A Description of Tunable Machine Translation Evaluation Systems in WMT13 Metrics Task

Aaron L.-F. Han
hanlifengaaron@gmail.com

Derek F. Wong
derekfw@umac.mo

Lidia S. Chao
lidiasc@umac.mo

Yi Lu
mb25435@umac.mo

Liangye He
wutianshui0515@gmail.com

Yiming Wang
mb25433@umac.mo

Jiaji Zhou
mb25473@uamc.mo

Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory
Department of Computer and Information Science
University of Macau, Macau S.A.R. China

Abstract

This paper is to describe our machine translation evaluation systems used for participation in the WMT13 shared Metrics Task. In the Metrics task, we submitted two automatic MT evaluation systems nLEPOR_baseline and LEPOR_v3.1. nLEPOR_baseline is an n-gram based language independent MT evaluation metric employing the factors of modified sentence length penalty, position difference penalty, n-gram precision and n-gram recall. nLEPOR_baseline measures the similarity of the system output translations and the reference translations only on word sequences. LEPOR_v3.1 is a new version of LEPOR metric using the mathematical harmonic mean to group the factors and employing some linguistic features, such as the part-of-speech information. The evaluation results of WMT13 show LEPOR_v3.1 yields the highest average-score 0.86 with human judgments at system-level using Pearson correlation criterion on English-to-other (FR, DE, ES, CS, RU) language pairs.

1 Introduction

Machine translation has a long history since the 1950s (Weaver, 1955) and gains a fast development in the recent years because of the higher level of computer technology. For instances, Och (2003) presents Minimum Error Rate Training (MERT) method for log-linear statistical machine translation models to achieve better translation quality; Menezes et al. (2006) introduce a syntactically informed phrasal SMT system for English-to-Spanish translation using a phrase translation model, which is based on global reordering and the dependency tree; Su et al. (2009) use the Thematic Role Templates model to improve the translation; Costa-jussà et al. (2012) develop the phrase-based SMT system for Chinese-Spanish translation using a pivot language. With the rapid development of Machine Translation (MT), the evaluation of MT has become a challenge in front of researchers. However, the MT evaluation is not an easy task due to the fact of the diversity of the languages, especially for the evaluation between distant languages (English, Russia, Japanese, etc.).

2 Related works

The earliest human assessment methods for machine translation include the intelligibility and fidelity used around 1960s (Carroll, 1966), and the adequacy (similar as fidelity), fluency and comprehension (improved intelligibility) (White et al., 1994). Because of the expensive cost of manual evaluations, the automatic evaluation metrics and systems appear recently.

The early automatic evaluation metrics include the word error rate WER (Su et al., 1992) and position independent word error rate PER (Tillmann et al., 1997) that are based on the Levenshtein distance. Several promotions for the MT and MT evaluation literatures include the ACL's annual workshop on statistical machine translation WMT (Koehn and Monz, 2006; Callison-Burch et al., 2012), NIST open machine translation (OpenMT) Evaluation series (Li, 2005) and the international workshop of spoken language translation IWSLT, which is also organized annually from 2004 (Eck and Hori, 2005;

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 414–421,
Sofia, Bulgaria, August 8-9, 2013 ©2013 Association for Computational Linguistics

Paul, 2008, 2009; Paul, et al., 2010; Federico et al., 2011).

BLEU (Papineni et al., 2002) is one of the commonly used evaluation metrics that is designed to calculate the document level precisions. NIST (Doddington, 2002) metric is proposed based on BLEU but with the information weights added to the n-gram approaches. TER (Snover et al., 2006) is another well-known MT evaluation metric that is designed to calculate the amount of work needed to correct the hypothesis translation according to the reference translations. TER includes the edit categories such as insertion, deletion, substitution of single words and the shifts of word chunks. Other related works include the METEOR (Banerjee and Lavie, 2005) that uses semantic matching (word stem, synonym, and paraphrase), and (Wong and Kit, 2008), (Popovic, 2012), and (Chen et al., 2012) that introduces the word order factors, etc. The traditional evaluation metrics tend to perform well on the language pairs with English as the target language. This paper will introduce the evaluation models that can also perform well on the language pairs that with English as source language.

