozaki et al., 2010) and IMPACT (H. Echizen-ya and K. Araki, 2007)(H. Echizen-ya et al., 2012). In this paper, we specifically examine a metric that corresponds to the lightweight linguistic and non-linguistic approaches because they are useful and are very easily built.

Among these metrics, METEOR and IMPACT are based on the f-measure, which combines precision and recall between the reference and candidate texts. The metrics' simple f-measure (P. Koehn, 2010) obtains precision and recall using Eqs. (1)–(3):

$$precision = \frac{matching\ \ words}{length\ \ of\ \ candidate} \quad (1)$$

$$recall = \frac{matching\ \ words}{length\ \ of\ \ reference} \quad (2)$$

Then f-measure is calculated using Eq. (3):

$$f\text{-}measure = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

For example, in the reference "doctor cured a patient" and candidate "doctor treated a patient", the precision and the recall are respectively 0.75 ($=\frac{3}{4}$). Therefore, the f-measure is 0.75 ($=\frac{2\times0.75\times0.75}{0.75+0.75}$), even though there is only one non-matching word. This is because the denominator is so small, since the sentences

Proceedings of Recent Advances in Natural Language Processing, pages 230–236,
Hissar, Bulgaria, 7-13 September 2013.

are short: the weight of each non-matched word is 0.25 ($=\frac{1}{4}$) in this example. In general, the relative influence of each non-matching word increases when sentences are short, distorting the overall score. This problem is especially serious in short sentences. On the other hand, the weight of each mismatched word is small when the number of words is large. For example, the weight of each word is 0.05 ($=\frac{1}{20}$) when the sentence length is 20. Therefore, an automatic evaluation metric in which the weight of each mismatched word does not depend heavily on sentence length would be highly desirable.

In this paper, we propose a new automatic evaluation metric in which the weight of each mismatched word does not depend heavily on sentence length. In our metric, the weight of each mismatched word is kept small even in short sentences. Therefore, our metric can obtain a stable evaluation score without regard to sentence length. We designate the metric as **A**utomatic Evaluation Metric that is **I**ndependent of Sentence **Le**ngth (AILE). Through experimentally obtained results, we confirmed that AILE indicates the highest correlation with human judgment among several leading metrics.

2 AILE: Automatic Evaluation Metric Independent of Sentence Length

In AILE, a chunk sequence is decided using **L**ongest **C**ommon **S**ubsequence (LCS) between the reference and candidate. A chunk is a string of consecutive words. In "doctor cured a patient" and "doctor treated a patient", the value of LCS is 3 because the matching words are "doctor", "a" and "patient". Therefore, the chunks are "doctor" and "a patient".

Moreover, AILE obtains $AILEscore$ as the evaluation score using the following Eqs. (4)–(8).

$$P = \left(\frac{\sum_{i=0}^{RN-1} (\alpha^i \times C\_score) + weight}{m^\beta + weight} \right)^{\frac{1}{\beta}} \quad (4)$$

$$R = \left(\frac{\sum_{i=0}^{RN-1} (\alpha^i \times C\_score) + weight}{n^\beta + weight} \right)^{\frac{1}{\beta}} \quad (5)$$

$$C\_score = \sum_{c \in c\_num} length(c)^\beta \quad (6)$$

$$weight = \begin{cases} \left(\frac{\delta}{log(m+n)}\right)^\beta, & C\_score > 0.0 \\ 0.0, & C\_score = 0.0 \end{cases} \quad (7)$$

$$AILE\ \ score = \frac{(1+\gamma^2)RP}{R+\gamma^2 P} \quad (8)$$

In Eq. (6), c and c_num mean each chunk and the number of chunks, respectively. Moreover, $length(c)$ means the number of words in each chunk and β is a parameter for the weight of chunk length. In "doctor cured a patient" and "doctor treated a patient", two chunks (*i.e.*, "doctor" and "a patient") exist. Therefore, C_score is 5.0 ($=1^{2.0}+2^{2.0}$) when β is 2.0. The $weight$ of Eq. (7) controls the weight of each matching word according to the sentence length. The m and n mean respectively the candidate length and reference length. The δ and β are parameters. The value of $weight$ is 0.0 when C_score is 0.0 because it means that the matching words between the reference and candidate do not exist. In "doctor cured a patient" and "doctor treated a patient", the value of $weight$ in Eq. (7) is 1.2261 ($=(\frac{1.0}{log(4+4)})^{2.0}$) when δ and β are respectively 1.0 and 2.0.