3 Description of Systems

3.1 Sub Factors

Firstly, we introduce the sub factor of modified length penalty inspired by BLEU metric.

$$LP = \begin{cases} e^{1-\frac{r}{c}} & if \ \ c < r \\ 1 & if \ \ c = r \\ e^{1-\frac{c}{r}} & if \ \ c > r \end{cases} \quad (1)$$

In the formula, LP means sentence length penalty that is designed for both the shorter or longer translated sentence (hypothesis translation) as compared to the reference sentence. Parameters c and r represent the length of candidate sentence and reference sentence respectively.

Secondly, let's see the factors of n-gram precision and n-gram recall.

$$P_n = \frac{\#ngram_{matched}}{\#ngram\ chunks\ in\ hypothesis} \quad (2)$$

$$R_n = \frac{\#ngram_{matched}}{\#ngram\ chunks\ in\ reference} \quad (3)$$

The variable $\#ngram_{matched}$ represents the number of matched *n*-gram chunks between hypothesis sentence and reference sentence. The *n*-gram precision and *n*-gram recall are firstly calculated on sentence-level instead of corpus-level that is used in BLEU (P_n). Then we define the weighted *n*-gram harmonic mean of precision and recall (*WNHPR*).

$$WNHPR = exp(\sum_{n=1}^{N} w_n logH(\alpha R_n, \beta P_n))(4)$$

Thirdly, it is the *n*-gram based position difference penalty (*NPosPenal*). This factor is designed to achieve the penalty for the different order of successfully matched words in reference sentence and hypothesis sentence. The alignment direction is from the hypothesis sentence to the reference sentence. It employs the n-gram method into the matching period, which means that the potential matched word will be assigned higher priority if it also has nearby matching. The nearest matching will be accepted as a backup choice if there are both nearby matching or there is no other matched word around the potential pairs.

$$NPosPenal = e^{-NPD} \quad (5)$$

$$NPD = \frac{1}{Length_{hyp}} \sum_{i=1}^{Length_{hyp}} |PD_i| \quad (6)$$

$$|PD_i| = |MatchN_{hyp} - MatchN_{ref}| \quad (7)$$

The variable $Length_{hyp}$ means the length of the hypothesis sentence; the variables $MatchN_{hyp}$ and $MatchN_{ref}$ represent the position number of matched words in hypothesis sentence and reference sentence respectively.

3.2 Linguistic Features

The linguistic features could be easily employed into our evaluation models. In the submitted experiment results of WMT Metrics Task, we used the part of speech information of the words in question. In grammar, a part of speech, which is also called a word class, a lexical class, or a lexical category, is a linguistic category of lexical items. It is generally defined by the syntactic or morphological behavior of the lexical item in question. The POS information utilized in our metric LEPOR_v3.1, an enhanced version of LEPOR (Han et al., 2012), is extracted using the Berkeley parser (Petrov et al., 2006) for English, German, and French languages, using COMPOST Czech morphology tagger (Collins, 2002) for Czech language, and using TreeTagger (Schmid, 1994) for Spanish and Russian languages respectively.

Ratio	other-to-English				English-to-other			
	CZ-EN	DE-EN	ES-EN	FR-EN	EN-CZ	EN-DE	EN-ES	EN-FR
HPR:LP:NPP(word)	7:2:1	3:2:1	7:2:1	3:2:1	7:2:1	1:3:7	3:2:1	3:2:1
HPR:LP:NPP(POS)	NA	3:2:1	NA	3:2:1	7:2:1	7:2:1	NA	3:2:1
$\alpha:\beta$ (word)	1:9	9:1	1:9	9:1	9:1	9:1	9:1	9:1
$\alpha:\beta$ (POS)	NA	9:1	NA	9:1	9:1	9:1	NA	9:1
$w_{hw}:w_{hp}$	NA	1:9	NA	9:1	1:9	1:9	NA	9:1

Table 1. The tuned weight values in LEPOR_v3.1 system

System	Correlation Score with Human Judgment								
	other-to-English				English-to-other				Mean score
	CZ-EN	DE-EN	ES-EN	FR-EN	EN-CZ	EN-DE	EN-ES	EN-FR	
LEPOR_v3.1	0.93	0.86	0.88	0.92	0.83	0.82	0.85	0.83	**0.87**
nLEPOR_baseline	0.95	0.61	0.96	0.88	0.68	0.35	0.89	0.83	0.77
METEOR	0.91	0.71	0.88	0.93	0.65	0.30	0.74	0.85	0.75
BLEU	0.88	0.48	0.90	0.85	0.65	0.44	0.87	0.86	0.74
TER	0.83	0.33	0.89	0.77	0.50	0.12	0.81	0.84	0.64