In Eqs. (4) and (5), P and R respectively indicate precision and recall. Moreover, RN means the repetition number for the decision of C_score. For example, in "doctor cured a patient" and "A patient helped doctor", the appearance order of chunks (*i.e.*, "doctor" and "a patient") between two sentences is different. In this case, $RN-1$ is 1 because $\alpha^0 \times C\_score$ for the chunk "a patient" is firstly calculated and $\alpha^1 \times C\_score$ for the chunk "doctor" is secondly calculated. That is, α is used as the parameter for the penalty when the appearance order of chunks between reference and candidate is different. In "doctor cured a patient" and "doctor treated a patient", the value of $\sum_{i=0}^{RN-1} (\alpha^i \times C\_score)$ is

Table 1: Spearman's rank correlation coefficient of system-level in AILE using NTCIR-7.

Parameters	Adequacy (14 systems)	Fluency (14 systems)	Avg.
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 2.0$	0.9912	0.9253	**0.9583**
$\alpha = 0.3$, $\beta = 1.2$, $\delta = 2.0$	0.9868	0.9297	**0.9583**
$\alpha = 0.5$, $\beta = 1.2$, $\delta = 2.0$	0.9780	0.9253	0.9517
$\alpha = 0.7$, $\beta = 1.2$, $\delta = 2.0$	0.9560	0.9033	0.9297
$\alpha = 0.9$, $\beta = 1.2$, $\delta = 2.0$	0.9473	0.8945	0.9209
$\alpha = 0.1$, $\beta = 1.0$, $\delta = 2.0$	0.9912	0.9253	**0.9583**
$\alpha = 0.1$, $\beta = 1.4$, $\delta = 2.0$	0.9780	0.9165	0.9473
$\alpha = 0.1$, $\beta = 1.6$, $\delta = 2.0$	0.9780	0.9165	0.9473
$\alpha = 0.1$, $\beta = 1.8$, $\delta = 2.0$	0.9780	0.9165	0.9473
$\alpha = 0.1$, $\beta = 2.0$, $\delta = 2.0$	0.9736	0.9121	0.9429
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 1.0$	0.9780	0.9253	0.9517
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 3.0$	0.9868	0.9297	**0.9583**
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 4.0$	0.9768	0.9297	**0.9583**
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 5.0$	0.9834	0.9241	0.9538
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 6.0$	0.9780	0.9165	0.9473
square root	0.9780	0.9253	0.9517
arctangent	0.9912	0.9253	**0.9583**

5.0 ($=0.5^0 \times 5.0$) when α is 0.5 because $RN-1$ is 0. The value of P and R in Eqs. (4) and (5) is respectively 0.6012 ($=\sqrt{\frac{5.0+1.2261}{4^{2.0}+1.2261}}$). Eq. (8) indicates f-measure using P and R. The γ is obtained as P/R. In "doctor cured a patient" and "doctor treated a patient", the value of $AILEscore$ is 0.6012 ($=\frac{(1+1.0^2)\times 0.6012\times 0.6012}{0.6012+1.0^2\times 0.6012}$) because the value of γ is 1.0 ($=\frac{0.6012}{0.6012}$).

The evaluation score increases from 0.5590 to 0.6012 using *weight* in Eq. (7). The $AILEscore$ without *weight* is 0.5590 because the value of P and R is respectively 0.5590 ($=\sqrt{\frac{5.0}{4^{2.0}}}$). This means that AILE can increase the evaluation score in short sentences using *weight* in Eq. (7). The value of *weight* is 1.2261 ($=(\frac{1.0}{log(4+4)})^{2.0}$) when m and n are respectively 4. The value of *weight* is 0.3896 ($=(\frac{1.0}{log(20+20)})^{2.0}$) when m and n are respectively 20. That is, the weight of non-matched words decreases in short sentences adding the large value (*e.g.*, 1.2261) of *weight* to the matching words (*i.e.*, $\sum_{i=0}^{RN-1}(\alpha^i \times C\_score)$ in Eqs. (4) and (5)). On the other hand, the weight of non-matched words does not change in long sentences, adding only the small value (*e.g.*, 0.3869) of *weight* to the matching words (*i.e.*, $\sum_{i=0}^{RN-1}(\alpha^i \times C\_score)$ in Eqs. (4) and (5)). Therefore, AILE can obtain a stable evaluation score without depending on sentence length.