Table 2. The performances of nLEPOR_baseline and LEPOR_v3.1 systems on WMT11 corpora

3.3 The nLEPOR_baseline System

The nLEPOR_baseline system utilizes the simple product value of the factors: modified length penalty, *n*-gram position difference penalty, and weighted *n*-gram harmonic mean of precision and recall.

$$nLEPOR = LP \times PosPenalty \times WNHPR \quad (8)$$

The system level score is the arithmetical mean of the sentence level evaluation scores. In the experiments of Metrics Task using the nLEPOR_baseline system, we assign *N*=1 in the factor WNHPR, i.e. weighted unigram harmonic mean of precision and recall.

3.4 The LEPOR_v3.1 System

The system of LEPOR_v3.1 (also called as hLEPOR) combines the sub factors using weighted mathematical harmonic mean instead of the simple product value.

$$hLEPOR = \frac{w_{LP}+w_{NPosPenal}+w_{HPR}}{\frac{w_{LP}}{LP}+\frac{w_{NPosPenal}}{NPosPenal}+\frac{w_{HPR}}{HPR}} \quad (9)$$

Furthermore, this system takes into account the linguistic features, such as the POS of the words. Firstly, we calculate the hLEPOR score on surface words $hLEPOR_{word}$ (the closeness of the hypothesis translation and the reference translation). Then, we calculate the hLEPOR score on the extracted POS sequences $hLEPOR_{POS}$ (the closeness of the corresponding POS tags between hypothesis sentence and reference sentence). The final score $hLEPOR_{final}$ is the combination of the two sub-scores $hLEPOR_{word}$ and $hLEPOR_{POS}$.

$$hLEPOR_{final} = \frac{1}{w_{hw}+w_{hp}}(w_{hw}hLEPOR_{word} + w_{hp}hLEPOR_{POS}) \quad (10)$$

4 Evaluation Method

In the MT evaluation task, the Spearman rank correlation coefficient method is usually used by the authoritative ACL WMT to evaluate the correlation of different MT evaluation metrics. So we use the Spearman rank correlation coefficient ρ to evaluate the performances of nLEPOR_baseline and LEPOR_v3.1 in system level correlation with human judgments. When there are no ties, ρ is calculated using:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (11)$$

The variable d_i is the difference value between the ranks for $system_i$ and n is the number of systems. We also offer the Pearson correlation coefficient information as below. Given a sample of paired data (X, Y) as (x_i, y_i), $i = 1\ to\ n$, the Pearson correlation coefficient is:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i-\mu_y)^2}} \quad (12)$$

where μ_x and μ_y specify the mean of discrete random variable X and Y respectively.

Directions	EN-FR	EN-DE	EN-ES	EN-CS	EN-RU	Av
LEPOR_v3.1	*.91*	*.94*	*.91*	*.76*	***.77***	***.86***
nLEPOR_baseline	*.92*	*.92*	*.90*	***.82***	*.68*	*.85*
SIMP-BLEU_RECALL	**.95**	.93	.90	**.82**	.63	.84
SIMP-BLEU_PREC	.94	.90	.89	**.82**	.65	.84
NIST-mteval-inter	.91	.83	.84	.79	.68	.81
Meteor	.91	.88	.88	**.82**	.55	.81
BLEU-mteval-inter	.89	.84	.88	.81	.61	.80
BLEU-moses	.90	.82	.88	.80	.62	.80
BLEU-mteval	.90	.82	.87	.80	.62	.80
CDER-moses	.91	.82	.88	.74	.63	.80
NIST-mteval	.91	.79	.83	.78	.68	.79
PER-moses	.88	.65	.88	.76	.62	.76
TER-moses	.91	.73	.78	.70	.61	.75
WER-moses	.92	.69	.77	.70	.61	.74
TerrorCat	.94	**.96**	**.95**	na	na	.95
SEMPOS	na	na	na	.72	na	.72
ACTa	.81	-.47	na	na	na	.17
ACTa5+6	.81	-.47	na	na	na	.17

Table 3. System-level Pearson correlation scores on WMT13 English-to-other language pairs

5 Experiments

5.1 Training

In the training stage, we used the officially released data of past WMT series. There is no Russian language in the past WMT shared tasks. So we trained our systems on the other eight language pairs including English to other (French, German, Spanish, Czech) and the inverse translation direction. In order to avoid the overfitting problem, we used the WMT11 corpora as training data to train the parameter weights in order to achieve a higher correlation with human judgments at system-level evaluations. For the nLEPOR_baseline system, the tuned values of α and β are 9 and 1 respectively for all language pairs except for ($\alpha = 1, \beta = 9$) for CS-EN language pair. For the LEPOR_v3.1 system, the tuned values of weights are shown in Table 1. The evaluation scores of the training results on WMT11 corpora are shown in Table 2. The designed methods have shown promising correlation scores with human judgments at system level, 0.87 and 0.77 respectively for nLEPOR_baseline and LEPOR_v3.1 of the mean score on eight language pairs. As compared to METEOR, BLEU and TER, we have achieved higher correlation scores with human judgments.

5.2 Testing

In the WMT13 shared Metrics Task, we also submitted our system performances on English-to-Russian and Russian-to-English language pairs. However, since the Russian language did not appear in the past WMT shared tasks, we assigned the default parameter weights to Russian language for the submitted two systems. The officially released results on WMT13 corpora are shown in Table 3, Table 4 and Table 5 respectively for system-level and segment-level performance on English-to-other language pairs.

Directions	EN-FR	EN-DE	EN-ES	EN-CS	EN-RU	Av
SIMP-BLEU_RECALL	.92	.93	.83	.87	.71	**.85**
LEPOR_v3.1	*.90*	*.9*	*.84*	*.75*	***.85***	***.85***
NIST-mteval-inter	**.93**	.85	.80	.90	.77	**.85**
CDER-moses	.92	.87	.86	.89	.70	**.85**
nLEPOR_baseline	*.92*	*.90*	*.85*	*.82*	*.73*	*.84*
NIST-mteval	.91	.83	.78	.92	.72	.83
SIMP-BLEU_PREC	.91	.88	.78	.88	.70	.83
Meteor	.92	.88	.78	**.94**	.57	.82
BLEU-mteval-inter	.92	.83	.76	.90	.66	.81
BLEU-mteval	.89	.79	.76	.90	.63	.79
TER-moses	.91	.85	.75	.86	.54	.78
BLEU-moses	.90	.79	.76	.90	.57	.78
WER-moses	.91	.83	.71	.86	.55	.77
PER-moses	.87	.69	.77	.80	.59	.74
TerrorCat	**.93**	**.95**	**.91**	na	na	.93
SEMPOS	na	na	na	.70	na	.70
ACTa5+6	.81	-.53	na	na	na	.14
ACTa	.81	-.53	na	na	na	.14

Table 4. System-level Spearman rank correlation scores on WMT13 English-to-other language pairs

Table 3 shows LEPOR_v3.1 and nLEPOR_baseline yield the highest and the second highest average Pearson correlation score 0.86 and 0.85 respectively with human judgments at system-level on five English-to-other language pairs. LEPOR_v3.1 and nLEPOR_baseline also yield the highest Pearson correlation score on English-to-Russian (0.77) and English-to-Czech (0.82) language pairs respectively. The testing results of LEPOR_v3.1 and nLEPOR_baseline show better correlation scores as compared to METEOR (0.81), BLEU (0.80) and TER-moses (0.75) on English-to-other language pairs, which is similar with the training results.

On the other hand, using the Spearman rank correlation coefficient, SIMPBLEU_RECALL yields the highest correlation score 0.85 with human judgments. Our metric LEPOR_v3.1 also yields the highest Spearman correlation score on English-to-Russian (0.85) language pair, which is similar with the result using Pearson correlation and shows its robust performance on this language pair.