3 Experiments

3.1 Experimental Procedure

We performed experiments to confirm the effectiveness of AILE. The correlations between the scores by automatic evaluation and the scores by human judgments are calculated, respectively, at the system level and the sentence level. Spearman's rank correlation coefficient is used at the system level and the Kendall tau rank correlation coefficient is used in the sentence level. In the first experiment, the references and candidates were obtained from patent data in NTCIR-7 (A. Fujii et al., 2008). We used as candidates the machine translation system's translation of Japanese sentences into English sentences. In NTCIR-7 data, 14 machine translation systems were used and each machine translation system translated 100 Japanese sentences into 100 English sentences. Therefore, we obtained 1,400 candidates. We used single references. The median value in the evaluation results of three human judges was used as the scores of 1–5. The experiments determined suitable values for the three parameters α, β and δ. Moreover, the

Table 2: Kendall tau rank correlation coefficient of sentence-level in AILE using NTCIR-7.

Parameters	Adequacy (1,400 sentences)	Fluency (1,400 sentences)	Avg.	Total Avg.
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 2.0$	0.4304	0.3627	0.3965	**0.6774**
$\alpha = 0.3$, $\beta = 1.2$, $\delta = 2.0$	0.4231	0.3596	0.3914	0.6749
$\alpha = 0.5$, $\beta = 1.2$, $\delta = 2.0$	0.4095	0.3533	0.3814	0.6666
$\alpha = 0.7$, $\beta = 1.2$, $\delta = 2.0$	0.3862	0.3414	0.3638	0.6468
$\alpha = 0.9$, $\beta = 1.2$, $\delta = 2.0$	0.3449	0.3156	0.3303	0.6256
$\alpha = 0.1$, $\beta = 1.0$, $\delta = 2.0$	0.4058	0.3400	0.3729	0.6656
$\alpha = 0.1$, $\beta = 1.4$, $\delta = 2.0$	0.4300	0.3645	0.3973	0.6723
$\alpha = 0.1$, $\beta = 1.6$, $\delta = 2.0$	0.4211	0.3605	0.3908	0.6691
$\alpha = 0.1$, $\beta = 1.8$, $\delta = 2.0$	0.4116	0.3550	0.3833	0.6653
$\alpha = 0.1$, $\beta = 2.0$, $\delta = 2.0$	0.4040	0.3503	0.3772	0.6601
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 1.0$	0.3993	0.3467	0.3730	0.6624
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 3.0$	0.4178	0.3588	0.3883	0.6733
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 4.0$	0.4239	0.3624	0.3932	0.6758
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 5.0$	0.4278	0.3647	0.3963	0.6751
$\alpha = 0.1$, $\beta = 1.2$, $\delta = 6.0$	0.4303	0.3457	**0.3980**	0.6727
square root	0.4182	0.3537	0.3860	0.6689
arctangent	0.4288	0.3617	0.3953	0.6768

Table 3: Spearman's rank correlation coefficient of system-level in NTCIR-7.

Metrics	Adequacy (14 systems)	Fluency (14 systems)	Avg.
AILE	0.9912	0.9253	**0.9582**
BLEU	0.8505	0.8242	0.8374
IMPACT	0.9912	0.9253	**0.9582**
METEOR	0.8022	0.7538	0.7780
RIBES	0.9121	0.8374	0.8747
TER	-0.9473	-0.8769	-0.9121

correlations in both system-level and sentence-level were obtained using AILE. In the second and third experiments, the references and candidates were respectively obtained from WMT10 (C. Callison-Burch et al., 2010) and WMT11 (C. Callison-Burch et al., 2011). In these experiments, as candidate we used the machine translation system's translations of European (*i.e.*, Czech, German, Spanish and French) sentences into English sentences, compared to a single reference. The correlations with system-level translations were obtained using AILE in these experiments.

Moreover, we used the following automatic evaluation metrics: BLEU (ver. 12), METEOR (ver. 1.4), RIBES (ver. 1.02.3), TER (tercom ver. 0.7.25), and IMPACT (ver. 4.0.2) to compare with AILE. In all experiments, the software "tokenizer.perl" and "lowercase.perl" (P. Koehn, 2011) were used for all references and candidates before the evaluation scores were calculated using the metrics.