Directions	EN-FR	EN-DE	EN-ES	EN-CS	EN-RU	Av
SIMP-BLEU_RECALL	**.16**	**.09**	**.23**	**.06**	**.12**	**.13**
Meteor	.15	.05	.18	**.06**	.11	.11
SIMP-BLEU_PREC	.14	.07	.19	**.06**	.09	.11
sentBLEU-moses	.13	.05	.17	.05	.09	.10
LEPOR_v3.1	*.13*	*.06*	*.18*	*.02*	*.11*	*.10*
nLEPOR_baseline	*.12*	*.05*	*.16*	*.05*	*.10*	*.10*
dfki_logregNorm-411	na	na	.14	na	na	.14
TerrorCat	.12	.07	.19	na	na	.13
dfki_logregNormSoft-431	na	na	.03	na	na	.03

Table 5. Segment-level Kendall's tau correlation scores on WMT13 English-to-other language pairs

However, we find a problem in the Spearman rank correlation method. For instance, let two evaluation metrics MA and MB with their evaluation scores $\overrightarrow{MA} = \{0.95,0.50,0.45\}$ and $\overrightarrow{MB} = \{0.77,0.75,0.74\}$ respectively reflecting three MT systems $\vec{M} = \{M_1, M_2, M_3\}$. Before the calculation of correlation with human judgments, they will be converted into $\breve{\overrightarrow{MA}} = \{1,2,3\}$ and $\breve{\overrightarrow{MB}} = \{1,2,3\}$ with the same rank sequence using Spearman rank method; thus, the two evaluation systems will get the same correlation with human judgments whatever are the values of human judgments. But the two metrics reflect different results indeed: MA gives the outstanding score (0.95) to M_1 system and puts very low scores (0.50 and 0.45) on the other two systems M_2 and M_3 while MB thinks the three MT systems have similar performances (scores from 0.74 to 0.77). This information is lost using the Spearman rank correlation methodology.

The segment-level performance of LEPOR_v3.1 is moderate with the average Kendall's tau correlation score 0.10 on five English-to-other language pairs, which is due to the fact that we trained our metrics at system-level in this shared metrics task. Lastly, the officially released results on WMT13 other-to-English language pairs are shown in Table 6, Table 7 and Table 8 respectively for system-level and segment-level performance.

Directions	FR-EN	DE-EN	ES-EN	CS-EN	RU-EN	Av
Meteor	**.98**	.96	.97	**.99**	**.84**	**.95**
SEMPOS	.95	.95	.96	**.99**	.82	.93
Depref-align	.97	.97	.97	.98	.74	.93
Depref-exact	.97	.97	.96	.98	.73	.92
SIMP-BLEU_RECALL	.97	.97	.96	.94	.78	.92
UMEANT	.96	.97	**.99**	.97	.66	.91
MEANT	.96	.96	**.99**	.96	.63	.90
CDER-moses	.96	.91	.95	.90	.66	.88
SIMP-BLEU_PREC	.95	.92	.95	.91	.61	.87
LEPOR_v3.1	*.96*	*.96*	*.90*	*.81*	*.71*	*.87*
nLEPOR_baseline	*.96*	*.94*	*.94*	*.80*	*.69*	*.87*
BLEU-mteval-inter	.95	.92	.94	.90	.61	.86
NIST-mteval-inter	.94	.91	.93	.84	.66	.86
BLEU-moses	.94	.91	.94	.89	.60	.86
BLEU-mteval	.95	.90	.94	.88	.60	.85
NIST-mteval	.94	.90	.93	.84	.65	.85
TER-moses	.93	.87	.91	.77	.52	.80
WER-moses	.93	.84	.89	.76	.50	.78
PER-moses	.84	.88	.87	.74	.45	.76
TerrorCat	**.98**	**.98**	.97	na	na	.98

Table 6. System-level Pearson correlation scores on WMT13 other-to-English language pairs

METEOR yields the highest average correlation scores 0.95 and 0.94 respectively using Pearson and Spearman rank correlation methods on other-to-English language pairs. The average performance of nLEPOR_baseline is a little better than LEPOR_v3.1 on the five language pairs of other-to-English even though it is also moderate as compared to other metrics. However, using the Pearson correlation method, nLEPOR_baseline yields the average correlation score 0.87 which already wins the BLEU (0.86) and TER (0.80) as shown in Table 6.

Directions	FR-EN	DE-EN	ES-EN	CS-EN	RU-EN	Av
Meteor	.98	.96	**.98**	.96	.81	**.94**
Depref-align	**.99**	**.97**	.97	.96	.79	**.94**
UMEANT	**.99**	.95	.96	**.97**	.79	.93
MEANT	.97	.93	.94	**.97**	.78	.92
Depref-exact	.98	.96	.94	.94	.76	.92
SEMPOS	.94	.92	.93	.95	**.83**	.91
SIMP-BLEU_RECALL	.98	.94	.92	.91	.81	.91
BLEU-mteval-inter	**.99**	.90	.90	.94	.72	.89
BLEU-mteval	**.99**	.89	.89	.94	.69	.88
BLEU-moses	**.99**	.90	.88	.94	.67	.88
CDER-moses	**.99**	.88	.89	.93	.69	.87
SIMP-BLEU_PREC	**.99**	.85	.83	.92	.72	.86
nLEPOR_baseline	*.95*	*.95*	*.83*	*.85*	*.72*	*.86*
LEPOR_v3.1	*.95*	*.93*	*.75*	*0.8*	*.79*	*.84*
NIST-mteval	.95	.88	.77	.89	.66	.83
NIST-mteval-inter	.95	.88	.76	.88	.68	.83
TER-moses	.95	.83	.83	0.8	0.6	0.80
WER-moses	.95	.67	.80	.75	.61	.76
PER-moses	.85	.86	.36	.70	.67	.69
TerrorCat	.98	.96	.97	na	na	.97

Table 7. System-level Spearman rank correlation scores on WMT13 other-to-English language pairs

Once again, our metrics perform moderate at segment-level on other-to-English language pairs due to the fact that they are trained at system-level. We notice that some of the evaluation metrics do not submit the results on all the language pairs; however, their performance on submitted language pair is sometimes very good, such as the dfki_logregFSS-33 metric with a segment-level correlation score 0.27 on German-to-English language pair.

6 Conclusion

This paper describes our participation in the WMT13 Metrics Task. We submitted two systems nLEPOR_baseline and LEPOR_v3.1. Both of the two systems are trained and tested using the officially released data. LEPOR_v3.1 yields the highest Pearson correlation average-score 0.86 with human judgments on five English-to-other language pairs, and nLEPOR_baseline yields better performance than LEPOR_v3.1 on other-to-English language pairs. Furthermore, LEPOR_v3.1 shows robust system-level performance on English-to-Russian language pair, and nLEPOR_baseline shows best system-level performance on English-to-Czech language pair using Pearson correlation criterion. As compared to nLEPOR_baseline, the experiment results of LEPOR_v3.1 also show that the proper use of linguistic information can increase the performance of the evaluation systems.

Directions	FR-EN	DE-EN	ES-EN	CS-EN	RU-EN	Av
SIMP-BLEU_RECALL	**.19**	**.32**	**.28**	.26	.23	**.26**
Meteor	.18	.29	.24	**.27**	**.24**	.24
Depref-align	.16	.27	.23	.23	.20	.22
Depref-exact	.17	.26	.23	.23	.19	.22
SIMP-BLEU_PREC	.15	.24	.21	.21	.17	.20
nLEPOR_baseline	*.15*	*.24*	*.20*	*.18*	*.17*	*.19*
sentBLEU-moses	.15	.22	.20	.20	.17	.19
LEPOR_v3.1	*.15*	*.22*	*.16*	*.19*	*.18*	*.18*
UMEANT	.10	.17	.14	.16	.11	.14
MEANT	.10	.16	.14	.16	.11	.14
dfki_logregFSS-33	na	.27	na	na	na	.27
dfki_logregFSS-24	na	.27	na	na	na	.27
TerrorCat	.16	.30	.23	na	na	.23

Table 8. Segment-level Kendall's tau correlation scores on WMT13 other-to-English language pairs

Acknowledgments

The authors wish to thank the anonymous reviewers for many helpful comments. The authors are grateful to the Science and Technology Development Fund of Macau and the Research Committee of the University of Macau for the funding support for our research, under the reference No. 017/2009/A and RG060/09-10S/CS/FST.

References

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the 43th Annual Meeting of the*

Association of Computational Linguistics (ACL- 05), pages 65–72, Ann Arbor, US, June. Association of Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar F. Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (WMT '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 22-64.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montreal, Canada. Association for Computational Linguistics.