3.2 Experimental Results

Tables 1 and 2 respectively provide Spearman's rank correlation coefficients of system-level and Kendall tau rank correlation coefficients of sentence-level in AILE based on the various values of parameters. In Table 2, "Total Avg." indicates the average value between "Avg." in Table 1 and "Avg." in Table 2. Moreover, "square root" and "arctangent" respectively indicate the correlation coefficients obtained by replacing $log(m + n)$ in Eq. (7)

Table 4: Kendall tau rank correlation coefficient of sentence-level in NTCIR-7.

Metrics	Adequacy (1,400 sentences)	Fluency (1,400 sentences)	Avg.
AILE	0.4304	0.3627	**0.3965**
BLEU	0.1146	0.1491	0.1319
IMPACT	0.4138	0.3503	0.3820
METEOR	0.1838	0.2060	0.1949
RIBES	0.3558	0.2950	0.3254
TER	-0.2664	-0.2605	-0.2635

Table 5: Spearman's rank correlation coefficient of system-level in WMT10.

Metrics	cz-en (12 systems)	de-en (25 systems)	es-en (14 systems)	fr-en (24 systems)	Avg.
AILE	0.6573	0.6769	0.6029	0.5878	0.6312
BLEU	0.7203	0.7885	0.3890	0.6862	**0.6460**
IMPACT	0.6643	0.7115	0.6381	0.5635	0.6443
METEOR	0.5594	0.8538	0.4330	0.4957	0.5855
RIBES	0.4895	0.5423	0.6615	0.5200	0.5533
TER	-0.8042	-0.3700	-0.5429	-0.3983	-0.5288

with $\sqrt{m+n}$ and $\arctan(x+y)$. In these case, 0.1, 1.2 and 2.0 were respectively used as the values of parameters α, β and δ.

Tables 3 and 4 respectively provide Spearman's rank correlation coefficients of system-level and Kendall tau rank correlation coefficients of sentence-level in NTCIR-7(A. Fujii et al., 2008). Table 5 provides the Spearman's rank correlation coefficient of system-level in WMT10 (C. Callison-Burch et al., 2010). Table 6 provides the Spearman's rank correlation coefficient of system-level in WMT11(C. Callison-Burch et al., 2011). In Table 6, "indiv" and "comb" respectively indicate a single machine translation system and the combination of two machine translation systems.

3.3 Discussion

Through Table 2, the value 0.6774 was the highest value in "Total Avg.". Therefore, 0.1, 1.2, and 2.0 were determined as the most suitable values of parameters α, β and δ respectively. In AILE of Tables 3-6, their values were used as the values.

AILE provided the highest correlation with human judgments, except for Table 5. These results show the effectiveness of AILE. Moreover, we investigated the effectiveness of AILE in short sentences and long sentences. The AILE can obtain a high correlation by decreasing the weight of mismatched words in short sentences. We performed the experiments using two data sets in which the numbers of word in the pairs of the reference and candidate are respectively small and large. In NTCIR-7 data, the average of word number in all pairs of the reference and candidate is 61.59. Therefore, we divided all pairs in two kinds of data. One is the pairs of short sentences (numbers of words in reference and candidate under 60). Another is the pairs of long sentences (numbers of words in reference and candidate over 61). The number of short sentence pairs is 763 and the number of long sentence pairs is 637. Moreover, we used AILE with *weight* and AILE without *weight* to confirm the effectiveness of *weight* in Eq. (7). Tables 7 and 8 provide Kendall tau rank correlation coefficients of sentence-level using short sentences and long sentences. In system-level, the Spearman's rank correlation coefficients of AILE using *weight* are the same as those of AILE without *weight*.

Through Table 7, the correlation coefficients of AILE using *weight* are higher them of AILE without *weight*. The value of "Avg." improved 0.0043 (from 0.3729 to 0.3772) using *weight* of Eq. (7) in long sentences. On the

Table 6: Spearman's rank correlation coefficient of system-level in WMT11.