Carroll, John B. 1966. An Experiment in Evaluating the Quality of Translations, *Mechanical Translation and Computational Linguistics*, vol.9, nos.3 and 4, September and December 1966, page 55-66, Graduate School of Education, Harvard University.

Chen, Boxing, Roland Kuhn and Samuel Larkin. 2012. PORT: a Precision-Order-Recall MT Evaluation Metric for Tuning, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 930–939, Jeju, Republic of Korea, 8-14 July 2012.

Collins, Michael. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Volume 10 (EMNLP 02), pages 1-8. Association for Computational Linguistics, Stroudsburg, PA, USA.

Costa-jussà, Marta R., Carlos A. Henríquez and Rafael E. Banchs. 2012. Evaluating Indirect Strategies for Chinese-Spanish Statistical Machine Translation. *Journal of artificial intelligence research*, Volume 45, pages 761-780.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research* (HLT '02). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 138-145.

Eck, Matthias and Chiori Hori. 2005. Overview of the IWSLT 2005 Evaluation Campaign. *Proceedings of IWSLT 2005*.

Federico, Marcello, Luisa Bentivogli, Michael Paul, and Sebastian Stiiker. 2011. Overview of the IWSLT 2011 Evaluation Campaign. In *Proceedings of IWSLT 2011*, 11-27.

Han, Aaron Li-Feng, Derek F. Wong and Lidia S. Chao. 2012. LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors. *Proceedings of the 24th International Conference on Computational Linguistics* (COLING 2012: Posters), Mumbai, India.

Koehn, Philipp and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. *Proceedings of the ACLWorkshop on Statistical Machine Translation*, pages 102–121, New York City.

Li, A. (2005). Results of the 2005 NIST machine translation evaluation. In *Machine Translation Workshop.*

Menezes, Arul, Kristina Toutanova and Chris Quirk. 2006. Microsoft Research Treelet Translation System: NAACL 2006 Europarl Evaluation, *Proceedings of the ACLWorkshop on Statistical Machine Translation*, pages 158-161, New York City.

Och, Franz Josef. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (ACL-2003). pp. 160-167.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 311-318.

Paul, Michael. 2008. Overview of the IWSLT 2008 Evaluation Campaign. *Proceeding of IWLST 2008*, Hawaii, USA.

Paul, Michael. 2009. Overview of the IWSLT 2009 Evaluation Campaign. In *Proc. of IWSLT 2009*, Tokyo, Japan, pp. 1–18.

Paul, Michael, Marcello Federico and Sebastian Stiiker. 2010. Overview of the IWSLT 2010 Evaluation Campaign, *Proceedings of the 7th International Workshop on Spoken Language Translation*, Paris, December 2nd and 3rd, 2010, page 1-25.

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguis-*

tics (ACL-44). Association for Computational Linguistics, Stroudsburg, PA, USA, 433-440.

Popovic, Maja. 2012. Class error rates for evaluation of machine translation output. *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 71–75, Canada.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Snover, Matthew, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06)*, pages 223–231, USA. Association for Machine Translation in the Americas.

Su, Hung-Yu and Chung-Hsien Wu. 2009. Improving Structural Statistical Machine Translation for Sign Language With Small Corpus Using Thematic Role Templates as Translation Memory, *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, VOL. 17, NO. 7.

Su, Keh-Yih, Wu Ming-Wen and Chang Jing-Shin. 1992. A New Quantitative Quality Measure for Machine Translation Systems. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 433–439, Nantes, France.

Tillmann, Christoph, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated DP Based Search For Statistical Translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology* (EUROSPEECH-97).

Weaver, Warren. 1955. Translation. In William Locke and A. Donald Booth, editors, *Machine Translation of Languages: Fourteen Essays*. John Wiley & Sons, New York, pages 15–23.

White, John S., Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the Conference of the Association for Machine Translation in the Americas* (AMTA 1994). pp193-205.

Wong, Billy and Chunyu Kit. 2008. Word choice and word position for automatic MT evaluation. In Workshop: *MetricsMATR of the Association for Machine Translation in the Americas (AMTA)*, short paper, Waikiki, Hawai'I, USA. Association for Machine Translation in the Americas.