Metrics	cz-en indiv (8 systems)	de-en indiv (20 systems)	es-en indiv (15 systems)	es-en comb (6 systems)
AILE	0.9048	0.1729	0.7571	-0.0857
BLEU	0.8333	0.2309	0.8204	-0.1739
IMPACT	0.9048	0.1722	0.7857	-0.3714
METEOR	0.9286	0.5308	0.8321	-0.6000
RIBES	0.8333	0.0406	0.5393	-0.0667
TER	-0.9524	-0.1985	-0.7250	0.8286

Metrics	fr-en indiv (18 systems)	fr-en comb (6 systems)	Avg.
AILE	0.7503	0.7714	**0.5451**
BLEU	0.7730	-0.1449	0.3898
IMPACT	0.7750	0.6377	0.4840
METEOR	0.7998	0.0857	0.4295
RIBES	0.7337	-0.0857	0.3324
TER	-0.7564	0.0286	-0.2959

Table 7: Kendall tau rank correlation coefficient of sentence-level in long sentences.

Metrics	Adequacy (637 sentences)	Fluency (637 sentences)	Avg.
AILE using *weight*	0.4011	0.3532	0.3772
AILE without *weight*	0.3975	0.3482	0.3729

other hand, in Table 8, the value of "Avg." improved 0.0096 (from 0.3461 to 0.3557) using *weight* of Eq. (7) in short sentences. These results indicate the effectiveness of the use of *weight* in Eq. (7). Especially, *weight* is effective in short sentences described in Section 2. The improved value 0.0096 in short sentences is higher than 0.0043 in long sentences. Therefore, we confirmed that *weight* of Eq. (7) is especially effective in short sentences. As a result, AILE can obtain stable evaluation scores without depending on sentence length.

4 Conclusion

In this paper, we proposed a new automatic evaluation metric, in which the weight of each mismatched word does not depend heavily on sentence length. Our metric can obtain stable evaluation scores that are not distorted by sentence length. Our experimental results indicated that the correlation coefficient of our metric is the highest among some leading metrics. Therefore, we confirmed the effectiveness of our metric.

Future studies will work to increase the correlation coefficients. Moreover, we will use our metric as tuning in SMT. The AILE software will be released as IMPACT version 4.0.3 by `http://www.lst.hokkai-s-u.ac.jp/~echi/impact.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

C. Liu, D. Dahlmeier and H. Tou Ng. 2010. TESLA: Translation Evaluation of Sentences with Linear-programming-based Analysis. Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR. pp.354–359.

S. Padó, M. Galley, D. Jurafsky and C. D. Manning. 2009. Textual Entailment Features for

Table 8: Kendall tau rank correlation coefficient of sentence-level in short sentences.

Metrics	Adequacy (763 sentences)	Fluency (763 sentences)	Avg.
AILE using *weight*	0.3897	0.3217	0.3557
AILE without *weight*	0.3774	0.3147	0.3461

Machine Translation Evaluation. Proceedings of the Fourth Workshop on Statistical Machine Translation.

J. Giménez and L. Márquez. Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations. 2007. Proceedings of IJCNLP. pp.319–326.

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. pp.228–231.

Y. Seng Chan and H. Tou Ng. 2008. MAXSIM: An Automatic Metric for Machine Translation Evaluation Based on Maximum Similarity. Proceedings of the Metrics-MATR Workshop of AMTA-2008. pp.319–326.

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.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas (AMTA). pp.223–231.

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.

H. Echizen-ya and K. Araki. 2007. Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum. Proceedings of the Eleventh Machine Translation Summit. pp.151–158.

H. Echizen'ya, K. Araki and H. Hovy. 2012. Optimization for Efficient Determination of Chunk in Automatic Evaluation for Machine Translation. Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology. pp.17–30.

P. Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, Cambridge, UK.

H. Echizen-ya, T. Ehara, S. Shimohata, A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, N. Kando. 2009. Meta-Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NTCIR-7. Proceedings of the Third Workshop on Patent Translation. pp.9–16.

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.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki and O. F. Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. Proceedings of the Join Fifth Workshop on Statistical Machine Translation and Metrics MATR. pp.17–53.

C. Callison-Burch, P. Koehn, C. Monz and O. F. Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation, Proceedings of the Sixth Workshop on Statistical Machine Translation. Proceedings of the Sixth Workshop on Statistical Machine Translation. pp.22–64.

H. Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing. pp.389–400.

P. Koehn. 2011. `http://www.statmt.org/wmt11/translation-task.html`.